# Scalable clustering of segmented trajectories within a continuous time framework. Application to maritime traffic data.

Pierre Gloaguen, Laetitia Chapel, Chloé Friguet, Romain Tavenard

**Preprint**
**18/05/2020**

# Scalable clustering of segmented trajectories within a continuous time framework
# Application to maritime traffic data

Pierre Gloaguen[1], Laetitia Chapel[2], Chloé Friguet[2], Romain Tavenard[3]

**Abstract**
In the context of the surveillance of the maritime traffic, a major challenge is the automatic identification of traffic flows from a set of observed trajectories, in order to derive good management measures or to detect abnormal or illegal behaviours for example. In this paper, we propose a new modelling framework to cluster sequences of a large amount of trajectories recorded at potentially irregular frequencies. The model is specified within a continuous time framework, being robust to irregular sampling in records and accounting for possible heterogeneous movement patterns within a single trajectory. It partitions a trajectory into sub-trajectories, or *movement modes*, allowing a clustering of both individuals' movement patterns and trajectories. The clustering is performed using non parametric Bayesian methods, namely the hierarchical Dirichlet process, and considers a stochastic variational inference to estimate the model's parameters, hence providing a scalable method in an easy-to-distribute framework. Performance is assessed on both simulated data and on our motivational large trajectory dataset from the Automatic Identification System (AIS), used to monitor the world maritime traffic: the clusters represent significant, atomic motion-patterns, making the model informative for stakeholders.

**Keywords**
AIS data; Continuous time; Hierarchical Dirichlet process; Scalability; Stochastic variational inference; Trajectory clustering

[2] *Univ. Bretagne-Sud, IRISA-OBELIX, Vannes, France*
[1] *AgroParisTech, MIA-Paris, Paris, France*
[3] *Université de Rennes, LETG/IRISA, Rennes, France*

## 1. Introduction

For the last thirty years, the tracking of movements of individuals or objects has been eased by the increasing development of several tracking devices, such as the Global Positioning System (GPS) that allows recording geographical positions through time. Such devices have led to large movement databases that store trajectory data, in which each point represents a position in space at a given time. These databases contain a great deal of knowledge and require analysis [5] such as the extraction of movement patterns across diverse trajectories in order to perform human mobility or traffic monitoring [15, 36], motion prediction [25], human action recognition [32] or detection of abnormal behavior in maritime routes [29].

A striking example of large movement database comes from the worldwide maritime traffic surveillance. This monitoring relies on several sources of data, in a rising context of maritime big data [8]. Among these sources lies the Automatic Identification System (AIS), which automatically collects messages from vessels around the world, at a high frequency. AIS data basically consist in GPS-like data, together with the instantaneous speed and heading, and some vessel specific static information [4]. An example of such data is presented in Figure 1 (left), considering 6 months of AIS data of vessels steaming in the Ushant traffic separation scheme in Brittany, west of France [7]. These data are characterized by their diversity as they (1) are collected at different frequencies (2) have different lengths (3) are not necessarily regularly sampled (4) represent very different behaviors, but (5) share common trends or similar subparts (called hereafter *movement modes*).

One major challenge in this context is the extraction of movement patterns emerging from the observed data, considering trajectories that share similar movement modes (see Figure 1 (right)). This issue can be restated from a machine learning point of view as a large-scale clustering task involving the definition of clustering methods that can handle such complex data while being efficient on large databases, and that are able to both cluster trajectories as a whole and detect common sub-trajectories.

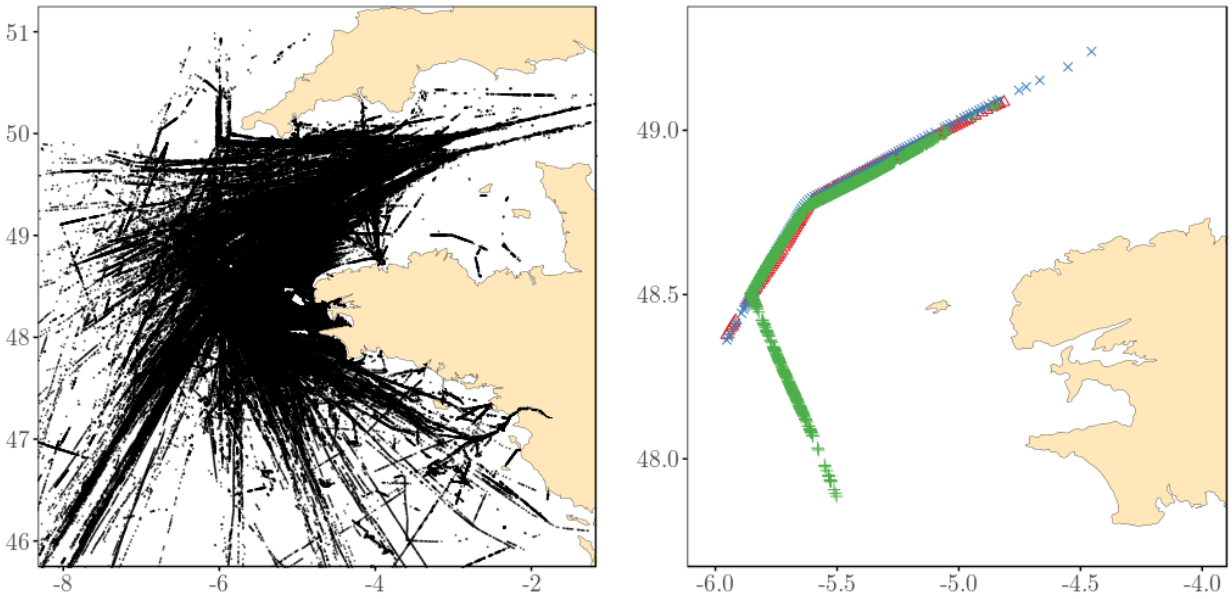


**Figure 1.** (Left) AIS data in the Ushant traffic separation scheme (Brittany, France), gathered during 6 months (7M. GPS obs.). Source: CLS, Brest, France and ANR SESAME [7] (Right) Three trajectories extracted from the dataset, sharing common *movement modes* (characterized by the directions West-South-West and South-South-West). The green (+ shaped points) trajectory, however, subsequently adopts another movement mode (South-South-East). The red (triangles) and blue (crosses) trajectories belong to the same cluster, which is different from the one of the green (+ shaped) trajectory.

Regarding the problem of clustering entire trajectories, existing methods in the literature follow two main trends.

- The first class of methods is a density-based clustering framework that relies on the definition of typical motions in the trajectories [6, 21]. Similarities between full trajectories are considered to take into account local details in the segments.
- The second class of methods is a rule-based clustering framework, that relies on pre-conceived decision rules to build each cluster. The main advantage of these methods is their simplicity, but they suppose the prior knowledge of decision rules, and afterwards make strong independence assumption on observed points, that might not hold in practice.

In a more parametric approach, one can aim at modelling the heterogeneous nature of movement data [16] to define a proper normality model, whose structure is made of clusters.

In this context, trajectories would belong to a cluster, and, moreover, a trajectory could be the bringing together of different moving patterns, hereafter called movement modes. In this context, different trajectories may share similar movement modes in a given portion. To learn common sub-trajectories in the clustering process, [14] propose a partition-and-group framework based on a dedicated distance. [31] proposed the definition of a normality model for trajectories through an unsupervised clustering framework inspired by topic models [24]. Initially developed in the field of text analysis, topic models aim at classifying documents depending on their main topics. The clustering approach then estimates the different topics in each document as well as clusters documents together regarding their topics. In their transposition of topic models to trajectory analysis, [31] consider trajectories as documents and quantized GPS observations (position and velocity) as words. It is assumed that each point of the trajectory belongs to a semantic region (a topic), called hereafter a *movement mode*. A movement mode is therefore a specific distribution to be estimated from trajectory data. Trajectories are then considered as a mixture of movement modes, and a trajectory cluster is a set of trajectories with a common mixture distribution. Estimation of both movement modes and trajectory clusters can be performed on a discretized space within a Bayesian framework using Gibbs sampling. If this approach still assumes independence of GPS observations (conditionally to their semantic region), it takes into account different heterogeneity levels for movement data. However, it requires a quantization of the space whose influence over results is not investigated, and relies on Gibbs sampling for inference, which is known to be hardly scalable [3].

Nevertheless, none of the aforementionned methods are able to deal with large databases, neither consider the continuous nature of the data. For an in-depth review of spatial trajectory clustering algorithms, one can refer to [34] or [35, Section 6].

Given the challenges mentionned above, we propose in this paper a novel model, based on topic models, to cluster large and complex trajectory datasets in a scalable framework.

Following these works, the approach proposed here distinguishes yet from existing works in the definition of GPS observations distributions as it is now assumed that (1) the distribution lives in a continuous space, avoiding the quantization of the (position/velocity) space, and (2) within a movement mode (or a topic), observations are not independent. Each movement mode is modeled through the bivariate velocity as a continuous time Gaussian process, namely the Ornstein Uhlenbeck Process (OUP) [28], that has been proposed in movement ecology as a flexible framework to model the fundamental units of movement [10]. Thanks to its continuous time formulation, the OUP allows handling irregularly sampled data. Moreover, as a Gaussian process, it accounts for observation autocorrelation. The unsupervised clustering is made in a non parametric Bayesian context using a Hierarchical Dirichlet Process framework [27]. To perform movement mode estimation in a scalable approach, we use asymptotic properties of the OUP together with stochastic variational inference [11]. Contrary to MCMC methods that obtain samples from the posterior distribution, the variational approach solves an optimization problem, allowing the use of distributed computation and stochastic optimization to scale and speed up inference [3].

The remainder of this paper is organized as follows: in Section 2, the hierarchical parametric framework that models trajectory data is fully specified, and the proposed scalable approach to estimate model parameters from data is described in Section 3. The method is evaluated in Section 4, both on simulated data sets, to assess inference performance, and on a new dataset of current AIS data that is released with the paper, to show the interest of our scalable approach in a realistic context.

## 2. Movement model

In this section, we define a parametric framework to model trajectory data, *i.e.* sequences of geographical positions recorded through time. The modelling framework aims to account for two levels of heterogeneity possibly present in trajectory data: (1) heterogeneity of an individual's movement within a single trajectory, and (2) heterogeneity between observed trajectories of several individuals.

In the following, the fundamental modelling of movement is made through the individual's velocity process. Velocity is observed directly (as in the AIS context) or is at least computed from the successive geographical locations of the individuals. The positions and velocities of an individual are modeled as two continuous time processes in $\mathbb{R}^2$, denoted by $(X_t)_{t\geq 0}$ and $(V_t)_{t\geq 0}$. Following a common paradigm, we assume that a moving individual's trajectory might be the collection of heterogeneous patterns [16], namely the *movement modes*. Different movement modes along a trajectory refer to different ways of moving in terms of velocity distribution, reflecting different behaviors, activities, or routes. It is assumed that the set of possible movement modes of tracked individuals during their trajectories is countable, and that a given movement mode can be adopted by several individuals.

### 2.1 Fundamental unit of movement

Similarly to [10], a movement mode is assumed to be characterized by a specific correlated velocity model, defined in a continuous-time framework. Formally, during a time segment $[\tau_1;\tau_2]$, if an individual adopts the movement mode $k$, then its velocity process $(V_t)_{\tau_1\leq t\leq\tau_2}$ is assumed to be the solution of the following Stochastic Differential Equation (SDE) (see [18] for a formal definition of SDEs):

$$\begin{aligned} &\mathrm{d}V_t = -\Gamma_k\,(V_t-\mu_k)\,\mathrm{d}t + \Sigma_k \mathrm{d}W_t,\ \ \tau_1\leq t\leq\tau_2,\\ &V_{\tau_1} = v_{\tau_1} \end{aligned} \tag{1}$$

where:

- $\mu_k\in\mathbb{R}^2$ is the asymptotic mean velocity of the $k$-th movement mode;
- $\Gamma_k$ is a $2\times 2$ real-valued matrix, and is an autocorrelation parameter, modelling the recall force of the process towards the mean $\mu_k$;
- $\Sigma_k$ is a $2\times 2$ real-valued matrix, and is a diffusion term, modelling the variability of the process around the mean $\mu_k$;
- $W_t$ is a bivariate standard Brownian motion, modelling the stochasticity of the individual's velocity process;
- $v_{\tau_1}$ is the initial condition of the SDE: the individual's velocity at time $\tau_1$.

The solution to Eq. (1) is a well known continuous time stochastic process, the Ornstein Uhlenbeck Process (OUP) [28], also known as the *mean reverting* process. This mean reverting property is controlled by the recall parameter $\Gamma_k$ and makes the OUP suitable to describe movement modes of individuals, that are often characterized by a mean velocity regime, which is reached by rapid and brutal accelerations (or decelerations, see Figure 2). The OUP satisfies the three following properties:

1. $(V_t)_{\tau_1\leq t\leq\tau_2}$ is a continuous time Markov process;

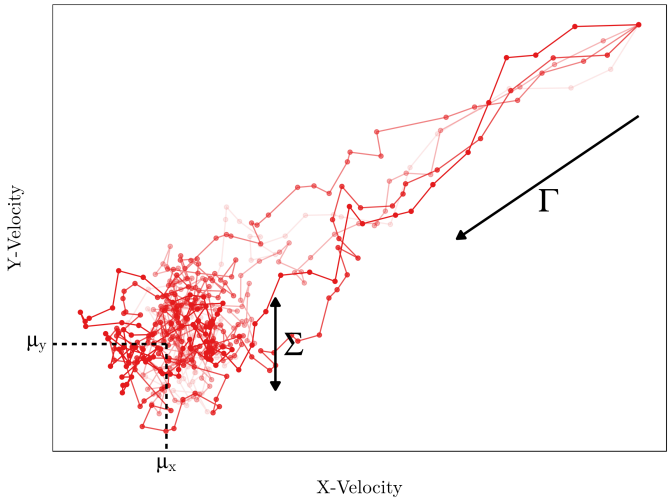


**Figure 2.** Five realizations of bivariate Ornstein Ulhenbeck Processes of parameters $\mu$, $\Gamma$ and $\Sigma$, solution of the SDE eq.(1), starting at position $v_0 = (0,0)$ (top right corner).

2. $(V_t)_{\tau_1 \leq t \leq \tau_2}$ is a Gaussian process, *i.e.* the law of the random variable $V_t | V_{\tau_1} = v_{\tau_1}$ is Gaussian, with explicit mean and covariance (given in Appendix A);

3. if both eigenvalues of $\Gamma_k$ are positive, then $(V_t)_{t \geq \tau_1}$ is an asymptotically stationary Gaussian Process, *i.e.*, the law of $V_t | V_{\tau_1} = v_{\tau_1}$ satisfies:

$$V_t | V_{\tau_1} \xrightarrow{t \to \infty} \mathcal{N}(\mu_k, \Lambda_k), \tag{2}$$

where $\Lambda_k$ is the asymptotic covariance matrix of the OUP (given in Appendix A).

As discussed in [10], the OUP offers a flexible framework to model wide range of autocorrelated velocity processes, such as highly directed movements or rotational ones. Knowing the individual's velocity process, its position process, starting at position $x_{\tau_1}$, is obtained by deterministic integration over time:

$$X_t = X_{\tau_1} + \int_{\tau_1}^{t} V_s \mathrm{d}s.$$

The resulting process $(X_t)_{\tau_1 \leq t \leq \tau_2}$ is known as an Integrated Ornstein Ulhenbeck Process (IOUP), which remains a Gaussian Process.

## 2.2 Within trajectories heterogeneity

To account for heterogeneous ways of moving during a single trajectory, we assume that an individual's trajectory in a time segment $[0,T]$ is a sequence of IOUPs. Formally, there exists a sequence of times $\tau_0 = 0 < \tau_1 < \cdots < \tau_K = T$ such that the trajectory is a sequence of $K$ successive IOUPs, each IOUP being defined over a time segment $[\tau_i, \tau_{i+1}]$. A simulated example of such a sequence is shown on Figure 3.

Therefore, a trajectory is characterized by (1) the movement modes adopted within the trajectory, (2) the time spent by the tracked individual within each movement mode and (3) the transitions from one movement mode to another. Here, we do not impose specific modelling for the last two points. Our first goal is then to estimate the different movement modes within trajectories from a given dataset.

## 2.3 Between trajectories heterogeneity

In addition to this segmentation problem, we assume that the dataset is composed of different (unknown) clusters of (entire) trajectories. A cluster is characterized by similar distributions over movement modes. In other words, two trajectories belong to the same cluster if they are composed of the same movement modes in similar proportions. This clustering problem is illustrated in Figure 4. Our second goal is to cluster together such trajectories.

To summarize, the three following problems must be solved:

1. Characterizing different movement modes present in the dataset;

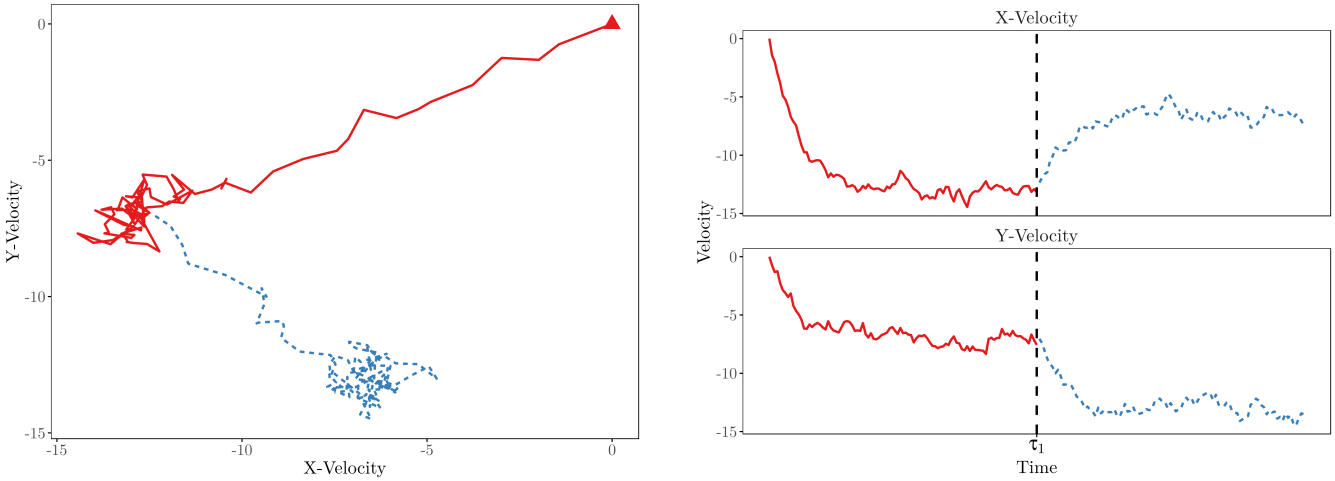


**Figure 3.** (Left) Simulation of a sequence of Ornstein Uhlenbeck Processes defining a trajectory. The red triangle marks the starting velocity. The trajectory is made of two movement modes (identified by their colors and line styles), characterized by two different sets of parameters for the velocity process from Eq. (1). (Right) Bivariate velocity processes, together with the instant $\tau_1$ at which parameters change (vertical lines).

2. For each GPS observation, estimating in which movement modes it belongs;
3. Clustering together trajectories that have the same distribution over movement modes.

They are addressed in an unsupervised and scalable approach in the following section.

## 3. Scalable two step clustering

In the previous section, we defined a realistic framework to model movement data, introducing a continuous time autocorrelated model over velocity process. The resulting framework therefore consists in a continuous time regime switching diffusion model. Estimating both parameters and movement mode assignation of data is in this case a statistical challenge, even in the case of a single trajectory [19]. For instance, [2] use thinning Poisson process and MCMC methods to perform parameter estimation from movement data, resulting in highly time demanding approach. A big challenge for estimation methods therefore remains their scalability, especially in the dual clustering context that we target here.

In order to perform scalable parameter inference and clustering of both trajectories and GPS observations (into movement modes), we adopt a pragmatic two step approach that takes advantage of the inherent properties of the OUP and are described with more details hereafter:

1. A first dual clustering is performed based on a simpler independent Gaussian mixture model, in order to estimate potential movement modes and trajectory clusters: it allows getting rid of within mode autocorrelation in the inference, and therefore eases the computations. The Gaussian hypothesis in this case is rather natural, as the OUP stationary distribution is Gaussian.
2. Among the estimated movement modes, only those meeting a temporal consistency constraint are kept. Parameters of these consistent movement modes are then estimated, and used to reassign observations that were assigned to inconsistent movement modes. It ensures that only trajectory segments for which this stationary distribution was reached are kept to estimate movement modes.

### 3.1 Step 1: simplified model based on an independence assumption

#### 3.1.1 Hierarchical model

The data set is composed of $J$ independent trajectories that we aim to cluster into $C$ groups. This number of clusters is unknown and has to be inferred from data. Each trajectory $j$ is a sequence of $n_j$ velocities $(V_i^j)_{i=1,\dots,n_j}$.

For a trajectory $j$, we denote $F_j = c$ if it belongs to the cluster $c$ ($1 \leq c \leq C$). We suppose that:

$$\mathbb{P}(F_j = c) = w_c$$

where $w_c$ is an unknown parameter to be estimated and we denote $\boldsymbol{w} = (w_1, \dots, w_C)$. In this first step, it is assumed that each cluster $c$ is a Gaussian mixture distribution having $K$ components and mixture weights $\boldsymbol{\pi}^c = (\pi_1^c, \dots, \pi_K^c)$. Again, this number

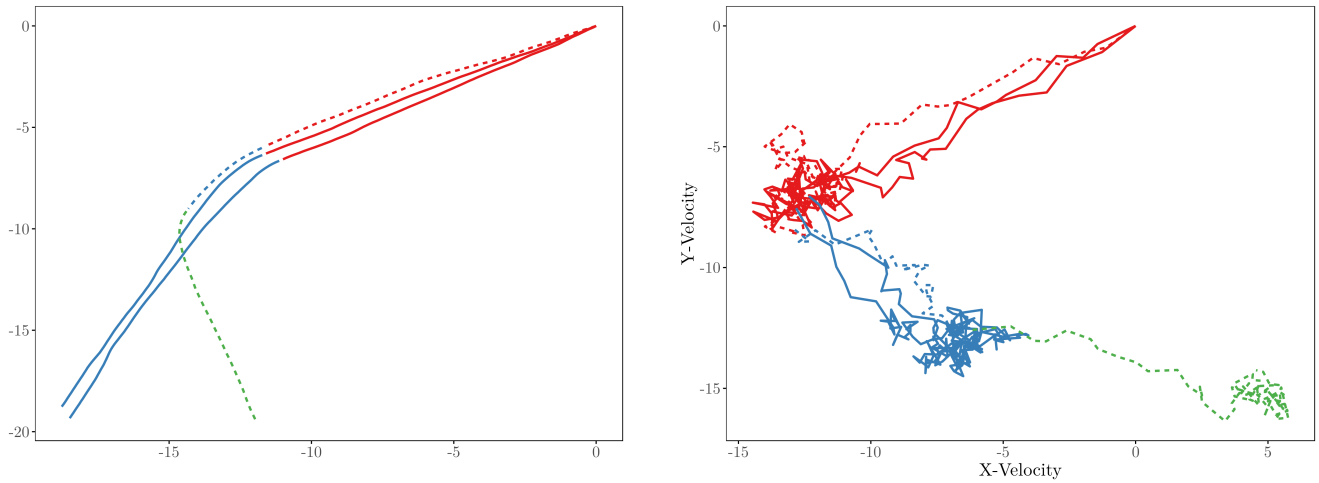


**Figure 4.** Simulation of three different trajectories with the same model as in Figure 3. The left panel shows the position processes, the right panel shows the velocity processes. All processes start from (0, 0). (Left) Two trajectories (in plain lines) are made of the same movement modes so they belong to the same cluster. The other trajectory (dashed line) has a different distribution over movement modes so it belongs to another cluster. (Right) The trajectories are plotted on the bivariate velocity space, showing the three possible movement modes. Two of them (red and blue) are shared by the two trajectory clusters.

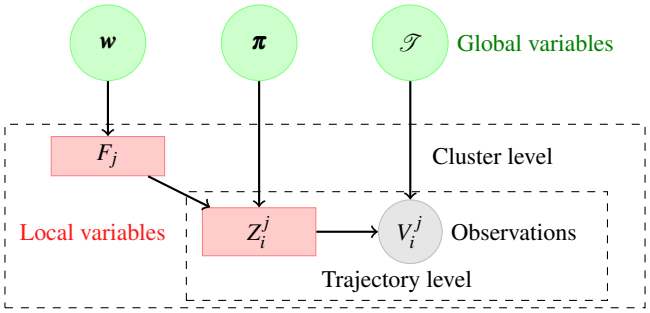


**Figure 5.** Graphical representation of the hierarchical structure of the simplified model of Section 3.1.

$K$ is unknown and has to be inferred from data. Each mixture component models a movement mode. For each $V_i^j$, we denote $Z_i^j = k$ if the velocity belongs to the $k$-th movement mode. Conditionnally to the cluster assignation of the trajectory $j$, we therefore have:

$$\mathbb{P}\left(Z_i^j = k | F_j = c\right) = \pi_k^c$$

We denote $\boldsymbol{\pi} = \left(\pi^1, \ldots, \pi^C\right)$ the set of all unknown movement modes weights to be estimated.

In this simplified model, conditionnally to their movement mode, all the velocities are mutually independent (as done in 31). This strong hypothesis eases the computation and ensures the scalability of the method as a first step, but will be relaxed in the second step. We therefore suppose that:

$$V_i^j | \left\{Z_i^j = k\right\} \sim \mathscr{N}\left(\mu_k, \Lambda_k^{-1}\right), \ 1 \leq k \leq K$$

where $\mu_k \in \mathbb{R}^2$ (resp. $\Lambda_k \in \mathscr{M}_{2\times 2}$) is the mean velocity (resp. the precision matrix) of the $k$-th movement mode. The set of unknown movement parameters $\{(\mu_k, \Lambda_k)\}_{k=1\cdots K}$ is denoted $\mathscr{T}$.

The movement parameter set $\mathscr{T}$, the trajectory clusters weights $\boldsymbol{w}$ and, for each cluster $c$, the movement mode weights $\boldsymbol{\pi}^c$ are the *global variables* (reusing the terminology used in [11]) of the model that have to be estimated. Moreover, for each trajectory $j$, the cluster label $F_j$ has to be estimated. Finally, for each observation $V_i^j$, the movement mode label $Z_i^j$ has to be estimated. Sets $\boldsymbol{F} = \left\{F_j, \ j = 1, \ldots, J\right\}$ and $\boldsymbol{Z} = \left\{Z_i^j, \ j = 1, \ldots, J, \ i = 1, \ldots, n_j\right\}$ are the *local variables* of the model to be estimated. The resulting hierarchical structure of this model is depicted on Figure 5. The next section describes the scalable inference approach to estimate both local and global variables from the data.

### 3.1.2 Bayesian estimation of the parameters using stochastic variational inference

In the first step, the inference of the local and global variables is made within a Bayesian context, considering unknown parameters as random variables. The inference therefore aims at obtaining the joint posterior distribution of local and global variables, denoted by $p(\mathscr{T}, \boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{F}, \boldsymbol{Z}|\boldsymbol{V})$.

Together with the prior specification, a classical problem in Bayesian clustering is to set the number of clusters. In the following, we adopt a Bayesian non parametric approach by considering a possibly infinite set of movement modes and clusters in the data.

First, for a given movement $k$, we consider as prior distribution for $(\mu_k, \Lambda_k)$ a Gaussian-Whishart having hyperparameters $m_0, \rho_0, \gamma_0$ and $W_0$. This distribution is a common prior for Gaussian parameters as it provides nice conjugacy properties that we'll use in our scalable purpose. The hyperparameters $m_0$ and $W_0$ are location hyperparameters for the mean and the precision matrix, while $\rho_0$ and $\gamma_0$ are concentration parameters, setting the amount of information carried by the prior.

As we do not to specifify *a priori* the number of possible movement modes, we assume that this number is infinite.

Each trajectory cluster $c$ is therefore an infinite mixture distribution whose components are Gaussian distributions. The mixture weights prior is given by a stick breaking distribution, denoted by $GEM(1, \beta)$, where $\beta$ is a concentration hyperparameter (see 23 as well as Appendix B.2 for details).

Finally, the amount of such clusters is itself infinite, and the prior distribution over cluster weights is again a stick breaking distribution $GEM(1, \alpha)$ where $\alpha$ is a concentration hyperparameter. The full prior specification is then given by the following equations:

$$\begin{cases} \left(\mu_k, \Lambda_k^{-1}\right) & \overset{i.i.d.}{\sim} & GW\left(m_0, \rho_0, \gamma_0, W_0\right) & k \geq 1, \\ \boldsymbol{w} & \sim & GEM(\alpha) & \\ \boldsymbol{\pi}^c & \overset{i.i.d.}{\sim} & GEM(\beta) & c \geq 1 \\ F_j|\boldsymbol{w} & \overset{i.i.d.}{\sim} & \boldsymbol{w} & 1 \leq j \leq J \\ Z_i^j|\boldsymbol{\pi}, F_j & \overset{i.i.d.}{\sim} & \boldsymbol{\pi}^{F_j} & 1 \leq j \leq J,\ 1 \leq i \leq n_j \end{cases}$$

The model depicted in Figure 5 together with these priors is known as a Hierarchical Dirichlet Process [27].

In this Bayesian context, the expression of the posterior density $p(\mathscr{T}, \boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{F}, \boldsymbol{Z}|\boldsymbol{V})$ has no known analytical form. A classical way to obtain Bayesian estimators of quantities of interest would be alternatively to obtain samples from this posterior distribution. This can be done using MCMC methods, that can be implemented for Dirichlet processes using a Gibbs sampler [17], and were used in [31] for the inference in their discrete-space framework. However, it is known that the Gibbs sampler does not scale properly [3], and therefore could not be used in the context of this paper.

A more scalable alternative to this simulation-based approach is the use of variational inference (see [3] for a recent review). Variational methods reduce the inference to an optimization problem by minimizing a divergence (typically the Kullback-Leibler divergence [13]) between the target posterior distribution and the members of a simpler family of distributions, the *variational* family. An appropriate and common variational family $\mathscr{Q}$ is the one which satisfies the mean field assumption, *i.e.* the set of $q$ distributions that can be fully factorized. Our variational family is therefore such that:

$$q(\mathscr{T}, \boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{F}, \boldsymbol{Z}) = q(\boldsymbol{w}) \prod_c q(\boldsymbol{\pi}^c) \prod_k q(\mu_k, \Lambda_k) \prod_j q(F_j) \prod_{i,j} q(Z_i^j). \tag{3}$$

The mean field inference problem reduces to find $q^*$ such that:

$$q^* = \underset{q \in \mathscr{Q}}{\operatorname{argmax}} \left\{ \mathbb{E}_q\left[\log p(\mathscr{T}, \boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{F}, \boldsymbol{Z}, \boldsymbol{V})\right] - \mathbb{E}_q\left[\log q(\mathscr{T}, \boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{F}, \boldsymbol{Z})\right] \right\}, \tag{4}$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to the p.d.f. $q$. The right hand side of (4) is known as the evidence lower bound (ELBO), and can be computed for appropriate families of distributions $\mathscr{Q}$, and thus can be maximized.

In addition to this mean field property, the variational approximation of the posterior distribution is restricted to finite sets of parameters. Formally, the posterior distributions of the infinite cluster (resp. movement mode) weights are approximated by a distribution on a finite set of weights having $C_{max}$ (resp. $K_{max}$) elements. Here, $C_{max}$ and $K_{max}$ are variational parameters given by the user. This variational approximation is known as the truncated stick breaking distribution [11] (see Appendix B for details).

A known algorithm to compute $q^*$ is the coordinate ascent variational inference (CAVI, [1]). This iterative algorithm starts from an initial guess $q^{(0)}$ for the optimal variational distribution, and successively updates each of its components by supposing the others known, and computes an expectation with respect to their distribution (Algorithm 1). For the model presented in Section 3.1 and the chosen prior distributions, all needed expectations can be computed explicitly (see Appendix B for details).

**Algorithm 1** Coordinate ascent variational inference algorithm

Denote $p = p(\mathbf{V}, \boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{F}, \boldsymbol{Z})$
Set $q^{(0)} = q_{\boldsymbol{w}}^{(0)}(\cdot) q_{\boldsymbol{\pi}}^{(0)}(\cdot) q_{\mathscr{T}}^{(0)}(\cdot) q_{\boldsymbol{F}}^{(0)}(\cdot) q_{\boldsymbol{Z}}^{(0)}$
**while** convergence is not reached **do**
    **Local variables update**
    $\log q_{\boldsymbol{Z}}^{(i+1)}(\cdot) = \mathbb{E}_{q_{\boldsymbol{w}}^{(i)} q_{\boldsymbol{\pi}}^{(i)} q_{\mathscr{T}}^{(i)} q_{\boldsymbol{F}}^{(i)}}[\log p]$
    $\log q_{\boldsymbol{F}}^{(i+1)}(\cdot) = \mathbb{E}_{q_{\boldsymbol{w}}^{(i)} q_{\boldsymbol{\pi}}^{(i)} q_{\mathscr{T}}^{(i)} q_{\boldsymbol{Z}}^{(i+1)}}[\log p]$
    **Global variables update**
    $\log q_{\boldsymbol{w}}^{(i+1)}(\cdot) = \mathbb{E}_{q_{\boldsymbol{\pi}}^{(i)} q_{\mathscr{T}}^{(i)} q_{\boldsymbol{F}}^{(i+1)} q_{\boldsymbol{Z}}^{(i+1)}}[\log p]$
    $\log q_{\boldsymbol{\pi}}^{(i+1)}(\cdot) = \mathbb{E}_{q_{\boldsymbol{w}}^{(i+1)} q_{\mathscr{T}}^{(i)} q_{\boldsymbol{F}}^{(i+1)} q_{\boldsymbol{Z}}^{(i+1)}}[\log p]$
    $\log q_{\mathscr{T}}^{(i+1)}(\cdot) = \mathbb{E}_{q_{\boldsymbol{w}}^{(i+1)} q_{\pi}^{(i+1)} q_{\boldsymbol{F}}^{(i+1)} q_{\boldsymbol{Z}}^{(i+1)}}[\log p]$
At last iteration $M$, set $q^* = q_{\boldsymbol{w}}^{(M)}(\cdot) q_{\boldsymbol{\pi}}^{(M)}(\cdot) q_{\mathscr{T}}^{(M)}(\cdot) q_{\boldsymbol{F}}^{(M)}(\cdot) q_{\boldsymbol{Z}}^{(M)}$

As described in [11], this algorithm can be seen as a gradient ascent, and is therefore suitable for stochastic approximations, resulting in the stochastic variational inference (SVI).

The estimation algorithm therefore reduces to:

1. sample uniformly a batch from the data set,
2. compute the expectations (as given in Appendix B) using only this batch,
3. update variational distributions using these expectations, as described in [11].

This procedure can be performed *online*, therefore with no need of storing the data [30]. Performances of these stochastic methods are widely discussed in [30] and [11].

It is worth noting here that all the needed computations can naturally be distributed (thanks to the independance simplification), as they are essentially a sum over simple operations involving single observations. Therefore, the SVI algorithm proposed here for the simplified model is widely scalable, unlike Gibbs sampling procedures.

As the optimization is done in a high dimensional space, and the algorithm only guarantees convergence to a local optimum, it is crucial to initialize the algorithm from different starting points, to ensure a good exploration of the space[1]. Again, different runs of the algorithm can be distributed.

### 3.2 Step 2: Estimation of OUP parameters from clustering outputs

The previous section described the first step of our two-step approach to perform trajectory clustering and inference of movement modes. In this section, we define how the parameters of the movement modes, defined by Ornstein Uhlenbeck processes, are re-estimated from this first step.

This first step gives as an output the set of optimal variational distributions, namely:

- $q_{\mathscr{T}}^*$, the posterior distribution of $\{(\mu_k, \Lambda_k)\}_{k=1,\dots K_{max}}$, the parameters (under the Gaussian mixture assumption) of the $K_{max}$ possible movement modes in the data set;
- $q_{\boldsymbol{w}}^*$, the posterior distribution of the weights of the $C_{max}$ possible trajectory clusters in the data set;
- $q_{\boldsymbol{\pi}}^* = \{q_{\boldsymbol{\pi}^c}^*,\ c = 1, \dots C_{max}\}$, the posterior distribution of the weights of the $K_{max}$ possible movement modes in each trajectory cluster $c$ in the data set;
- $q_{\boldsymbol{F}}^* = \left\{q_{F_j}^*,\ j = 1, \dots J\right\}$, for each trajectory $j$, the posterior probability of being in each possible trajectory cluster;
- $q_{\boldsymbol{Z}}^* = \left\{q_{Z_i^j}^*,\ j = 1, \dots J, i = 1, \dots n_j\right\}$, for each observation $i$ of trajectory $j$, the posterior probability of being in each possible movement mode.

[1] In this purpose also, it should be pointed out that any stochastic approach is better than the fully deterministic CAVI algorithm, as discussed in [12], for instance.

From the last two distributions, one can get estimators of clusters and movement modes present in the data. A classical estimator would be the maximum a posteriori (MAP), *i.e.* the cluster (resp. movement mode) label giving the maximum weight of the posterior multinomial distribution $q^*_{F_j}$ (resp. $q^*_{Z^j_i}$) .

In order to estimate the OUP parameters of the $k$-th movement mode, as defined in Section 2, a filter is applied on movement mode sequences based on their temporal consistency. Formally, for a trajectory $j$, let us consider a sequence of $m$ successively recorded velocities $V^j_{i_1},\dots,V^j_{i_m}$, that were all estimated to belong to a same movement mode $k$. This sequence is said to be temporally consistent if its length $t_{i_m}-t_{i_1}$ is larger than $\delta$, a user chosen parameter, representing the minimal time lag for a movement mode. All consistent sequences estimated in a same movement mode are considered as independent realizations of OUP with common parameters. From the Markov and Gaussian properties of the OUP, the likelihood related to this data set can be easily maximized to obtain estimates of parameters $\mu_k$, $\Gamma_k$ and $\Sigma_k$ (see Appendix A). If an estimated movement mode $k$ (*i.e.* a movement mode containing at least one observation) has no consistent sequence, this movement mode is considered as inconsistent. Finally, to refine the movement modes classification, observation sequences belonging to inconsistent movement modes are reallocated to the consistent movement mode whose parameters maximize their likelihood.

This second step therefore depends on a parameter $\delta$, whose value is surely data dependent, but has a clear and easy interpretation. The resulting consistent movement mode concept allows one to (1) have a good estimation of OUP parameters within a movement mode (as a consistent sequence will often be related to a large amount of points) and (2) filter out "noise" movement modes gathering few observations in a temporally inconsistent manner.

### 3.3 Algorithmic complexity

The goal of this two-step inference scheme is to make the approach tractable while still using fine-grained continuous time models inside movement modes.

Time complexity for our first step is at worse $O(J\times t_{max}\times K_{max}+J\times C_{max})$ where $t_{max}$ maximal number of time step for a trajectory in the data set. It should be pointed out that our approach relies on a variational approximation of the stick breaking distribution, that imposes to set two parameters $K_{max}$ and $C_{max}$, as the maximal number of movement modes and trajectory clusters present in the data. These numbers should increase with $n$. It is worth noting that for our prior specification, a Dirichlet process of parameter $\alpha$, the expected number of classes in a $n$ data set is $\alpha\log n$ [9, Proposition 4.8]. The time complexity for this first step is hence $O(\beta n\log n+\alpha J\log J)$ where $n=\sum_j n_j$ is the total number of points in the data set.

Then, estimating OU parameters for a given movement mode in the second step of our process is linear in the number of observations assigned to this movement mode. Since this has to be done for all movement modes, complexity for this second step is linear in the total number of observations in the dataset.

Overall, computational complexity of the inference step is then quasilinear in $n$ and, as discussed above, parts of the computations involved can be distributed. We show in Section 4 that these two properties allow tackling large scale datasets of GPS data in reasonable time.

## 4. Experiments

### 4.1 Experiments on simulated data

To ensure that the inference proposed in Section 3 suits the model defined in Section 2, a numerical experiment is performed on simulated data.

**Simulation set up.** A data set of 40 trajectories containing overall 8,000 observations is simulated, according to a model with 2 trajectory clusters, the two clusters being composed of respectively 2 and 3 movement modes.

Within each movement mode, velocities are drawn from an OUP whose parameters are movement mode dependent. All trajectories start at velocity (0; 0), and spend 50% of the whole time in the first movement mode before switching to a second movement mode. Trajectories of the first cluster spend then the remaining time in the second movement mode. The trajectories in the second cluster spend 12.5% in the second movement mode before switching to the third movement mode for the remaining time (37.5% of the overall time). Simulated data are illustrated in Figure 6.

**Movement mode estimation: results after the first step.** A first clustering is performed using the variational inference approach within a hierarchical Dirichlet process, as described in Section 3.1. The variational approximations of the stick breaking distributions are made with truncated breaking distribution with $C_{max}=5$ and $K_{max}=12$. Optimization is made independently from 200 starting points chosen randomly, performing 1 000 iterations of the SVI, with a decreasing gradient step along iterations. The best estimate is chosen as the final point maximizing the ELBO defined in Eq. (4). The estimated movement mode for a velocity $V^j_i$ is computed as the maximum a posteriori of the weights distribution $q^*_{Z^j_i}$. It results that among the $K_{max}=12$ possible movement modes, only 5 of them contain at least one observation. Overall, 88.3% of the points

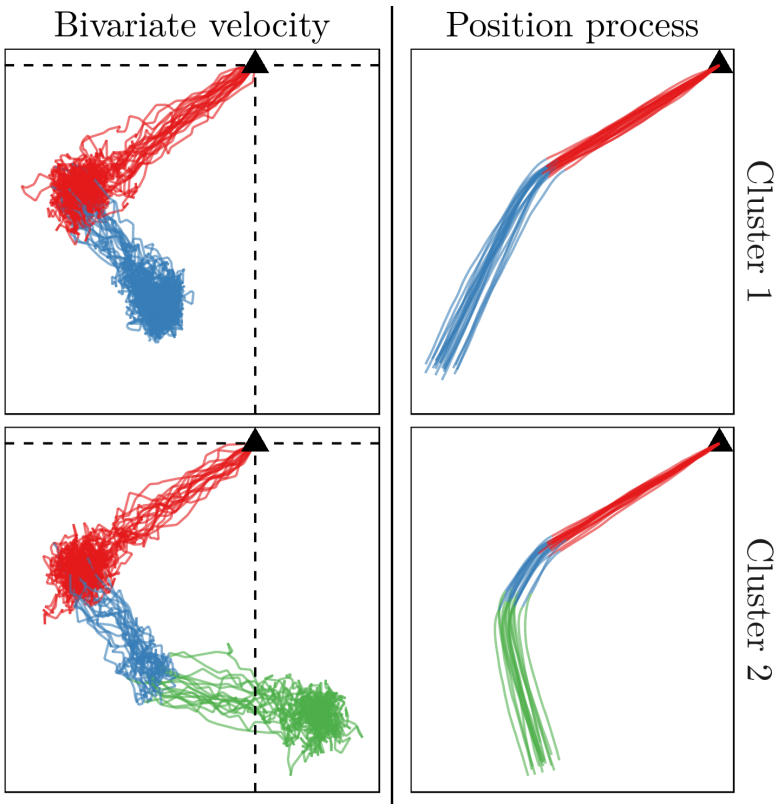


**Figure 6.** Simulated trajectory data. Three movement modes are shared by two clusters. Each cluster corresponds to two different routes. The black triangle denotes the beginning of each trajectory. The first two movement modes are present in both clusters.

are attributed to the correct movement mode (Table 1). However, two extra movement modes are estimated, corresponding to the transition phases towards the light blue movement mode and leaving from it (see Figure 7).

It is worth noting here that this first phase is a continuous space version of the work of [31]. This first phase of the presented approach, however, already improves on this baseline since it does not require to decide on a space quantization grid.

**Movement mode estimation: results after the second step.** Including the second step described in Section 3.2, movement modes for inconsistent sequences are re-estimated. Using $k$-consistent sequences, OUP parameters are estimated and used to reallocate inconsistent ones. This results in a good reallocation of the problematic transitory phases having now an overall 96.1% good classification rate, as shown in Figure 7 and Table 1.

One can see here that coming back to the OU property enables correct classification of transitory phases of the movement. This is due to the recall force in the OU that enables attaching any observed velocity that is not close to the mean velocity in the movement mode as soon as there is a trend for velocities to move toward that asymptotic mean.

**Trajectory clusters estimation.** As expected on such a simple example, the trajectory cluster assignation is 100% right. It is worth noting here that it would not be the case on trickier examples (not shown here), for instance, when the clusters are distinguished by the order of movement modes sequence, which cannot be captured by the HDP for Gaussian mixtures used here.

**Table 1.** Contingency table (counting the number of points) between true movement modes (in row) and estimated movement modes (in columns, the label assignment was made *a posteriori*), after the first step.

| True MM | Est. MM: Step 1 | | | | | Est. MM: Step 2 | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 |
| 1 | 3406 | 0 | 0 | 634 | 0 | 4026 | 12 | 2 |
| 2 | 203 | 2506 | 0 | 0 | 31 | 213 | 2517 | 10 |
| 3 | 0 | 74 | 1186 | 0 | 0 | 0 | 74 | 1186 |

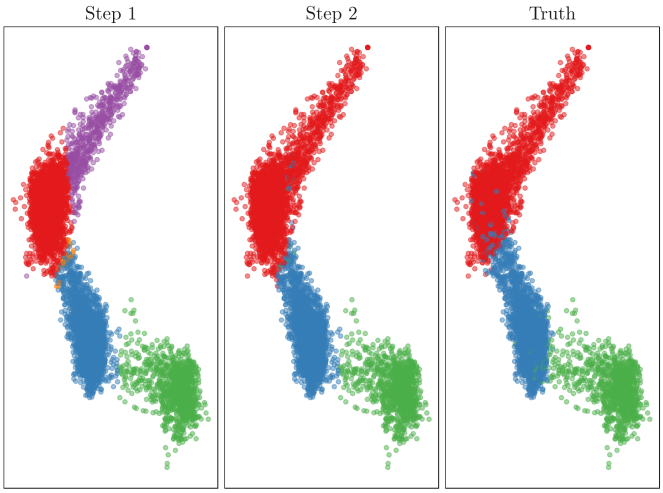


**Figure 7.** Estimated movement modes (in the bivariate velocity space) after one step (left) and two steps (middle). The ground truth is shown on the right. After one step, 5 movement modes are present in data, after step 2, only three are remaining. The related contingency table for good labelling is shown on Table 1

## 4.2 Experiments on marine traffic data

**Dataset.** We now validate our clustering approach analysing real data. The considered dataset records 6 months of AIS data of vessels steaming in the area of the Ushant traffic separation scheme (in Brittany, West of France). This is an area with one of the highest maritime traffic density in the world, with a clear separation scheme of two navigation lanes. Different kinds of vessels are sailing in the area, from cargos and tankers with high velocity and straight routes to sailboats or fishing vessels with lower speed and different sailing directions. As such, the area is highly monitored to avoid collision or grounding, and a better analysis and understanding of the different ship behaviors is of prime importance.

The whole trajectory dataset is shown on Figure 1 and is available online[2]. It consists in 18,603 trajectories, gathering at all more than 7 millions GPS observations. Only trajectories having more than 30 points are kept, time lag between two consecutive observations ranges between 5 seconds and 15 hours, with 95% of time lags below 3 minutes.

**Two step movement modes estimation.** Model fit is performed considering a maximum of 90 movement modes and 30 trajectory clusters, using non-informative priors. 200 runs of 50,000 iterations for SVI are performed independently, each run taking approximately 8 hours with SVI computations parallelized on a 8-core machine. The run leading to the best ELBO is kept as the estimate.

The SVI leads to the estimation of 81 different movement modes containing (in the sense of maximum a posteriori probability) at least one observation, with 50 movement modes containing 95% of observations. The shapes (mean and covariance) of these movement modes are shown on Figure 8. One can note that a lot of movement modes lead to a same steaming direction, but at different speeds. In this context, choosing more informative prior or adding a penalty on movement modes covariance parameters in the ELBO optimization could lead to a model with less movement modes.

**Trajectory clusters estimation.** In addition to the movement modes estimated above, 29 (non-empty) trajectory clusters are identified. Three emblematic clusters are represented on Figure 9, plotting all their trajectories with the same colour code as in Figure 8. The first two subfigures present the largest clusters (gathering respectively 18% and 12% of trajectories) and are composed of two preponderant movement modes: they represent the two main marine roads of the Ushant traffic separation scheme, the first one entering the English Channel, the second one exiting it. The third cluster gathers 5% of trajectories, and mostly puts together trips that are performed at low speed (below 5 knots), involving both sailing boats and fishing vessels.

**Comparison with other clustering method.** Our algorithm can be seen as an improved (continuous-space) version of the method presented in [31]. As such, in order to validate the results of the clustering, we rather rely on a $k$-means algorithm using Dynamic Time Warping (DTW, 22) as a distance between trajectories considering them as sequences of 2D velocity vectors (using DTW Barycenter Averaging method from [20] for centroid estimation). It is worth mentionning here that this method as a complexity at worse $O(JC_{max} \times t_{max}^2)$ (where $t_{max}$ is the maximal number of points for a trajectory in the data set). Thus, running this algorithm on the full dataset was too costly and we subsampled one tenth of the original dataset to feed the clustering. We used the implementation from the python library `tslearn` [26]. We also run our algorithm on the same

[2] `https://github.com/rtavenar/ushant_ais`

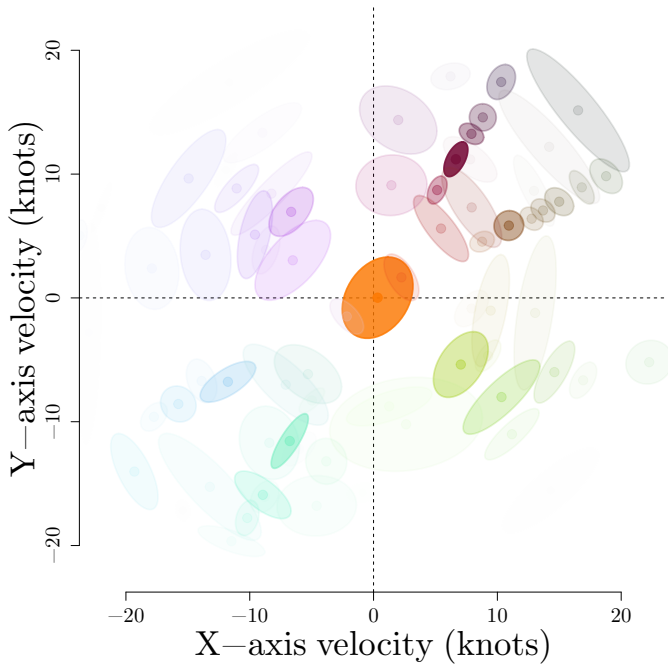


**Figure 8.** Movement modes estimated distributions (on the velocity space). Each ellipsoid represents a movement mode distribution, the dot being the expected mean, and the contour the 50% centered quantile. The transparency shows the estimated relative weight. The dotted lines show the 0 values.

dataset and we obtain a Normalized Mutual Information between the clusterings equal to 0.63. The resulting five main clusters are presented in Figure 2 in the appendix, together with a confusion matrix between those five clusters. As no ground truth is available, we cannot determine which solution fits better but we note by a visual inspection of the clusters that the DTW-based clustering fails at differenciating ships that follow the same traffic lane (*i.e.* identifying sub-routes in the main traffic lanes). One should also note that, by assigning movement modes to the observations, our methods provides extra explanation for cluster assignments, compared to $k$-means that solely relies on inertia minimization.

**Interest of the post-filtering step.** In the present case, the main advantage of the second step, which consists in a re-estimate of inconsistent sequences, is to avoid the estimation of transitory movement modes. Figure 10 shows a trajectory belonging to the central cluster of Figure 9. The transitory phase of the vessel, corresponding to its change of direction, is then no longer estimated as a dedicated movement mode, but rather belongs to one of the main movement modes.

## 5. Conclusion

In this work, we have defined a generic framework for the clustering of large trajectory data sets. This framework is specified in continuous time and space, which makes its formulation insensitive to GPS sampling. Inference is done in a two-step scalable approach using stochastic variational inference for a conjugate hierarchical Dirichlet process. This framework is easy to distribute as demonstrated experimentally, and has a quasilinear complexity in the number of observations. We provide a dataset of several millions of observations in the AIS context to both validate our model and allow future competitive methods to compare on a real-world large-scale trajectory dataset. Note however that the proposed framework is generic and suits, to our opinion, to a wide range of trajectory data sets.

Future works include extension of our framework to model sojourn time in movement modes and/or transitions between modes. This would make anomaly detection possible, both at the observation level (abnormal observation given the estimated movement mode) and at the trajectory level (abnormal trajectory as an unlikely sequence of movement modes). Such an extended model could also be used as a fully generative model.

## Acknowledgements

This work has been supported by DGA through the ANR/Astrid SESAME project (ref: ANR-16-ASTR-0026). Authors would like to thank CLS (Collecte Localisation Satellites) and Erwan Guegueniat for providing the raw data that allowed building the AIS dataset used in this paper.

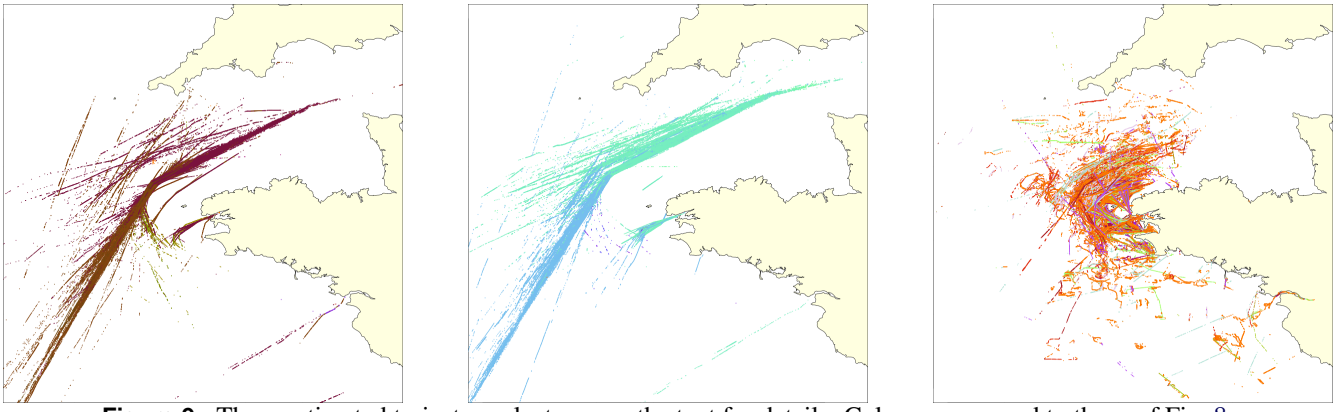

**Figure 9.** Three estimated trajectory clusters, see the text for details. Colors correspond to those of Fig. 8.

## References

[1] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[2] Paul Blackwell, Mu Niu, Mark S Lambert, and Scott D LaPoint. Exact bayesian inference for animal movement in continuous time. *Methods in Ecology and Evolution*, 7(2):184–195, 2016.

[3] David Blei, Alp Kucukelbir, and Jon McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

[4] Federico Clazzer, Andrea Munari, Matteo Berioli, and Francisco Lazaro Blasco. On the characterization of ais traffic at the satellite. In *OCEANS 2014-TAIPEI*, pages 1–9. IEEE, 2014.

[5] Urška Demšar, Kevin Buchin, Francesca Cagnacci, Kamran Safi, Bettina Speckmann, Nico Van de Weghe, Daniel Weiskopf, and Robert Weibel. Analysis and visualisation of movement: an interdisciplinary review. *Movement ecology*, 3(1):5, 2015.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96(34), pages 226–231, 1996.

[7] Ronan Fablet, Nicolas Bellec, Laetitia Chapel, Chloé Friguet, René Garello, Pierre Gloaguen, Guillaume Hajduch, Sébastien Lefèvre, François Merciol, Pascal Morillon, Christine Morin, Matthieu Simonin, Romain Tavenard, Cédric Tedeschi, and Rodolphe Vadaine. Next Step for Big Data Infrastructure and Analytics for the Surveillance of the Maritime Traffic from AIS & Sentinel Satellite Data Streams. In *2017 Conference on Big Data from Space (BiDS'17)*, pages 371–374, November 2017. Poster.

[8] Bernard Garnier and Aldo Napoli. Exploiting the potential of the future "maritime big data". In *Maritime Knowledge Discovery and Anomaly Detection Workshop*, 2016.

[9] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[10] Eliezer Gurarie, Christen H Fleming, William F Fagan, Kristin L Laidre, Jesús Hernández-Pliego, and Otso Ovaskainen. Correlated velocity models as a fundamental unit of animal movement: synthesis and applications. *Movement ecology*, 5(1):13, 2017.

[11] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[12] Holger H Hoos and Thomas Stützle. *Stochastic local search: Foundations and applications*. Elsevier, 2004.

[13] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[14] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.

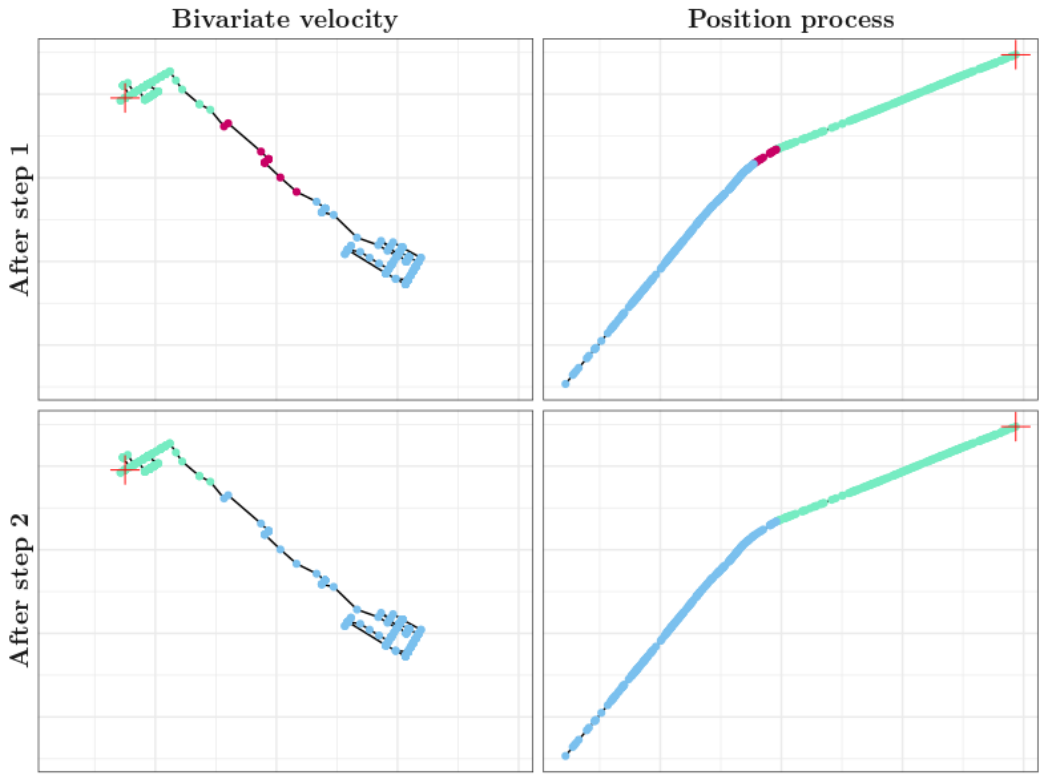


**Figure 10.** Example of result of the proposed second step in movement mode estimation. Both the velocity (left) and position (right) processes are shown (units are of no importance here). The red cross shows the first point of both processes. Colors corresponds to different movement modes. The purple phase in the middle of the trajectory is inconsistent (as defined in the text). It is a transitory phase in the vessel's velocity process. After the second step, it is assigned as the beginning of the second movement mode.

[15] Xiaolei Li, Jiawei Han, Jae-Gil Lee, and Hector Gonzalez. Traffic density-based discovery of hot routes in road networks. In *International Symposium on Spatial and Temporal Databases*, pages 441–459. Springer, 2007.

[16] Ran Nathan. An emerging movement ecology paradigm. *Proceedings of the National Academy of Sciences*, 105(49):19050–19051, 2008.

[17] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[18] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

[19] Toby Patterson, Alison Parton, Roland Langrock, Paul Blackwell, Len Thomas, and Ruth King. Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges. *Advances in Statistical Analysis*, 101(4):399–438, 2017.

[20] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. 44(3):678 – 693, 2011.

[21] Salvatore Rinzivillo, Dino Pedreschi, Mirco Nanni, Fosca Giannotti, Natalia Andrienko, and Gennady Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3-4):225–239, 2008.

[22] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[23] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[24] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[25] Cynthia Sung, Dan Feldman, and Daniela Rus. Trajectory clustering for motion prediction. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 1547–1552. IEEE, 2012.

[26] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. tslearn: A machine learning toolkit dedicated to time-series data, 2017. `https://github.com/rtavenar/tslearn`.

[27] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

[29] Michele Vespe and Fabio Mazzarella, editors. *Maritime Knowledge Discovery and Anomaly Detection Workshop Proceedings*, July 2016.

[30] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of AISTATS*, pages 752–760, 2011.

[31] Xiaogang Wang, Keng Teck Ma, Gee-Wah Ng, and W Eric L Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International journal of computer vision*, 95(3):287–312, 2011.

[32] Tingting Yao, Zhiyong Wang, Zhao Xie, Jun Gao, and David Dagan Feng. Learning universal multiview dictionary for human action recognition. *Pattern Recognition*, 64:236–244, 2017.

[33] Shun-Zheng Yu. Hidden semi-markov models. *Artificial intelligence*, 174(2):215–243, 2010.

[34] Guan Yuan, Penghui Sun, Jie Zhao, Daxing Li, and Canwei Wang. A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review*, 47(1):123–144, 2017.

[35] Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

[36] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.

# Appendix

## 1. Characteristics of the Ornstein Uhlenbeck process

Notations:

- $I_d$ is the $d \times d$ identity matrix;
- For a matrix $A$, $A^T$ is the transposed matrix of A;
- For two $2 \times 2$ matrices $A$ and $B$, $A \oplus B$ denotes the Kronecker sum of $A$ and $B$, defined by : $A \oplus B = A \otimes I_2 + I_2 \otimes B$, where $\otimes$ denotes the Kronecker product. Note that in our case, $A \oplus B$ is a $4 \times 4$ matrix;
- For a matrix $A = \begin{pmatrix} a_1 & a_3 \\ a_2 & a_4 \end{pmatrix}$, $\mathbf{vec}(A)$ denotes the vectorization of $A$, *i.e.* the vector $\mathbf{vec}(A) = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}$.

**Transition density** Suppose that the process $(V_t)_{\tau_1 \leq t \leq \tau_2}$ is solution to the equation (1), then,

$$\forall \tau_1 \leq t \leq \tau_2, \;\; V_t | V_{\tau_1} = v_{\tau_1} \sim \mathcal{N}(m_{k,\Delta}^{v_{\tau_1}}, \Lambda_{k,\Delta})$$

where $\Delta = t - \tau_1$ and

$$m_{k,\Delta}^{v_{\tau_1}} = \mathrm{e}^{-\Gamma_k \Delta} v_{\tau_1} + (I_2 - \mathrm{e}^{-\Gamma_k \Delta}) \mu_k$$
$$\mathbf{vec}\left(\Lambda_{k,\Delta}\right) = (\Gamma_k \oplus \Gamma_k)^{-1} \left(I_4 - \mathrm{e}^{-(\Gamma_k \oplus \Gamma_k)\Delta}\right) \mathbf{vec}\left(\Sigma_k \Sigma_k^T\right)$$

**Likelihood** Let $\mathbf{v}$ be a sequence of observations $v_{t_1} = v_{\tau_1}, v_{t_2}, \ldots, v_{t_n} = v_{tau_2}$, discrete time observations of a OUP starting at $V_{\tau_1}$ and such that $V_{\tau_1}$ is a random variable with p.d.f. (possibly depending on $\mu_k, \Gamma_k$ and $\Sigma_k$, $p_{\tau_1}(\cdot)$. Then the likelihood of the observed sequence is given by:

$$L(\mu_k, \Gamma_k, \Sigma_k; \mathbf{v}) = p_{\tau_1}(v_{\tau_1}) \prod_{i=1}^{n-1} \phi_{k,\Delta_i}(v_{t_{i+1}}),$$

where $\Delta_i = t_{i+1} - t_i$ and $\phi_{k,\Delta_i}(v_{t_{i+1}})$ is the p.d.f. of a Gaussian distribution with mean $m_{k,\Delta_i}^{v_{t_i}}$ and covariance $\Lambda_{k,\Delta_i}$

**Stationary distribution** From above, one can see that if both $e^{-\Gamma_k \Delta}$ and $e^{-(\Gamma_k \oplus \Gamma_k)\Delta}$ vanishes to 0 when $\Delta$ increase, then the process $(V_t)_{t \geq \tau_1}$ is asymptotically stationary, and

$$V_t | V_{\tau_1} = v_{\tau_1} \xrightarrow{distrib.} \mathcal{N}(\mu_k, \Lambda_k)$$

where

$$\mathbf{vec}(\Lambda_k) = (\Gamma_k \oplus \Gamma_k)^{-1} \mathbf{vec}\left(\Sigma_k \Sigma_k^T\right)$$

The vanishing condition is satisfied when both eigenvalues of $\Gamma_k$ are positive.

## 2. Computation for variational inference

### B.1 Prior specification

Overall, the variables prior distribution are defined as follows:

$$\begin{cases} \left(\mu_k, \Lambda_k^{-1}\right) & \overset{i.i.d.}{\sim} & GW(m_0, \rho_0, \gamma_0, W_0) & k \geq 1, \\ \boldsymbol{w} & \sim & GEM(\alpha) & \\ \boldsymbol{\pi}^c & \overset{i.i.d.}{\sim} & GEM(\beta) & c \geq 1 \\ F_j | \boldsymbol{w} & \overset{i.i.d.}{\sim} & Mult(\boldsymbol{w}) & 1 \leq j \leq J \\ Z_i^j | \boldsymbol{\pi}, F_j & \overset{i.i.d.}{\sim} & Mult(\boldsymbol{\pi}^{F_j}) & 1 \leq j \leq J,\ 1 \leq i \leq n_j \end{cases}$$

where $GW(\cdot)$ denotes the Gaussian Wishart distribution, depending on 4 hyperparameters, $GEM(\cdot)$ denotes the stick breaking distribution depending on 1 hyperparameter, and $Mult(\cdot)$ denotes the multinomial distribution (with weights as parameters). The choice of these distributions is convenient as it leads to a conjugate framework for conditional distributions, *i.e.*, the conditional distribution of a hidden variable given the observations and the other hidden variables.

### B.2 About the stick breaking construction

To build a prior for an infinite sequence of weights summing to one, we use the stick breaking construction. The topic weights $w_c$ are build using an hyperparameter $\alpha$:

$$\begin{aligned} &\nu_c \overset{i.i.d}{\sim} \mathscr{B}eta(1, \alpha) && c = 1, 2, \ldots \\ &w_c = \nu_c \prod_{i=1}^{c-1} (1 - \nu_i) && \textbf{Init: } w_1 = \nu_1 \end{aligned}$$

Weights $\pi_k^c$ are sampled similarlrly, with an hyperparameter $\beta$:

$$\begin{aligned} &\eta_k^c \overset{i.i.d}{\sim} \mathscr{B}eta(1, \beta) && k = 1, 2, \ldots \\ &\pi_k^c = \eta_k^c \prod_{i=1}^{k-1} (1 - \eta_i^c) && \textbf{Init: } \pi_1^c = \eta_1^c \end{aligned}$$

In the following, we'd rather deal with the sequence of $\boldsymbol{\nu}$ and $\boldsymbol{\eta^c}$, as there are i.i.d. samples. We have, for each $c$ and each $k$:

$$\begin{aligned} p(\nu_c) &= (1 - \nu_c)^{\alpha - 1} \\ p(\eta_k^c) &= (1 - \eta_k^c)^{\beta - 1} \end{aligned}$$

## B.3 Likelihood

Keeping the notation of the main text, the likelihood of an observation $\mathbf{v}_i^j$ is given by:

$$p(\mathbf{v_i^j}|Z_i^j,\mu,\Lambda) = \prod_{k=1}^{\infty} \varphi(\mathbf{v_i^j}|\mu_k,\Lambda_k^{-1})^{z_{i,k}^j},$$

which results, for the complete data set $\mathbf{V}$

$$p(\mathbf{V}|Z,\mu,\Lambda) = \prod_{j=1}^{M}\prod_{i=1}^{n_j}\prod_{k=1}^{\infty} \varphi(\mathbf{v_i^j}|\mu_k,\Lambda_k^{-1})^{z_{i,k}^j}$$

The distribution of latent allocation vectors (or local variables) $F_j$ and $Z_j^i$ is given by an (infinite) multinomial distributions depending on weights $\boldsymbol{w}$ and $\boldsymbol{\pi}$:

$$p(Z_i^j|F_j,\boldsymbol{\pi}) = \prod_{c=1}^{\infty}\left(\prod_{k=1}^{\infty}(\pi_k^c)^{z_{i,k}^j}\right)^{f_{j,c}} \tag{5}$$

$$p(F_j|\boldsymbol{w}) = \prod_{c=1}^{\infty} w_c^{f_{j,c}} \tag{6}$$

## B.4 Variational approximations of the posterior distributions

Let $\mathscr{L} = \left\{\{Z_i^j\},\{F_j\},\boldsymbol{\eta},\boldsymbol{\nu},\boldsymbol{\mu},\boldsymbol{\Lambda}\right\}$ be the set of (local and global) hidden variables. For any hidden variable $U$ in $\mathscr{L}$ the corresponding $q^*(U)$ is given (Bishop) by

$$\ln q^*(U) = \mathbb{E}_{\mathscr{L}\setminus U}\left[\ln p(\mathbf{V},\mathbf{Z},\mathbf{F},\boldsymbol{\nu},\boldsymbol{\eta},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] + \text{ constant} \tag{7}$$

The complete joint distribution can be split in simpler terms:

$$p(\mathbf{V},\mathbf{Z},\mathbf{F},\boldsymbol{\nu},\boldsymbol{\eta},\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\mathbf{V}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\mathbf{Z}|\mathbf{F},\boldsymbol{\eta})p(\mathbf{F}|\boldsymbol{\nu})p(\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{\eta})p(\boldsymbol{\nu}) \tag{8}$$

### B.4.1 Optimal variational distribution for $\nu$

The variational approximation of the trajectory cluster is a truncated stick breaking distribution. It is therefore define by $C_{max}$ random variables, such that the first $C_{max}-1$ are beta distributed, and the last one equals to one almost surely. A simple computation using equations (7) and (8) shows that for the $C_{max}-1$ first terms, the variational distribution of $\nu_c$ is given by:

$$q^*(\nu_c) \sim \mathscr{B}eta\left(1+\sum_{j=1}^{J}\mathbb{E}_{q_{\boldsymbol{F}}}[\mathbf{1}_{F_j=c}], \alpha+\sum_{j=1}^{J}\mathbb{E}_{q_{\boldsymbol{F}}}[\mathbf{1}_{F_j>c}]\right),$$

where $\mathbf{1}$ denotes the classical indicator function.

### B.4.2 Optimal variational distribution for $\eta$

Similarly to the clustering weights, a truncated stick breaking distribution is used. For the cluster $c$ and the movement mode $1 \leq k \leq K_{max}-1$, we have:

$$q^*(\eta_k^c) \sim \mathscr{B}eta\left(1+\sum_{j=1}^{J}\mathbb{E}_{q_{F_j}}[f_{j,c}]\sum_{i=1}^{n_j}\mathbb{E}_{q_{Z_i^j}}[\mathbf{1}_{Z_i^j=k}], \beta+\sum_{j=1}^{J}\mathbb{E}_{q_{F_j}}[f_{j,c}]\sum_{i=1}^{n_j}\mathbb{E}_{q_{Z_i^j}}[\mathbf{1}_{Z_i^j>k}]\right)$$

where we have written $F_j = \left(f_{j,1},\ldots,f_{j,C_{max}}\right)$.

### B.4.3 Optimal variational distribution for $(\mu,\Lambda)$

This computation is done in [1]. For eack movement mode $1 \leq k \leq K_{max}$, the optimal variational distribution is a $\mathscr{GW}(m_k,\rho_k,\gamma_k,W_k)$ with:

$$\begin{aligned}
\rho_k &= \rho_0 + N_k \\
m_k &= \frac{\rho_0 m_0 + N_k\bar{\mathbf{v}}_{\mathbf{k}}}{\rho_k} \\
\gamma_k &= \gamma_0 + N_k \\
W_k &= \left(W_0^{-1} + N_k\mathbf{S}_{\mathbf{k}} + \frac{\rho_0 N_k}{\rho_0+N_k}(\bar{\mathbf{v}}_{\mathbf{k}})-m_0)(\bar{\mathbf{v}}_{\mathbf{k}})-m_0)\right)
\end{aligned}$$

where

$$N_k = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathbb{E}_{q_{Z_i^j}}[z_{i,k}^j]$$
$$\bar{\mathbf{v}}_\mathbf{k} = \frac{1}{N_k}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathbb{E}_{q_{Z_i^j}}[z_{i,k}^j]\mathbf{v}_\mathbf{i}^\mathbf{j}$$
$$\mathbf{S}_\mathbf{k} = \frac{1}{N_k}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathbb{E}_{q_{Z_i^j}}[z_{i,k}^j](\mathbf{v}_\mathbf{i}^\mathbf{j}-\bar{\mathbf{v}}_\mathbf{k})(\mathbf{v}_\mathbf{i}^\mathbf{j}-\bar{\mathbf{v}}_\mathbf{k})^T$$

#### B.4.4 Optimal variational distribution for $F_j$

The computation results in $q^*(F_j)$ being a multinomial distribution on the set $[[1, C_{max}]]$ with the $c$-th weight proportional to:

$$\mathbb{E}_{q_\mathbf{v}}[\log(w_c)] + \sum_{i=1}^{n}\sum_{k=1}^{K_{max}} \mathbb{E}_{q_{Z_i^j}}[z_{i,k}^j]\mathbb{E}_{q_\boldsymbol{\eta}}[\log \pi_k^c]$$

Here, one can see that

$$\mathbb{E}_{q_\mathbf{v}}[\log(w_c)] = \mathbb{E}_{\boldsymbol{\nu_c}}[\log(\nu_c)] + \sum_{i=1}^{c-1}\mathbb{E}_{\boldsymbol{\nu_i}}[\log(1-\nu_i)]$$

This expectation can therefore be computed using the two following properties of the beta distribution. If $X \sim \mathscr{B}eta(\alpha_1, \alpha_2)$, then

1. $(1-X) \sim \mathscr{B}eta(\alpha_2, \alpha_1)$;
2. $\mathbb{E}[\log X] = \psi(\alpha_1) - \psi(\alpha_1+\alpha_2)$, where $\psi$ is the digamma function.

Of course, the same remarks hold for $\mathbb{E}_{q_\boldsymbol{\eta}}[\log \pi_k^c]$.

#### B.4.5 Optimal variational distribution for $Z_i^j$

The computation results in $q^*(Z_i^j)$ being a multinomial distribution on the set $[[1, K_{max}]]$ with the $k$-th weight proportional to:

$$\sum_{c=1}^{C_{max}} \mathbb{E}_{q_{F_j}}[f_{j,c}]\mathbb{E}_{q_\boldsymbol{\eta}}[\log \pi_k^c] + \frac{1}{2}\mathbb{E}_{q_{\Lambda_k}}[\log(\det\Lambda_k)] - \frac{1}{2}\mathbb{E}_{q_{\mu_k,\Lambda_k}}\left[(\mathbf{v}_i^j-\mu_k)^T\Lambda_k(\mathbf{v}_i^j-\mu_k)\right] - \log 2\pi$$

## 3. Simulation design for experiments

We consider a framework with $K = 12$ movement modes, and $C = 10$ trajectory cluster. Each trajectory cluster $c$ is characterized by:

- Its initial movement mode density (the law of the first movement mode chosen during the cluster) denoted by $\boldsymbol{\pi}_0^c = (\boldsymbol{\pi}_{0,1}^c, \ldots, \boldsymbol{\pi}_{0,K}^c)$;
- Its transition densities between movement modes (the laws for the next movement mode after having been in a cluster $k$) denoted by denoted by $\boldsymbol{\pi}_0^c = (\boldsymbol{\pi}_{k,1}^c, \ldots, \boldsymbol{\pi}_{k,K}^c)$ for all $1 \leq k \leq K$;

A movement mode $k$ is characterized by:

- Its movement parameters $\mu_k, \Gamma_k, \Sigma_k$, *i.e.* the OUP parameters;
- Its sojourn time distribution, *i.e.* the distribution of time spent in this movement mode, denoted by $d_k$. In our example, $d_k$ is the p.d.f. of gamma distribution $\mathscr{G}(\Delta_k, \frac{1}{2})$ where $\Delta_k$ is a movement mode-specific parameter.

For each trajectory $1 \leq j \leq 500$, we independently repeat the following procedure:

- **Final time** Sample[3] a final time $T^j$;

[3] In our example, it was sampled uniformly in the continuous interval $[80, 500]$

- **Number of observations** Sample[4] a number of observed points $n_j$;
- **Observation times** Sample independently $\upsilon_1, \ldots, \upsilon_{n_j}$ points from a uniform $U[0, T^j]$, and sort them to set times as $t_1 := \upsilon_{(1)}, \ldots t_{n_j} = \upsilon_{(n_j)}$;
- **Trajectory cluster** Sample a trajectory cluster $c$ with known probabilities $\boldsymbol{w} = (w_1, \ldots, w_{10})$;
- **First movement mode** Sample $k_0$ with probabilities $\boldsymbol{\pi}_0^c$;
- **First duration** Set $\tau_0 = 0$; Sample $\delta_0$ from $d_{k_0}$, set $\tau_1 = \tau_0 + \delta_0$;
- **First OUP sampling** Set an initial value $V_0$. Sample a OUP $(V_t)_{\tau_0 \leq t \leq \tau_1}$ starting at $v_0$ at all observation times between $\tau_0$ and $\tau_1$ and at $\tau_1$;
- Set $i = 1$;
- **While** $\tau_i < T^j$
  - **Movement mode** Sample $k_i$ with probabilities $\boldsymbol{\pi}_{k_{i-1}}^c$;
  - **Duration** Sample $\delta_i$ from $d_{k_0}$, set $\tau_{i+1} = \tau_i + \delta_i$;
  - **OUP sampling** Sample a OUP $(V_t)_{\tau_i \leq t \leq \tau_{i+1}}$ starting at $V_{\tau_i}$ at all observation times between $\tau_i$ and $\tau_{i+1}$, and at $\tau_{i+1}$;
  - Set $i = i + 1$;

The resulting process is a continuous time hidden semi-Markov model [33] whose emission densities are autoccorelated OU processes.

## 4. Additional experimental results

Table 2 represents the clusters obtained by our algorithm and those of a DTW-based $k$-means and Table 3 compares the results. The clusters are ordered by the number of trajectories they contain. In Table 2, one may notice that DTW-based $k$-means algorithms puts in the 2 first clusters trajectories that enter the traffic scheme from the South. One also notices a correspondance between the clusters 1,4 (our) $\rightarrow$ 1,2 (DTW), 2 (our) $\rightarrow$ 3 (DTW), 3 (our) $\rightarrow$ 4 (DTW), 5 (our) $\rightarrow$ 5 (DTW).

[4] In our example, it was sampled uniformly in the discrete interval $[50, 450]$

**Table 2.** First five clusters for our and DTW-based kmeans algorithms. Blue color indicates the beginning of the trajectories, yellow color the end.

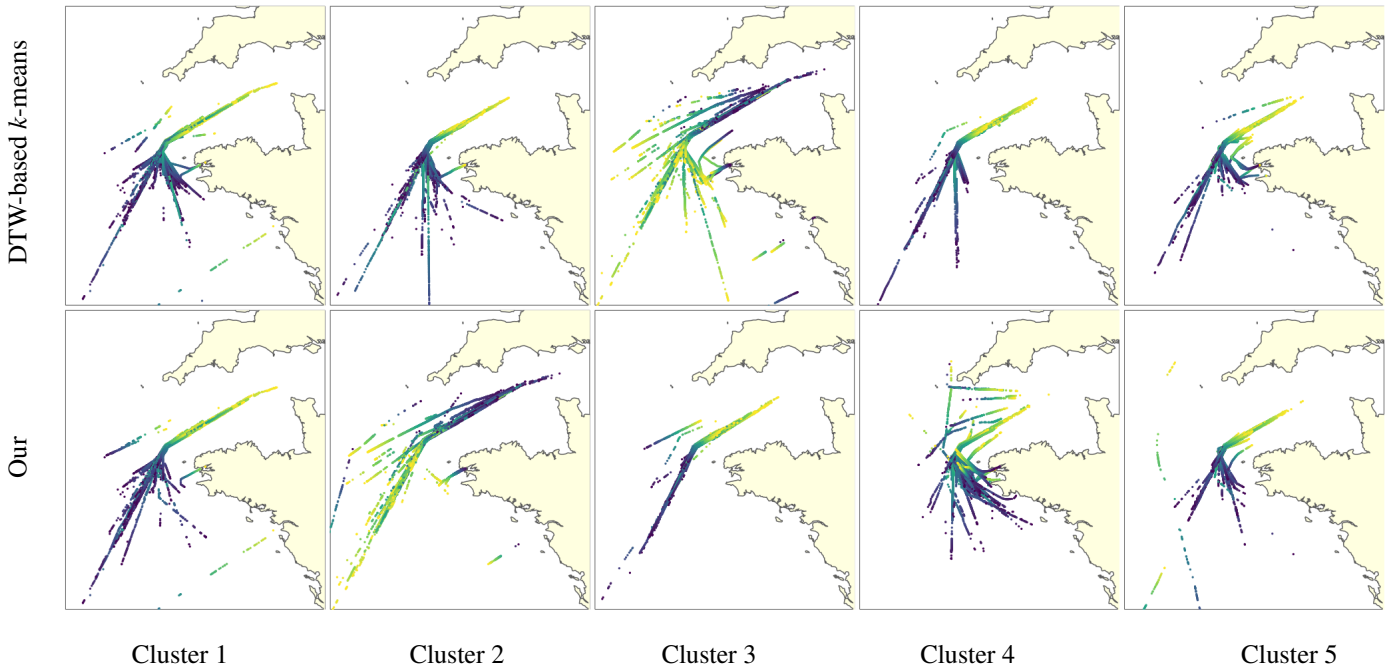


**Table 3.** Number of trajectories by cluster. The Normalized Mutual Information on those five clusters is equal to 0.68.

| Our/DTW | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 209 | 94 | 0 | 0 | 32 |
| 2 | 0 | 0 | 145 | 0 | 0 |
| 3 | 0 | 8 | 0 | 107 | 0 |
| 4 | 23 | 28 | 0 | 1 | 4 |
| 5 | 7 | 0 | 0 | 0 | 79 |