Celestial Mechanics and Dynamical Astronomy (2021) 133:51
https://doi.org/10.1007/s10569-021-10047-3

ORIGINAL ARTICLE

Applying data mining techniques to higher-dimensional Poincaré maps in the circular restricted three-body problem

Stefano Bonasera[1] · **Natasha Bosanac**[1]

Received: 28 April 2021 / Revised: 27 August 2021 / Accepted: 10 October 2021 /
Published online: 25 November 2021

Abstract
Poincaré maps are regularly used to facilitate rapid and informed trajectory design within multi-body systems. However, maps that capture a general set of spatial trajectories are often higher-dimensional and, as a result, challenging for a human to analyze. This paper addresses this challenge by employing techniques from data mining. Specifically, distributed clustering, dimension reduction and classification are used in combination to construct a data-driven approach to autonomously group higher-dimensional crossings on a Poincaré map according to the geometry of the associated trajectories generated over a short time interval. This procedure is demonstrated using a periapsis map that captures spatial trajectories at a single energy level in the Sun-Earth circular restricted three-body problem. Arcs along hyperbolic invariant manifolds associated with families of tori in the L_1 and L_2 gateways are also projected onto this clustering result to rapidly extract their fundamental geometries. Together, these examples demonstrate the potential for the presented data-driven approach to facilitate analysis of a complex solution space reflected on a higher-dimensional Poincaré map.

Keywords Circular restricted three-body problem · Higher-dimensional Poincaré maps · Clustering · Dimension reduction

1 Introduction

Poincaré maps are used in astrodynamics and celestial mechanics to analyze the complex solution space in a multi-body system. A Poincaré map displays the intersections of a set of trajectories with a surface of section, reducing continuous solutions to sequences of states and decreasing the dimensionality of the problem (Perko 1996). This technique has been used to study trajectories that admit a large number of returns to the surface of section, enabling

An earlier version of this paper was presented in January 2020 as Paper AIAA 2020-2178 at the 30th AIAA/AAS Space Flight Mechanics Meeting in Orlando, FL.

✉ Natasha Bosanac
natasha.bosanac@colorado.edu

[1] Smead Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO 80303, USA

analysis of the local neighborhood of a periodic orbit and the identification of fundamental solutions via patterns on the map (Perko 1996; Koon et al. 2011; Contopoulos 2002; Davis 2011). Poincaré maps are also commonly used in astrodynamics to examine only a small number of returns to the surface of section for trajectories that: lie on a segment of a global stable or unstable manifold of a periodic orbit or torus (Koon et al. 2011); escape the vicinity of a celestial body within several revolutions (Villac and Scheeres 2003; Paskowitz and Scheeres 2006; Haapala 2014; Davis 2011); and possess specific geometries or characteristics over short time intervals (Bosanac et al. 2018; Davis 2011). In fact, Poincaré maps have become a fundamental component of rapid trajectory design strategies applied to multi-body systems (Davis et al. 2018; Gómez et al. 2004; Haapala 2014; Koon et al. 2011; Bosanac et al. 2018).

For spatial trajectories in the circular restricted three-body problem (CR3BP), each intersection with a surface of section possesses a multi-dimensional description. Previous researchers have developed approaches to visualize and analyze the resulting higher-dimensional Poincaré maps during the trajectory design process. For instance, Haapala used multivariate representations of the map crossings, while Gómez et. al used additional constraints to filter the data represented on a lower-dimensional projection (Haapala 2014; Gómez et al. 2004). These techniques have supported using higher-dimensional Poincaré maps capturing spatial hyperbolic invariant manifolds in the CR3BP during transfer design and analysis. However, if a higher-dimensional Poincaré map is more complex, denser or associated with a nonautonomous dynamical model, data obscuration or a loss of information may still occur. Furthermore, it is challenging to extract from the map additional information about the geometries admitted by the associated trajectories due to the absence of generalizable descriptions of the geometry of nonlinear paths throughout a chaotic system; yet, such information is often valuable to a trajectory designer (Bosanac et al. 2018; Davis 2011). In this paper, data mining techniques are used to organize spatial trajectories by their geometry to facilitate analysis of a higher-dimensional Poincaré map generated in the CR3BP (Bosanac 2020; Bonasera and Bosanac 2020a, b).

Two well-known data mining techniques, clustering and dimension reduction, have been successfully applied to a wide variety of complex datasets to facilitate analysis and knowledge discovery. Clustering is used to discover groupings within a dataset without any predefined labeling criteria, while dimension reduction projects higher-dimensional data onto a lower-dimensional space that corresponds to either a fundamental geometrical structure or a set of latent features. Applying these techniques to complex datasets has guided knowledge discovery in a variety of scientific applications including visualizing proteins in single cell biology as well as pattern and outlier identification in time series data (Cao et al. 2019; Ali et al. 2019). Recent use in astrodynamics and applied mathematics includes locating coherent structures in nonlinear flows, detecting regions of stability near distant retrograde orbits via a Poincaré map, grouping periodic orbits, and extracting motion primitive sets from families of trajectories (Hadjighasem et al. 2016; Nakhjiri and Villac 2015; Villac et al. 2016; Smith and Bosanac 2019).

Recently, Bosanac has applied clustering to planar perigee maps by grouping the associated trajectories, at a fixed value of the Jacobi constant in the Sun-Earth CR3BP, by their geometry (Bosanac 2020). Specifically, each trajectory within a large set is described by a finite-dimensional feature vector that encodes information about several subsequent returns to an apsis surface of section. Then, the similarity in geometry between two trajectories is assessed using the Euclidean distance between their feature vectors. A dataset composed of the feature vectors for all trajectories in the set are input to the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) clustering algorithm to organize the solution space into clusters of geometrically similar trajectories. A two-dimensional map

is then used to visualize the initial perigees of these trajectories, colored by their cluster assignment and, therefore, their distinct geometry.

Building upon Bosanac's proof-of-concept, this paper uses data mining methods to reduce the complexity of analyzing a higher-dimensional prograde perigee map capturing spatial trajectories at a single energy level in the Sun-Earth CR3BP. First, a large set of trajectories are generated from prograde perigees, sampled near the Earth, for several returns to an apsis surface of section. Similar to the approach presented by Bosanac, a finite-dimensional feature vector is used to summarize a trajectory, while the Euclidean distance between two feature vectors is used to assess geometric similarity (Bosanac 2020). This paper then presents a new approach to grouping the trajectories according to their geometry in a computationally manageable process. Inspired by distributed clustering, the trajectories are first clustered in smaller partitions using their feature vectors as inputs to HDBSCAN. The clustering results from each partition are then sampled and combined using a cluster aggregation procedure that successively uses both clustering and dimension reduction in a binary tree structure to produce a global cluster summary. From this grouping, representative members of each cluster are used to summarize the geometries of trajectories across the entire dataset. A three-dimensional projection of the perigee map is also constructed to display the initial perigees of each trajectory, colored by their cluster assignment, in the configuration space. To mitigate the impact of data obscuration, an analyst may view only selected clusters of data on the higher-dimensional map focusing on solutions with a similar geometry as opposed to filtering with pre-defined analytical criteria. Furthermore, coloring each perigee by the geometry of their trajectories enhances the map while also reflecting the regions of existence of solutions with each type of geometry. This data-driven approach mitigates the burden on a human analyst who may otherwise manually construct such a summary or analyze a higher-dimensional map when identifying and assembling arcs of interest during trajectory design.

An additional application of a global summary of the solution space is demonstrated in this paper via a comparison with the hyperbolic invariant manifolds of spatial tori in the L_1 and L_2 gateways. Although these stable and unstable manifolds are known to govern natural transport within multi-body systems, it is currently difficult to visualize or assess their influence on the solution space, even when using Poincaré maps. Thus, a useful application of the presented data-driven approach emerges: arcs along these four-dimensional hyperbolic invariant manifolds are projected directly onto the previously generated clustering result at the same energy level using a classifier. This projection enables rapid extraction and visualization of the fundamental geometries of arcs along the manifolds as well as analysis of their connection to the characteristics of the solution space. This procedure, as well as the insights it enables, may eventually support further examination of natural transport mechanisms and trajectory design within multi-body systems.

2 Dynamical model

The CR3BP is used to describe the dynamics governing the motion of a spacecraft due to the point mass gravitational influence of two primary bodies. In this model, two primaries, P_1 and P_2, are assumed to follow circular orbits about their mutual barycenter. The third body, e.g., a spacecraft, is assumed to possess a negligible mass relative to the two primaries (Szebehely 1967). Three characteristic quantities are used to nondimensionalize mass, length and time quantities, respectively: m^*, equal to the sum of the masses of the primaries; l^*, equal to the constant distance between the two primaries; and t^* to set the mean motion

of the primaries to unity. This nondimensionalization scheme also enables definition of the parameter μ as the ratio between the mass of the smaller primary and the system mass. In the Sun-Earth system, the mass ratio is $\mu \approx 3.00348064 \times 10^{-6}$. Then, the nondimensional state of the spacecraft is expressed in an orthogonal reference frame $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$ that rotates with the two primaries: the $\hat{\boldsymbol{x}}$-axis is directed from the larger to the smaller primary, the $\hat{\boldsymbol{z}}$-axis is aligned with the orbital angular momentum of the primaries, while the $\hat{\boldsymbol{y}}$-axis completes the right-handed triad. The nondimensional state of the spacecraft is defined in this frame relative to the system barycenter as $\boldsymbol{x} = [x,\ y,\ z,\ \dot{x},\ \dot{y},\ \dot{z}]^T$. Using these definitions, the nondimensional equations of motion for a spacecraft in the CR3BP are expressed as

$$\ddot{x} - 2\dot{y} = \frac{\partial U}{\partial x}, \qquad \ddot{y} + 2\dot{x} = \frac{\partial U}{\partial y}, \qquad \ddot{z} = \frac{\partial U}{\partial z}, \tag{1}$$

where $U = (x^2+y^2)/2+(1-\mu)/r_1+\mu/r_2$ is the pseudo-potential function, and the distances of the spacecraft from the two primaries are, respectively, $r_1 = \sqrt{(x+\mu)^2 + y^2 + z^2}$ and $r_2 = \sqrt{(x-1+\mu)^2 + y^2 + z^2}$ (Szebehely 1967). The CR3BP, which is autonomous when formulated in the rotating frame, admits an integral of motion, labeled the Jacobi constant and equal to $C_J = 2U - \dot{x}^2 - \dot{y}^2 - \dot{z}^2$ (Szebehely 1967). At a single value of the Jacobi constant, trajectories are bound by zero velocity surfaces (ZVS), separating allowable and forbidden regions of motion. Within the ZVS, a wide variety of solutions exist including: equilibrium points, labeled L_i for integers $i = [1, 5]$; periodic orbits; quasi-periodic trajectories; and chaos (Szebehely 1967; Koon et al. 2011).

3 Poincaré mapping

In dynamical systems theory, Poincaré maps reduce the complexity of analyzing a set of trajectories. First, a surface of section is defined transverse to the flow (Perko 1996). Useful definitions for a surface of section include: hyperplanes defined by a specific value of a coordinate or a function; hyperspheres centered at one primary; stroboscopic sampling; and apses or other trajectory events (Verhulst 1996; Paskowitz and Scheeres 2006; Gómez and Mondelo 2001). Then, a continuous trajectory is reduced to a finite sequence of states via its intersections with the surface of section. The intersections are recorded to form the Poincaré map that is typically visualized via a lower-dimensional representation (Contopoulos 2002). Depending on the problem definition, the selected hyperplane and the map configuration, a set of map crossings may supply insight into the characteristics of the underlying solution space and facilitate identification of the dynamical mechanisms governing the associated flow (Contopoulos 2002).

This paper focuses on a periapsis map capturing spatial trajectories with initially prograde perigees in the Sun-Earth CR3BP at a single value of the Jacobi constant. In this scenario, a complete description of each initial perigee is four-dimensional. Consequently, a two- or three-dimensional projection does not supply a bijective representation of the mapping and results in data obscuration (Haapala 2014). Thus, visualization and analysis may be challenging and cumbersome for an astrodynamicist, impeding a rapid and thorough investigation of the solution space. Applying existing approaches to analyzing higher-dimensional Poincaré maps may require a priori insight into the structure of the solution space, result in the loss of information, or require a high analytical workload. However, data mining techniques offer a potential approach for addressing these challenges and effectively reducing the complexity of analyzing higher-dimensional Poincaré maps.

4 Clustering

Clustering techniques group the members of a dataset such that data in the same cluster are considered similar, while data in separate clusters are considered dissimilar (Han and Kamber 2006). In this paper, the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm, developed by Campello, Moulavi and Sander, is leveraged for cluster assignment following the approach presented by Bosanac (Campello et al. 2013; McInnes et al. 2017; Bosanac 2020). As Bosanac notes, this algorithm is appropriate for clustering the crossings on a Poincaré map according to the geometry of the associated trajectories in the CR3BP since it accommodates clusters of arbitrary shape and density; labels data that is not assigned to a cluster as noise; does not require a priori knowledge of the number of clusters; and accommodates an unknown or variable distance between data in a cluster (Bosanac 2020; Campello et al. 2013). To reduce the computational complexity of clustering a large dataset, techniques from distributed clustering are leveraged. This section presents a brief overview of HDBSCAN, followed by a discussion of a relevant approach to distributed clustering.

4.1 HDBSCAN

HDBSCAN is a density-based and hierarchical clustering algorithm developed by Campello, Moulavi and Sander (Campello et al. 2013). This algorithm takes as an input the dataset $[\boldsymbol{T}_i] = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_N\}$, composed of N members, each described by an M-dimensional feature vector $\boldsymbol{t}$. Then, HDBSCAN groups data in sufficiently dense regions of the multi-dimensional space into clusters. Two input parameters govern this clustering algorithm: m_{pts} and m_{clSize}. The first input parameter enables calculation of the core distance, d_{core}, which is the distance of a member from its $(m_{\text{pts}} - 1)$-th nearest neighbor in the M-dimensional feature vector space. The second parameter, m_{clSize}, sets the minimum number of members in a single cluster. Using these input parameters, HDBSCAN populates a distance matrix with the distance between the i-th and j-th data points calculated as $d_{\text{reach}}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \max\left\{d_{\text{core}}(\boldsymbol{t}_i), d_{\text{core}}(\boldsymbol{t}_j), d(\boldsymbol{t}_i, \boldsymbol{t}_j)\right\}$ where $d(\boldsymbol{t}_i, \boldsymbol{t}_j)$ is simply the distance between the two points. The quantity $d_{\text{reach}}(\boldsymbol{t}_i, \boldsymbol{t}_j)$ is labeled the mutual reachability distance and requires specifying a distance metric, e.g., the Euclidean norm, L^∞-norm or Hausdorff distance. A minimum spanning tree (MST) is then constructed by leveraging the computed mutual reachability distances as the weights of the edges between each pair of members of the dataset. A self-loop representing the core distance of each member is added at each node to generate an extended MST. HDBSCAN then condenses the MST to produce a dendrogram that supports cluster assignment: clusters are identified as those that both possess at least a minimum number of members and are considered sufficiently stable across the dendrogram. During the clustering process, each member of the dataset is either assigned to a cluster or considered noise (Campello et al. 2013). A schematic overview of the underlying algorithm for HDBSCAN appears in Algorithm 1 of Campello et al. (2013). As presented by Campello, Moulavi and Sander, the HDBSCAN algorithm is $\sim \mathcal{O}(MN^2)$ in time and $\sim \mathcal{O}(MN)$ in memory storage, when the clustering is performed on an $(N \times M)$-dimensional dataset (Campello et al. 2013). In this paper, the HDBSCAN algorithm is accessed via the *hdbscan* library in Python, which admits a computational complexity that approaches $\sim \mathcal{O}(N\log(N))$ under certain conditions (McInnes et al. 2017).

4.2 Distributed clustering

Distributed clustering focuses on developing strategies to efficiently and accurately cluster a large dataset that may be distributed across multiple computational machines and time. These approaches tend to be composed of four fundamental steps (Aggarwal and Reddy 2018). First, a dataset is split into multiple partitions: each partition is clustered to produce a local clustering result. Then, each local clustering result is reduced to a representative subset; these subsets are designed to support rapid data sharing, low storage requirements and sufficient representation of the structure of the data in the partition. In the third step, these representative solutions from the local clusters are aggregated to construct a global clustering result for the entire dataset. Then, this global model may be returned to each local partition to facilitate labeling any data that does not appear in the global clustering result. Bendechache, Le-Khac and Kechadi demonstrate that this approach effectively minimizes communication between partitions of a dataset, scales well when the dataset size increases and may outperform centralized clustering algorithms in both result quality and execution time (Bendechache et al. 2016).

5 Dimension reduction

Dimension reduction is used to project a higher-dimensional dataset onto a lower-dimensional space (Wenskovitch et al. 2018). One class of dimension reduction algorithms is manifold learning, which discovers a lower-dimensional manifold on which the data is assumed to lie (McInnes et al. 2018). When this embedding sufficiently captures the structure of the dataset, it may be leveraged for processing a dataset prior to clustering or visualization (Wenskovitch et al. 2018; Han and Kamber 2006). In this paper, the uniform manifold approximation and projection (UMAP) algorithm is used for nonlinear dimension reduction (McInnes et al. 2018). This algorithm is selected due to its demonstrated success in capturing the structure of complex datasets via lower-dimensional representations in a variety of disciplines (Becht et al. 2019; Li et al. 2019).

UMAP is a nonlinear dimension reduction technique that leverages concepts from algebraic and fuzzy topology (McInnes et al. 2018). A high-dimensional dataset that is input to UMAP is assumed to be uniformly distributed on a Riemannian manifold that is locally connected; UMAP approximates this manifold to identify a lower-dimensional representation of the data. First, UMAP forms a topological representation of the high-dimensional dataset. This process begins by constructing a local Riemannian metric in the vicinity of each member of the dataset, with the local neighborhood defined using the n_n-nearest neighbors. The fuzzy union of these local metric spaces is used to form a fuzzy simplicial set that is represented by an n_n-neighbor graph. This graph is constructed with a specified distance metric and represented via a force-directed layout to supply a fuzzy topological representation of the high-dimensional dataset. Then, stochastic gradient descent is used to identify an embedding onto a lower-dimensional Euclidean space with a similar fuzzy topological structure to the actual dataset. UMAP initializes this low-dimensional representation through spectral embedding. Then, the embedding is refined by minimizing the cross-entropy between the fuzzy topological structures of the original dataset and lower-dimensional representation. A schematic overview of the underlying algorithm for UMAP appears in Section 4.1 of McInnes et al. (2018). The overall computational complexity of UMAP is driven by the n_n-neighbor

search, empirically approximated as $\sim \mathcal{O}(N^{1.14})$, and the stochastic gradient descent step, $\sim \mathcal{O}(n_n N)$ (McInnes et al. 2018).

This work leverages three input parameters to influence the low-dimensional projection generated by UMAP (McInnes et al. 2018). First, n_n governs the size of the local neighborhood used to construct the weighted n_n-neighbor graph. A small integer value of n_n prioritizes the local structures across the manifold, while a higher integer value focuses on the global structure. Then, m_{dist} specifies the minimum allowable distance between any two points in the lower-dimensional space; as a result, this parameter also influences the topological structure of the data in the learned representation. Finally, n_c defines the dimension of the lower-dimensional embedding of the original dataset. In this paper, the UMAP algorithm is accessed via the *umap-learn* library in Python (McInnes et al. 2018).

6 Classification

Classification is a form of supervised learning that assigns unlabeled data to classes via a classifier constructed using already labeled data (Han and Kamber 2006). In general, the classification process is divided into two components: classifier training or construction, and testing (Han and Kamber 2006). First, a subset of members of a labeled dataset are used to construct a classifier that converts a feature vector into a predicted label. Then, during the testing phase, this classifier is used to assign the remaining members of the dataset to a class, thereby testing the classifier. Although there are a wide variety of classification algorithms, this paper employs one of the most foundational methods to support a proof-of-concept: weighted k-nearest neighbor classification. This classification scheme has been employed in various disciplines including, for example, processing large datasets in astronomy and celestial mechanics (Ivezić et al. 2019; Mommert et al. 2020). The incorporation of more complex or computationally efficient classifiers serves as an avenue for future work.

The weighted k-nearest neighbor classification algorithm assigns classes to unlabeled data using a straightforward k-nearest neighbor search applied to the labeled dataset (Han and Kamber 2006). Given a member of an unlabeled dataset, the algorithm searches for the k-nearest neighbors in the labeled dataset where k is a user-specified integer. This search is performed using a specified similarity metric; in this paper, selected as the Euclidean distance between the feature vectors of two members. Each of the k-nearest neighbors in the labeled dataset are assigned a weight; in this paper, equal to the inverse square of the distance from the unlabeled member in the higher-dimensional feature space. Among these k-nearest neighbors, the class that possesses the largest cumulative weight supplies the prediction for the class of the unlabeled member. This algorithm possesses a computational complexity $\sim \mathcal{O}(kN_C M_C)$ where N_C is the number of members in the training dataset and M_C is the dimension of the input feature vector (Han et al. 2014). In this paper, the weighted k-nearest neighbor classification algorithm is accessed via MATLAB® (MathWorks 2020).

7 General procedure for processing higher-dimensional Poincaré map data

This paper presents a framework for analyzing higher-dimensional Poincaré maps in the spatial CR3BP using techniques from data mining. This section presents a general outline of this procedure in the context of a prograde perigee map at a single Jacobi constant in the

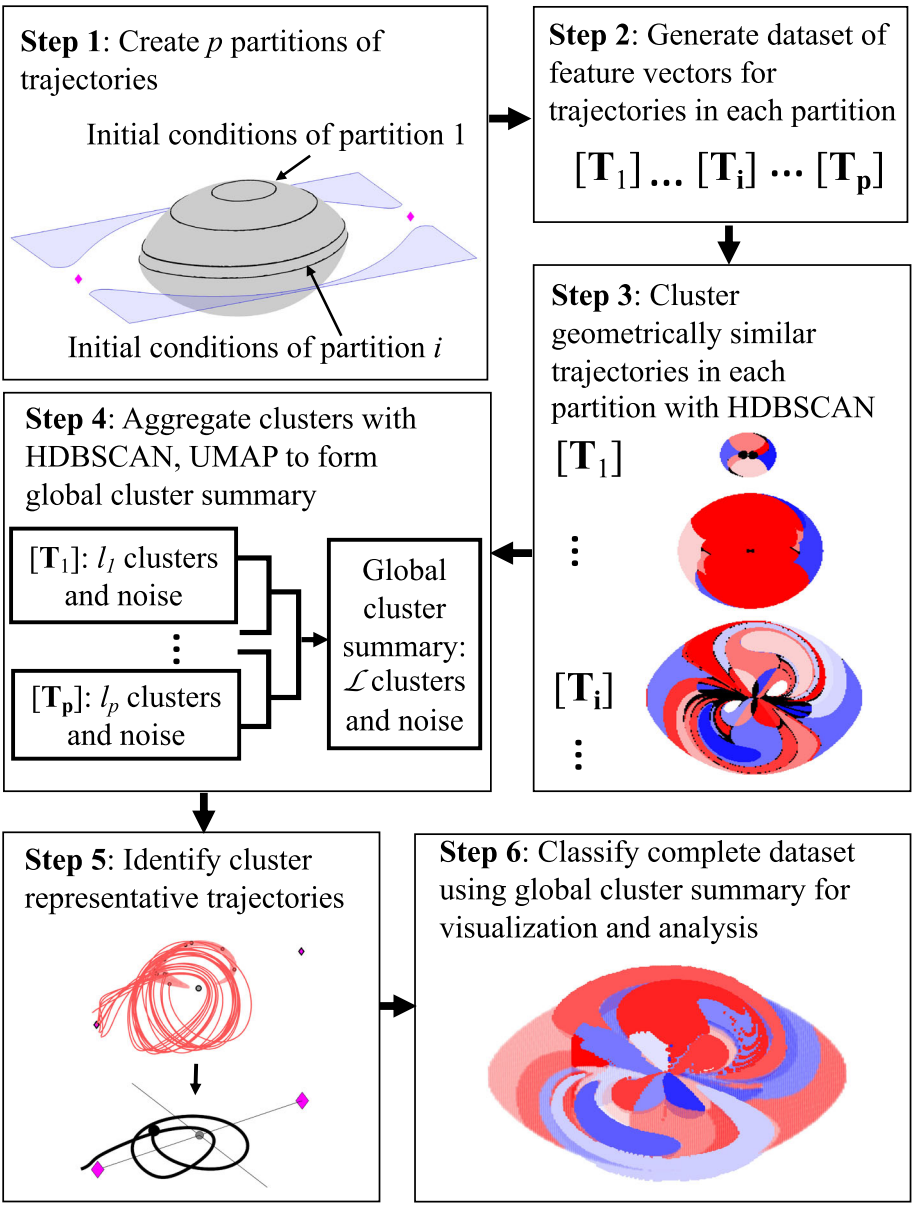

Fig. 1 Graphical overview of framework for analyzing higher-dimensional Poincaré maps in the spatial CR3BP using techniques from data mining

Sun-Earth CR3BP. A graphical summary of this framework appears in Fig. 1, with each step explained in detail in this section.

7.1 Step 1: constructing partitions of initial conditions

The complete set of initial conditions used to construct a periapsis map is seeded as prograde perigees near the Earth, within the ZVS at a selected value of C_J. First, a periapsis surface of section is defined relative to the Earth (Villac and Scheeres 2004). Each perigee satisfies the following two conditions in the rotating frame of the Sun-Earth CR3BP:

$$(x-1+\mu)\,\dot{x}+y\dot{y}+z\dot{z}=0 \quad \cup \quad (x-1+\mu)\,\ddot{x}+y\ddot{y}+z\ddot{z}+\dot{x}^2+\dot{y}^2+\dot{z}^2>0. \quad (2)$$

Candidate initial position coordinates are selected using a grid defined by N_x equally spaced values of x within the range $[x_{\min}, x_{\max}]$, N_y equally spaced values of y within the range $[y_{\min}, y_{\max}]$ and N_z equally spaced values of z within the range $[z_{\min}, z_{\max}]$. For each candidate position vector, the speed at perigee is calculated from the Jacobi constant as $v=\sqrt{2U-C_J}$. If v possesses a real value, the position coordinates x, y, z lie within the ZVS. Then, an associated velocity unit vector is calculated via a linear combination of two basis vectors defining the plane that is perpendicular to the position vector measured relative to the Earth. These basis vectors are defined as the following two vectors that lie in the two-dimensional nullspace of the vector $[x-1+\mu, y, z]$: $\hat{\boldsymbol{u}}_1$ possesses a z-component equal to zero and is directed to produce an instantaneous angular momentum vector with a positive z-component when the velocity vector is parallel to $\hat{\boldsymbol{u}}_1$; and $\hat{\boldsymbol{u}}_2$ is perpendicular to $\hat{\boldsymbol{u}}_1$, with a positive z-component. An initial velocity vector, $\boldsymbol{v}_0$, is then calculated from these basis vectors using a specified angle θ as

$$\boldsymbol{v}_0 = v\left(\frac{\cos(\theta)\hat{\boldsymbol{u}}_1+\sin(\theta)\hat{\boldsymbol{u}}_2}{||\cos(\theta)\hat{\boldsymbol{u}}_1+\sin(\theta)\hat{\boldsymbol{u}}_2||}\right). \quad (3)$$

If the resulting state vector satisfies the perigee conditions in Eq. (2), it is used to define a feasible initial condition.

The complete set of initial conditions is partitioned into p groups of prograde perigees near the Earth. In this work, perigees in the i-th partition possess the same value of the z-coordinate and θ. Note that for distinct values of the z-coordinate, the perigees that lie within the ZVS may encompass distinct ranges of values in the x and y coordinates. Thus, the number of initial conditions within each partition are not necessarily equivalent using this discretization approach.

7.2 Step 2: generating a dataset for each partition

The trajectories associated with the initial conditions in the i-th partition are generated in the Sun-Earth CR3BP. Specifically, each initial condition is propagated forward in time until satisfying any of the following termination conditions: completing N_{ret} subsequent returns to the surface of section defined by Eq. 2; passing within a nondimensional distance of 10^{-6} from the Earth; passing through either the L_1 or L_2 gateways, as determined by the trajectory extending beyond the locations of the equilibrium points with $x \leq 0.989$ and $x \geq 1.011$, respectively; or a perigee occurring within the L_1 or L_2 gateways rather than within the region encompassed by perigees near the Earth where $0.993 \leq x \leq 1.007$. Note that the final condition is only introduced to facilitate a clear comparison with the stable manifolds associated with libration point orbits in later sections.

Each trajectory that admits at least one additional perigee prior to satisfying any of the termination conditions is summarized via an M-dimensional feature vector that supports a straightforward assessment of geometric similarity between two trajectories. Although there

are a variety of existing approaches, this paper specifically focuses on assessing geometric similarity between two trajectories as the Euclidean distance between two finite-dimensional feature vectors that encode a summary of each trajectory (Zheng and Zhou 2011). Although this approach requires the definition of a single feature vector of fixed length across all members of the dataset, it is computationally efficient when applied to large datasets. Furthermore, Bosanac has demonstrated that this approach supports the successful differentiation of planar trajectories in the CR3BP by their geometry (Bosanac 2020). Alternative and well-known approaches to assess the similarity between two trajectories, such as dynamic time warping, may offer a more robust comparison between two sequences of spatiotemporal data of dissimilar lengths (Zheng and Zhou 2011). However, the definition of this trajectory similarity measure may result in a significant increase in data storage requirements and computational time when applying HDBSCAN and UMAP. A thorough examination of these alternative trajectory similarity measures in the context of the presented framework may, however, offer a valuable avenue of future work as the complexity or time horizon of the trajectories of interest increases.

Similar to the approach presented by Bosanac, each trajectory is summarized using a finite-dimensional feature vector that encodes information about several returns to an apsis surface of section (Bosanac 2020). Specifically, the j-th trajectory is represented by the unique feature vector $\boldsymbol{t}_j = \left[\boldsymbol{s}_{j,0},\ \boldsymbol{s}_{j,1},\ \ldots,\ \boldsymbol{s}_{j,2N_{\text{ret}}+1}\right]^T$, where $\boldsymbol{s}_{j,k}$ describes the k-th apsis along the j-th trajectory relative to the Earth (Bosanac 2020). An apsis, defined relative to the Earth, is calculated as a state along a trajectory that satisfies only the first condition in Eq. 2. Each apsis is then summarized to reflect its state as well as the time since the beginning of the trajectory via the vector $\boldsymbol{s}_{j,k}$, defined as

$$\boldsymbol{s}_{j,k} = \left[\tilde{t}_{j,k}, \tilde{x}_{j,k},\ \tilde{y}_{j,k},\ \tilde{z}_{j,k},\ 0.5\left(\hat{\boldsymbol{h}}_{j,k} \cdot \hat{\boldsymbol{z}} + \text{sign}\left(\hat{\boldsymbol{h}}_{j,k} \cdot \hat{\boldsymbol{z}}\right)\right)\right]^T, \tag{4}$$

where the tilde indicates the use of normalization, mitigating poor conditioning between components of the feature vector. Specifically, $\tilde{t}_{j,k}$ is the time at which the k-th apsis occurs, measured from zero at the initial condition and normalized by the total propagation time along the j-th trajectory. Then, $\tilde{x}_{j,k},\ \tilde{y}_{j,k},\ \tilde{z}_{j,k}$ are the position components of the state at the k-th apsis in the rotating frame, normalized by the nondimensional distance between the Earth and L_2. The final term in $\boldsymbol{s}_{j,k}$ is calculated using the z-component of the instantaneous angular momentum vector. This function is designed to: introduce separation between prograde and retrograde apses via the discrete-valued quantity $\text{sign}(\hat{\boldsymbol{h}}_{j,k} \cdot \hat{\boldsymbol{z}})$; produce a continuous range of values for states with the same direction of motion; and output values within the range $[-1, 1]$. Due to the apsis constraints, this feature vector reflects the direction of motion and the orientation of the orbital plane relative to the $\hat{\boldsymbol{z}}$-axis. If the trajectory terminates prior to the k-th apsis, each vector $\boldsymbol{s}_{j,k}$, with $k \leq 2N_{\text{ret}} + 1$, is assigned a placeholder value, equal to $[t_t, 10, 0, 0, 0]$ where $t_t = 1$ for the first apsis after termination and $t_t = 0$ for all remaining apses. This placeholder vector is designed to introduce a sufficient separation from members of the dataset that do not terminate early (Bosanac 2020). Then, the dataset $[\boldsymbol{T}_i]$ is populated with the M-dimensional feature vectors associated with each trajectory in the i-th partition.

7.3 Step 3: cluster each individual partition

In the application explored within this paper, there is limited a priori knowledge of an appropriate division of the dataset; thus, relative cluster validation is used when selecting the parameters governing HDBSCAN. Following the work of Bosanac, the density-based clus-

tering validation (DBCV) index introduced by Moulavi et al. is used in this paper as a scalar representation of the quality of a clustering result, relative to alternative configurations of the clustering procedure applied to the same dataset (Moulavi et al. 2014; Bosanac 2020). To calculate the DBCV index, Moulavi et al. defined two more quantities: $\mathrm{DSC}(C_j)$, the density sparseness of a cluster and $\mathrm{DSPC}(C_j,\ C_k)$, the density separation of clusters C_j and C_k. The validity index $V_C(C_j)$ of the j-th cluster is then calculated as

$$V_C(C_j) = \frac{\min_{k\in\{0,\ldots,l_i-1\},\, j\neq k}(\mathrm{DSPC}(C_j, C_k)) - \mathrm{DSC}(C_j)}{\max\left[\min_{k\in\{0,\ldots,l_i-1\},\, j\neq k}(\mathrm{DSPC}(C_j, C_k)), \mathrm{DSC}(C_j)\right]}, \tag{5}$$

where l_i is the total number of clusters for dataset $[\boldsymbol{T}_i]$. Then, the DBCV index, which summarizes the quality of the entire set of clusters, is computed as

$$\mathrm{DBCV} = \sum_{j=1}^{l_i} \frac{|C_j|}{|T_i|} V_C(C_j) \tag{6}$$

From this definition, $-1 < \mathrm{DBCV} < 1$, with positive values indicating a good clustering result. Computationally efficient approximations of the DBCV and validity indices, available in the *hdbscan* library in Python, are leveraged in this analysis for input parameter selection (McInnes et al. 2017).

The input parameters that govern the HDBSCAN algorithm to cluster each individual partition are selected through an iterative exploration for one single partition. The values of m_{pts} and m_{clSize} are selected to: obtain a positive value of the DBCV index; generate a low percentage of datapoints identified as noise; and avoid either an excessively large or negligibly small number of clusters. A demonstration of this input parameter selection process appears in the work of Bosanac (2020). To reduce the burden on a human analyst as the number of partitions increases and to ensure consistency across the dataset, the same combination of m_{pts} and m_{clSize} is used for all clustering steps and produces reasonable results in the application analyzed in this paper.

Clustering is performed independently on the data in each of the p partitions using the selected values for m_{pts} and m_{clSize} and the specified similarity measure. When applied to $[\boldsymbol{T}_i]$, HDBSCAN produces $l_i \in \mathbb{N}$ clusters along with one set of noise points. Each cluster contains trajectories that are geometrically similar, as assessed using the Euclidean distance between feature vectors that encode information about several returns of each trajectory to an apsis surface of section. Trajectories associated with two distinct clusters are, as a result of the selected similarity measure and the parameters governing HDBSCAN, considered geometrically dissimilar. Following application to all p partitions, a total of $\sum_{i=1}^{p} l_i$ clusters, each localized to a single partition of the dataset, and p sets of noise are identified.

7.4 Step 4: cluster aggregation

The clusters associated with each partition of the dataset are used to identify a minimal set of unique clusters for the entire dataset in this cluster aggregation step. The goal is to identify clusters of similar solutions that exist across multiple partitions in a computationally feasible and robust manner that requires little intervention from a human analyst. To implement this procedure, an approach from the discipline of distributed clustering, as described in Sect. 4.2, is employed. Specifically, the independent clustering of data within an individual partition is considered a local clustering result, or local model. The collection of p local clustering results are then gradually aggregated to form a

global clustering result for the entire dataset. Pseudocode for the cluster aggregation process is presented in Alg. 1 to supplement the detailed description within this subsection.

Algorithm 1 Cluster aggregation process.

Require: array of datasets DB, HDBSCAN parameters: $m_{\text{pts}}, m_{\text{clSize}}$, UMAP parameters: n_n, m_{dist} and $N_{\text{int}}, N_{\text{noise}}$
  while length$(DB) > 1$ **do**
    $p \leftarrow$ length(DB)
    $DB_n \leftarrow [\ ]$
    for $i = 0 \rightarrow p - 1$ **do**
      if $i < p$ or i is even **then**
        $l \leftarrow$ HDBSCAN$(DB[i], m_{\text{pts}}, m_{\text{clSize}}$).get_lb() ▷ get_lb(): label data by cluster
        $db \leftarrow [\ ]$
        for $j = \min(l) \rightarrow \max(l)$ **do**
          $l_{\text{clustj}} \leftarrow$ where$(l == j)$ ▷ where$(l == j)$: find data in cluster j
          if $j ==$ label for noise **then**
            $m \leftarrow \lfloor$length$(l_{\text{clustj}})/N_{\text{noise}} \rfloor$ ▷ $\lfloor x \rfloor$: floor of x
          else
            $m \leftarrow N_{\text{int}}$
          end if
          if length$(l_{\text{clustj}}) > m$ **then**
            $l_{samp} = l_{clustj}$.esamp(m) ▷ esamp(m): evenly sample up to m members
            $db_l \leftarrow DB[i][l_{\text{samp}}]$
          else
            $db_l \leftarrow DB[i][l_{clustj}]$
          end if
          db.append(db_l) ▷ x.append(y): append y to x
        end for
        $DB_n[\lfloor i/2 \rfloor] \leftarrow db$
      else
        $DB_n[\lfloor (p-1)/2 \rfloor] \leftarrow DB[p]$
      end if
    end for
    $DB \leftarrow DB_n$
  end while
  $DB_r \leftarrow$ UMAP$(DB, n_n, m_{\text{dist}})$ ▷ DB has 1 dataset
  $l_G \leftarrow$ HDBSCAN$(DB_r, m_{\text{pts}}, m_{\text{clSize}})$.get_lb() ▷ l_G labels data in global summary

To enable a computationally efficient aggregation process, each cluster in each partition is summarized by a reduced set of members. In this paper, these representative members of a cluster are selected as follows: for clusters of up to N_{int} members, all members form the representative set; for clusters of more than N_{int} members, the members are evenly sampled based on ordering in the dataset to produce an intermediate summary of the cluster, possessing up to N_{int} members. Each set of noise points is sampled to contribute every N_{noise}-th member.

Cluster aggregation is performed sequentially using a binary tree approach. At the first level of this binary tree, cluster aggregation is performed on unique pairs of neighboring partitions to produce p_1 intermediate cluster summaries, where $p_1 = \lceil p/2 \rceil$. At the next level of the binary tree, cluster aggregation is performed on unique pairs of these p_1 intermediate cluster summaries. When there is an odd number of partitions or intermediate clustering summaries at a single level of the binary tree, the last result is not clustered again; rather, its clusters are moved to the next level of the binary tree structure. At each aggregation step, HDBSCAN is used to produce a new set of clusters for each intermediate cluster summary. In this paper, the same values of the input parameters are used when individually clustering each partition and during cluster aggregation steps; this approach is observed to produce suitable results, while reducing the complexity of the input parameter selection process. At the final step of this cluster aggregation procedure, where only one intermediate cluster summary is used

to produce the final global clustering result, a modification to the clustering approach is employed.

UMAP is used during the final step of cluster aggregation to project the intermediate cluster summaries onto a lower-dimensional space prior to clustering. When a dataset is described by a high-dimensional feature vector, using dimension reduction to identify a smaller set of fundamental variables that sufficiently captures the characteristics of the original information offers numerous benefits during subsequent clustering steps, such as: reducing the potential for overfitting due to the curse of dimensionality (Aggarwal and Reddy 2018); reducing the computational resources required for storage and processing; and reducing the percentage of members in a dataset that are classified as noise when clustering with HDBSCAN. Advances in the discipline of clustering may reveal alternative algorithms which may not experience such issues; in that case, the use of UMAP for preprocessing at the final step of cluster aggregation may become unnecessary.

Care must be taken when applying dimension reduction methods such as UMAP to a dataset prior to clustering (Wenskovitch et al. 2018). Specifically, dimension reduction techniques may not preserve the density of a dataset; rather, only the structure of the dataset. In this paper, applying clustering to a projection of the dataset onto a lower-dimensional space calculated via UMAP produces a smaller percentage of members labeled as noise and a relative ease in selecting the input parameters, with only a small increase in the number of clusters compared with performing this final clustering step in the original M-dimensional space. Furthermore, this paper applies UMAP only at the final step because it requires a computational time of up to a few minutes for larger datasets, as opposed to a computational time on the order of seconds for clustering the same high-dimensional dataset via HDBSCAN on a computer with a 4 GHz Intel Core i7 processor.

When UMAP is used in the final step of cluster aggregation, the UMAP input parameters, n_n and m_{dist}, are set equal to a fixed set of values. These values are selected via iterative exploration to sufficiently separate distinct groups and create compact clusters. Once clustering has been applied to the dimensionally reduced dataset in the final step of cluster aggregation, a global clustering result is generated; the result is $\mathcal{L}$ unique clusters of map crossings associated with a summary of the dataset and one group of noise points.

7.5 Step 5: calculate global cluster representatives

To aid visualization and analysis, a single representative trajectory is generated using the medoid of each of the $\mathcal{L}$ clusters associated with the global clustering result; consistent with the approach used by Bosanac (2020). The medoid is defined as the member of a cluster that is most similar to the other members of the cluster (Cichosz 2015). For cluster $C_j = \{\boldsymbol{t}_1^{(j)}, \boldsymbol{t}_2^{(j)}, \ldots, \boldsymbol{t}_{M_j}^{(j)}\}$, with cluster cardinality $|C_j| = M_j \in \mathbb{N}$, the medoid of the j-th cluster is calculated as

$$\boldsymbol{t}_{\text{med}}^{(j)} = \text{argmin}_{\boldsymbol{t}_k^{(j)} \in C_j} \sum_{i=1,\, i \neq k}^{M_j} d\left(\boldsymbol{t}_i^{(j)}, \boldsymbol{t}_k^{(j)}\right), \tag{7}$$

where $d(\cdot, \cdot)$ is a distance metric, selected as the Euclidean distance in this paper.

7.6 Step 6: classification of the complete dataset

Weighted k-nearest neighbor classification is used to assign each member of the original dataset to a cluster, given information supplied by the global clustering result. Specifically, the M-dimensional feature vectors of each member of the $\mathcal{L}$ unique clusters of the final global clustering result supply predictor information, while the integer cluster assignment labels are used as the response information. Of course, this approach assumes that each perigee must lie in one of the recovered classes. Then, the classifier is configured to compare each member of an unlabeled dataset to its ten nearest neighbors, calculated using the Euclidean norm and weighted by the inverse square of the distance. Five-fold cross-validation of the classifier is applied to the labeled dataset. If the classifier possesses a high estimated accuracy, it is used to assign each map crossing in the original dataset to one of the $\mathcal{L}$ unique groups in the global clustering result.

8 Application: prograde perigee map in the Sun-Earth CR3BP

The presented data-driven approach to analyze higher-dimensional map crossings and the geometry of their associated trajectories is demonstrated for a prograde perigee map in the Sun-Earth CR3BP; this map displays the initial perigees of spatial trajectories at a Jacobi constant of $C_J = 3.00088$ with various initial values of the z-component of velocity. At this Jacobi constant, both the L_1 and L_2 gateways are open, producing a complex solution space. A subset of the solution space with an initial value of $\dot{z} = 0$ only possesses a three-dimensional description due to the additional constraint on $\dot{z}$. Accordingly, this subset of spatial perigees supports an initial analysis and verification of the resulting global cluster summary of the solution space. Then, for various initial values of the z-component of velocity, each initial perigee that is seeded from the map possesses a description that is four-dimensional. As a result, visualization of a wide variety of map crossings on a single Poincaré map is challenging due to both data obscuration and a loss of information when the map is displayed in the three-dimensional configuration space. It is in a scenario such as this that a data-driven approach may significantly reduce the workload required of a human analyst examining the solution space.

8.1 Generating a global cluster summary

Prograde perigees are sampled in the vicinity of the Earth at a Jacobi constant of $C_J = 3.00088$. The initial perigees in the i-th partition are constructed from up to 200 evenly sampled values of x in the range $[0.993, 1.007]$, up to 200 values of y within the range $[-0.0043, 0.0043]$ and a single value of z in the sequence of 64 evenly sampled z-coordinates within the range $[-0.0033, 0.0033]$, in nondimensional units. All perigees in the i-th partition possess a velocity vector with a value of θ that is equal to one value in the set $[-89°, -80°, -70°, \ldots, 0°, \ldots, 70°, 80°, 89°]$. The initial conditions within each partition are then propagated forward in time for up to three subsequent perigees to generate the associated trajectories. These trajectories, with up to seven apses relative to the Earth, are then described using 35-dimensional feature vectors, as defined in Sect. 7.2, to produce the dataset $[\boldsymbol{T}_i]$ for the i-th partition, containing between 402 and 25,204 members.

The input parameters used by HBSCAN at each instance of clustering are selected via iterative exploration in the context of the partition with the lowest positive value of z and

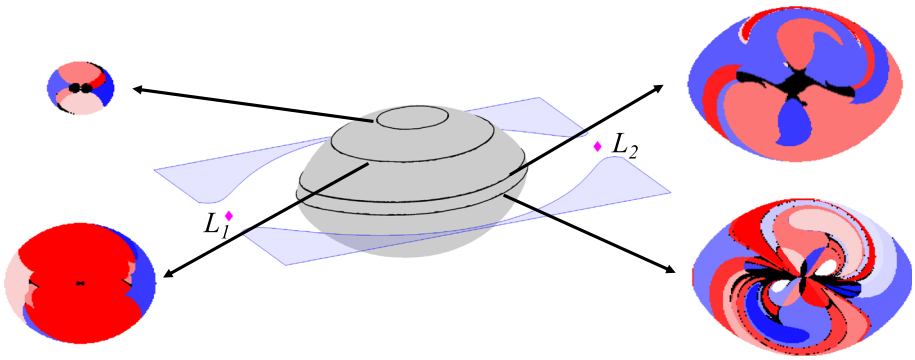

Fig. 2 Local clustering results for selected partitions of the prograde perigee map in the Sun-Earth CR3BP at $C_J = 3.00088$ with $\dot{z} = 0$, colored in shades of red and blue according to the cluster in individual partitions

$\dot{z} = 0$. The values $m_{\text{pts}} = 100$ and $m_{\text{clSize}} = 100$ are selected to produce a clustering result for this partition with a relatively high positive value of the DBCV index, equal to 0.527, and a low percentage of noise, equal to 6.7%. However, this combination of input parameters may not produce similarly low levels of noise across all partitions. Aggregating clusters via a binary tree approach and including a sample of the noise points at each step mitigates the impact of any substandard results across a single partition using these values of the input parameters. In addition, note that alternate combinations of these input parameters may produce either similar or distinct clustering results.

HDBSCAN is used to independently cluster the map crossings within each partition of the dataset. An example of the results of independently clustering each partition is shown in Fig. 2 for the partitions where $\dot{z} = \theta = 0$. The center of this figure displays in gray the feasible initial conditions across this subset of the data in the configuration space in the Sun-Earth rotating frame. In this center figure, the L_1 and L_2 equilibrium points are displayed as magenta diamonds, while the light blue region indicates the zero velocity curves that lie in the plane of the primaries at $C_J = 3.00088$. The black curves indicate the external boundaries of four partitions that correspond to the insets of this figure, each displaying the results of individually clustering each partition. In these insets, initial conditions are colored in distinct shades of red and blue that reflect the local cluster assignments. While some colors may be repeated across each inset figure to ensure sufficient visual differentiation, the clusters associated with each partition are, at this step, independent and distinct. Within each inset figure, there are some black crossings: these initial conditions are initially classified as noise, typically due to the associated members not meeting the minimum cluster size threshold.

Cluster aggregation is first implemented for each of the partitions that possess the same value of θ. This particular segmentation of the cluster aggregation procedure is employed to limit the computational load required during this proof of concept. First, each cluster from each dataset at a single level of the binary tree is represented by up to 400 members using the strategy described in Sect. 7.4. Next, HDBSCAN is applied to intermediate datasets with $m_{\text{pts}} = 100$ and $m_{\text{clSize}} = 100$, continuing until one final intermediate summary dataset remains for each value of θ. Then, UMAP is used to project the data onto a three-dimensional space, consistent with the dimension of the description of each initial perigee at fixed values of C_J and θ. During this step, the input parameters $m_{\text{dist}} = 0$ and $n_n = 500$ are used to produce compact clusters that preserve global structure, while limiting the computational time associated with computing the distance to n_n nearest neighbors. Then, HDBSCAN is

applied to the dataset using the embedding constructed by UMAP. The result is 19 sets of clustering results, each summarizing the solution space of perigees at a fixed Jacobi constant and a fixed value of θ that constrains the initial velocity vector.

Cluster aggregation is then applied to the 19 groups of clusters, each corresponding to fixed values of θ, to produce one global cluster summary, with three differences in the selected parameters. Each cluster in the intermediate summaries is constrained to possess up to 800 members and, in the final dimension reduction and clustering step, up to 200 members. This parameter selection balances summarizing each cluster via a sufficient number of members with reducing computational time and memory requirements for this larger dataset during the final steps of aggregation. In addition, the embedding constructed by UMAP in the final step is defined as four-dimensional, consistent with the dimension of the description of an initial perigee at a fixed Jacobi constant with an unconstrained value of θ. Finally, n_n is set equal to 250, consistent with the smaller number of members used to represent each cluster. Following this final cluster aggregation step, the global cluster summary consists of 166 clusters, containing a total of 44,377 perigees that exist across all partitions of the dataset, and one noise set, composed of 566 perigees. For information about the global clustering result, "Appendix 1" displays the representative trajectories for all 166 clusters in the configuration space.

To facilitate a detailed visualization and analysis, each member of the complete dataset is projected onto the global cluster summary. The weighted k-nearest neighbor classification scheme is used to assign the unlabeled data to one of the 166 clusters associated with the global clustering result; the classifier is estimated by MATLAB to possess an accuracy of 99.9%. Of course, this approach assumes that all trajectories in the complete dataset possess a geometry associated with one of the recovered clusters and, therefore, may result in some incorrect labels for trajectories that exist in groups that are too small to have been incorporated into the global cluster summary. It is observed through analysis of the data, however, that these cases comprise a negligibly small fraction of the periapses in the entire dataset. The result of this classification procedure is a large set of spatial perigees with an unconstrained value of $\dot{z}$ in the Sun-Earth CR3BP at $C_J = 3.00088$ that are clustered by the geometry of the associated trajectories.

8.2 Analyzing a subset of the solution space with $\dot{z} = 0$ at $C_J = 3.00088$

Analysis of a subset of the solution space corresponding to initially prograde perigees at $C_J = 3.00088$ with $\theta = \dot{z} = 0$ facilitates verification of the results in a scenario where the map crossings only possess a three-dimensional description. As a result, the associated map representation does not suffer from a loss of information when displayed in the three-dimensional configuration space; rather, the only issue is data obscuration, which may be mitigated by displaying only subsets of clusters. Figure 3 displays only the initial perigees with $\theta = \dot{z} = 0$ in the three-dimensional configuration space for a subset of the clusters that intersect the plane of the primaries. Note that visualizing only the initial conditions of specific sets of trajectories on a Poincaré map and, in some cases, using color to reflect additional useful information is an approach that has been used by several researchers studying short-term trajectories that admit a small number of returns to a surface of section (Davis 2011; Koon et al. 2011; Bosanac et al. 2018; Villac and Scheeres 2003). Overlaid in Fig. 3 using semi-transparent markers are map crossings with $z = \dot{z} = 0$ that are assigned to clusters using the classifier. Each map crossing in Fig. 3 is colored by the assigned cluster using distinct shades of red and blue. At the boundaries of the figure, the representative trajectories

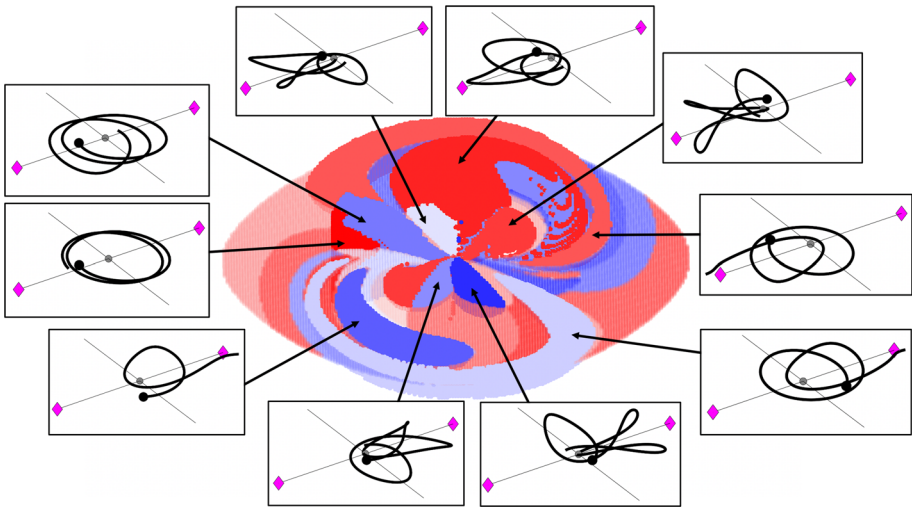

Fig. 3 Selected clusters of prograde perigees near the plane of the primaries in the Sun-Earth CR3BP at $C_J = 3.00088$ with $\dot{z} = 0$. Clusters are assigned based on geometric similarity of the associated trajectories and colored in distinct shades of red and blue. Representative trajectories of selected perigees in each cluster appear in the insets

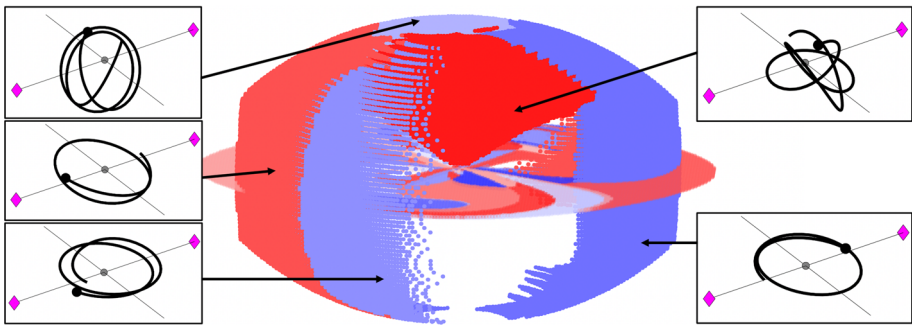

Fig. 4 Selected clusters of prograde perigees with large z-excursions in the Sun-Earth CR3BP at $C_J = 3.00088$ with $\dot{z} = 0$. Clusters are assigned based on geometric similarity of the associated trajectories and colored in distinct shades of red and blue. Representative trajectories of selected perigees in each cluster appear in the insets

for selected clusters are displayed. Some members of these clusters extend out of the plane of the primaries; the maximum z-extension of any perigee in these clusters is approximately 0.001938. Similar information is shown in Fig. 4 for selected clusters of trajectories with a large z-excursion; trajectories that are assigned to these clusters possess initial perigees with a maximum z-extension of 0.0033.

Examination of Figs. 3 and 4 as well as the representative trajectories of clusters that possess members within the $\dot{z} = \theta = 0$ partition reveals that the data-driven approach successfully differentiates trajectories by their geometry. In fact, these clusters separate: transiting and non-transiting trajectories, bounded and nonbounded trajectories over the short time horizon of up to three revolutions, trajectories with a different direction of motion at subsequent apoapses, trajectories that begin within distinct regions of the configuration

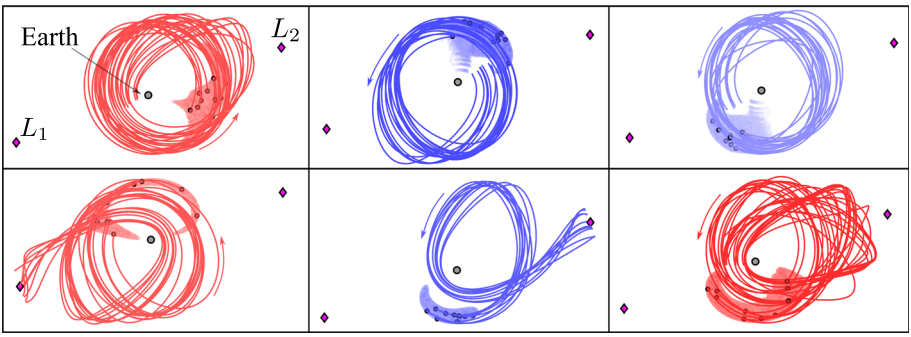

Fig. 5 Selected trajectories associated with prograde perigees within six clusters in the Sun-Earth CR3BP at $C_J = 3.00088$ with $\dot{z} = 0$, colored in distinct shades of red and blue according to the cluster

space, as well as trajectories with different apsis locations and apsidal rotations. In addition, trajectories with similar geometries are successfully grouped within the same cluster. Figure 5 displays 15 randomly selected trajectories within each of six clusters in the Sun-Earth rotating frame. Each trajectory is colored according to their cluster assignment and consistent with Figs. 3 and 4. Overlaid on these subfigures are the initial perigees associated with the entire cluster of interest, visualized using transparent markers, and the initial perigees of the selected trajectories, indicated with black circles. The trajectories across each of the six clusters in Fig. 5 are geometrically similar, i.e., they exhibit a similar shape as well as apsis location and apsidal rotations across each subsequent revolution around the Earth. Thus, the presented framework successfully organizes a large set of trajectories into clusters of distinct geometries, while grouping trajectories of similar geometry.

Partial visualization of the clusters, as shown in Figs. 3 and 4, also supports an initial verification of the global clustering result. The clusters that pass through $z = 0$ are close to symmetric about the plane of the primaries, consistent with the known symmetry in the CR3BP about the xy plane. Only negligibly small differences exist between the positive and negative z extensions of some clusters due to the data-driven nature of the presented framework, which leverages a global summary of a larger dataset, a classifier, and an assumption that all trajectories are associated with one of the 166 recovered clusters. In the CR3BP, these deviations could be eliminated entirely by only clustering trajectories associated with initial perigees that possess a positive z-component and exploiting the symmetry to analytically extend the clusters to initial perigees with a negative z-component. However, such an approach is not used in data generation to facilitate verification.

8.3 Analyzing the broader solution space at $C_J = 3.00088$

The global clustering result is examined across the entire dataset, which includes initial perigees with various values of θ at a fixed value of the Jacobi constant, $C_J = 3.00088$. A subset of clusters is selected from those that appear in Figs. 3 and 4. Figure 6 displays for each of seven clusters, organized along the horizontal axis: a) a representative trajectory, evaluated as the medoid of the associated members that exist at each value of θ, labeled on the vertical axis; and b) the region of existence for the cluster in the configuration space. If a cluster does not possess members at a specific value of θ, the associated cell in Fig. 6 is shaded in dark gray. In Fig. 6b, the selected map crossings are plotted in the same shades of

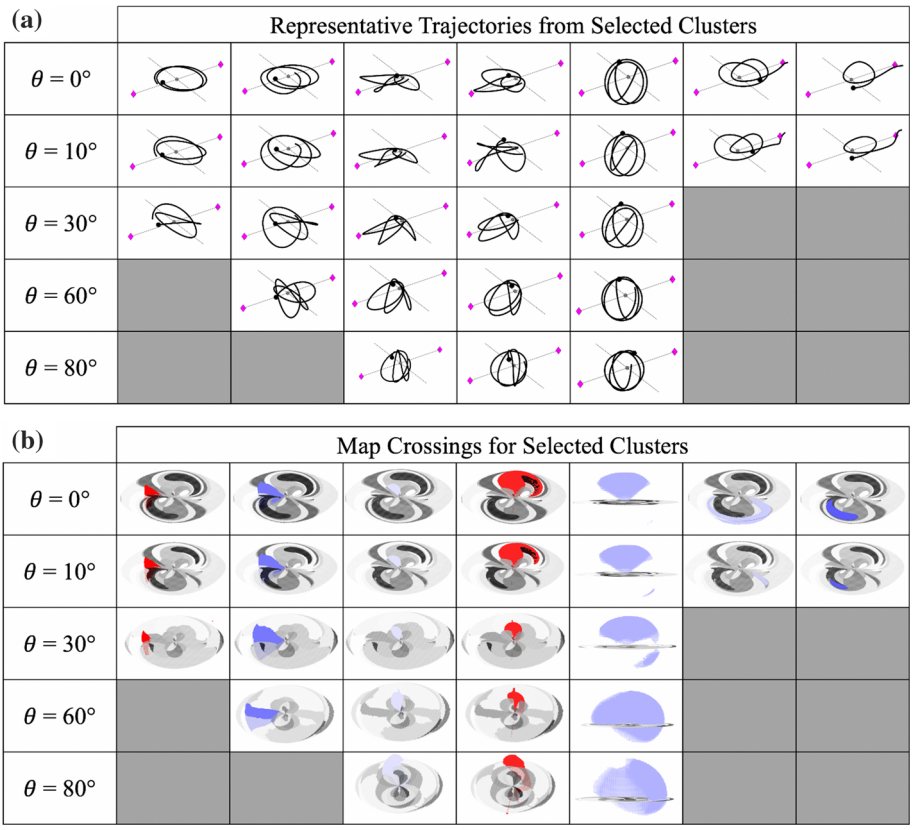

Fig. 6 Existence of selected clusters of prograde perigees at $C_J = 3.00088$ that produce geometrically similar trajectories across partitions described by specific values of θ: **a** representative trajectories generated from selected perigees in those partitions, and **b** the associated region of existence of the initial perigees colored in distinct shades of red and blue according to the cluster

red and blue that appear in Figs. 3 and 4 to indicate the cluster assignment. Overlaid in Fig. 6b in semi-transparent markers are perigees with $z = 0$ and an initial value of $\dot{z}$ calculated via the specified value of θ; these perigees are colored in shades of gray according to their cluster, assigned by the classifier, for visual clarity.

Analysis of Fig. 6 reveals insights into the regions of existence of perigees associated with each type of trajectory across the selected values of θ. First, the cluster that is summarized in the first column of Fig. 6 captures trajectories that complete two revolutions around the Earth with a small apsidal rotation. The trajectories in the second column exhibit a larger apsidal rotation over two and a half revolutions around the Earth. While the associated medoids of these clusters, as shown in Fig. 6a, evolve as θ increases, the general shape of the clusters of initial perigees, as shown in Fig. 6b in the configuration space, does not evolve significantly across the constructed partitions. Furthermore, these clusters do not exist at all high values of θ. The middle three columns correspond to clusters that exist across all values of θ that are represented in Fig. 6. For the clusters in the third and fourth columns, the regions of existence of the initial perigees evolve away from the Earth with a significantly different geometry that spans more out-of-plane members than in-plane members as θ increases. The

cluster of trajectories with a significant out-of-plane component in the fifth column evolves to encompass a wider array of initial perigees as θ increases. In fact, at $\theta = 80°$, the cluster encompasses a large array of initial perigees with positive and negative z-components. As the value of θ decreases toward zero, the regions of perigees contained within the cluster above and below the xy-plane shrink to form two distinct groupings. Finally, the trajectories that correspond to transits through the L_2 gateway before completing three revolutions around the Earth appear in the final two columns of this figure. These transits only occur for perigees close to the xy-plane and at low values of θ; their regions of existence shrink significantly as θ increases. A similar, straightforward analysis may be performed for other clusters associated with trajectories admitting geometries of interest.

The presented unsupervised organization of a large, higher-dimensional set of perigees by the geometry of the associated trajectories supports examination of a complex solution space. The global clustering result offers a summary of the fundamental geometries of arcs within the set via the representative trajectories for each cluster. This summary is constructed without burdening an analyst to manually group trajectories based on direct observations; a task that may become quite time-consuming for large sets of trajectories. In addition, this approach does not rely on predefined separation criteria that are often challenging to define in a generalizable manner for trajectories in regions of distinct sensitivities within a multi-body system. In addition, a map displaying only initial perigees, colored by their cluster assignment, supplies the analyst with valuable information about the type of trajectories associated with each map crossing; information that is otherwise challenging to extract from a Poincaré map. This visualization also enables the analyst to view only selected clusters that reflect trajectories with specific geometries, thereby mitigating the impact of data obscuration that occurs when projecting four-dimensional data onto a three-dimensional map.

9 Application: examining fundamental transport mechanisms

At a Jacobi constant of $C_J = 3.00088$ in the Sun-Earth CR3BP, trajectories that lie within the boundaries of the global stable and unstable manifolds of periodic orbits and tori near L_1 and L_2 transit through the L_1 and L_2 gateways, respectively (Delshams et al. 2016; Koon et al. 2011). As a result, these manifolds contribute valuable insight into the global structure of the phase space and natural transport between distinct regions of a multi-body system. Arcs along these stable and unstable manifolds that are generated for shorter time intervals than needed to transit through the L_1 and L_2 gateways also supply insight into the mechanisms that influence the geometries admitted by the solution space in the Sun-Earth CR3BP at the same energy level. Correspondingly, arcs along these manifolds are also often used to construct complex transfers in a multi-body system (Koon et al. 2011). Despite their importance, these complex four-dimensional manifolds are challenging to visualize either in the configuration space or via a Poincaré map. Such challenges in visualization and analysis may be mitigated by grouping arcs that lie along these manifolds according to their geometry.

This section focuses on leveraging the global cluster summary to rapidly organize arcs along the global stable and unstable manifolds associated with tori near L_1 and L_2 at a Jacobi constant of $C_J = 3.00088$ into groups according to their geometry. Specifically, these arcs are projected onto the global clustering result using the classifier constructed in Sect. 8. The resulting groupings are examined in this section in the context of two examples: 1) rapidly associating the arcs along the higher-dimensional stable manifold with the trajectories that they influence in the solution space over short time intervals; and 2) visualizing the geometry

of arcs that lie along the higher-dimensional unstable manifold. These examples demonstrate additional applications of the global cluster summary that is constructed via a data-driven approach in Sect. 8. The resulting insight into the fundamental geometries of arcs within these complex sets may eventually support initial guess construction for trajectory design and analysis of the dynamical mechanisms governing natural transport; both goals are the focus of ongoing work.

9.1 Projecting stable and unstable manifold arcs onto the global cluster summary

The first step in constructing the datasets used in this section is to generate the simply periodic orbits that exist near L_1 and L_2 in the Sun-Earth CR3BP. At a fixed value of the Jacobi constant, $C_J = 3.00088$, and in each of the L_1 and L_2 gateways, periodic Lyapunov and vertical orbits exist; at this Jacobi constant, neither the halo nor axial orbits exist in this system. These Lyapunov and vertical orbits are computed using a multiple-shooting scheme, formulated with constraints to enforce periodicity and the associated value of the Jacobi constant (Bosanac 2016). A stability analysis of each periodic orbit is performed by studying the eigenvalues of the monodromy matrix associated with a single fixed point (Koon et al. 2011). Each periodic orbit admits one set of stable and unstable modes and one set of oscillatory modes, indicating the existence of stable and unstable manifolds as well as nearby quasi-periodic orbits that trace out the surface of invariant tori. The initial conditions, orbit period and stable and unstable eigenvalues of the monodromy matrix for the Lyapunov and vertical orbits in each of the L_1 and L_2 gateways are listed in "Appendix 2."

The one-parameter family of tori that connect the Lyapunov and vertical orbits in each of the L_1 and L_2 gateways at $C_J = 3.00088$ are then generated using an approach presented by Olikara and Scheeres (Olikara and Scheeres 2012; Gómez et al. 2003). This approach, which leverages previous work by Jorba and Gómez and Mondelo, focuses on directly computing the underlying torus that is traced out by a quasi-periodic orbit (Jorba 2001; Gómez and Mondelo 2001; Olikara and Scheeres 2012). At the specified Jacobi constant, a single torus is computed near each of the L_1 and L_2 Lyapunov orbits via a multiple-shooting formulation of this algorithm. Additional tori that exist at the same Jacobi constant along each family are computed using pseudo-arclength continuation until reaching the nearby vertical orbits (Olikara and Scheeres 2012). These two families of 2-tori in each of the L_1 and L_2 gateways are then subsampled to retain only 50 members. A subset of the 50 tori in each family is shown in Fig. 7 in the Sun-Earth rotating frame with each torus colored to reflect its z-amplitude: black tori lie close to the Lyapunov orbits, while copper tori resemble the vertical orbits. In this figure, segments of the zero velocity surfaces are overlaid in blue for dimensional perspective.

The procedure presented by Olikara and Scheeres also enables a stability analysis of each torus, consistent with an approach previously presented by Jorba (Jorba 2001; Olikara and Scheeres 2012). Stability analysis of these tori near L_1 and L_2 reveals that they admit stable and unstable manifolds. As a result, segments of the three-dimensional stable and unstable manifolds associated with each of these 2-tori are generated to construct sets of trajectories that lie on the four-dimensional stable and unstable manifolds associated with the family of tori at this single energy level.

Segments of the global stable and unstable manifolds are generated for each of the tori computed near L_1 and L_2. To generate an approximation of the global stable manifold associated with each torus, the stable eigenvector associated with the differential of the invariance condition used in Olikara and Scheeres' algorithm is calculated for 12,525 states

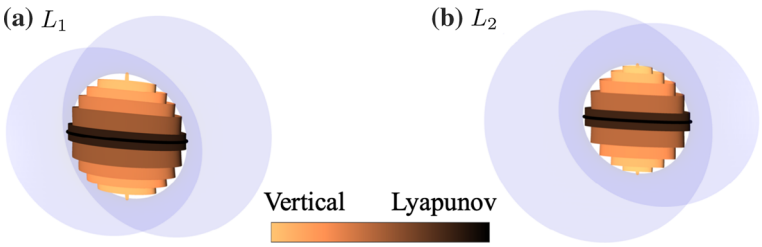

Fig. 7 Families of tori connecting the Lyapunov and vertical orbits near **a** L_1 and **b** L_2 in the Sun-Earth CR3BP at $C_J = 3.00088$

distributed over the entire torus. A state that lies on a torus is perturbed along the eigenvector associated with the stable mode and propagated backward in time for up to four subsequent perigees in the Earth vicinity to generate a trajectory in the global stable manifold. The perigees along a single trajectory admit velocity vectors that are described by values of θ that increase with each subsequent revolution around the Earth, all lying within the approximate range $[-22.2°, 22.6°]$ over the specified number of subsequent perigees. These perigees supply distinct initial conditions for arcs that lie along the stable manifold; when propagated forward in time, some arcs reach the L_1 or L_2 gateways before completing three revolutions around the Earth, while some arcs remain within the Earth vicinity over this time interval (Conley 1968). Each of these arcs that is seeded from a single trajectory along the global stable manifold associated with a torus is described by a 35-dimensional feature vector, constructed as defined in Sect. 7.2, when propagated forward in time. A similar procedure is used to generate the unstable manifold from the eigenvectors associated with the unstable mode and propagating forward in time.

The datasets capturing trajectories that lie along the global stable and unstable manifolds of the computed tori within each of the L_1 and L_2 gateways are projected onto the global cluster summary constructed in Sect. 8 by applying the classifier. The result is a set of labels for each arc, indicating the cluster of trajectories with a similar geometry. Manually performing a similar association between two large sets of trajectories via four-dimensional perigee maps would be time-consuming and challenging for a human analyst. Thus, the following subsections leverage the presented data-driven framework to study both the geometry of arcs along the hyperbolic invariant manifolds of tori in the L_1 and L_2 gateways and the nearby trajectories that they influence over a short time interval.

9.2 Analyzing the stable manifolds of tori near L_1 and L_2

Trajectories that lie along the computed stable manifolds and reach the L_1 and L_2 gateways within less than three revolutions are analyzed to support verification of both the association process and the constructed clusters. Figure 8 displays this subset of the initial perigees associated with arcs along the stable manifolds of tori near L_1 and L_2 that are assigned to clusters of trajectories with a similar geometry. The center of this figure displays the specific clusters that these perigees are assigned to using the global cluster summary constructed in Sect. 8. Each of the perigees is projected onto the three-dimensional configuration space and colored in unique shades of red and blue, consistent with the color scheme used in Figs. 3 and 4 to reflect their cluster assignment; overlaid on this figure using semi-transparent markers are the clustered map crossings with $\dot{z} = z = 0$. The insets of this figure then supply

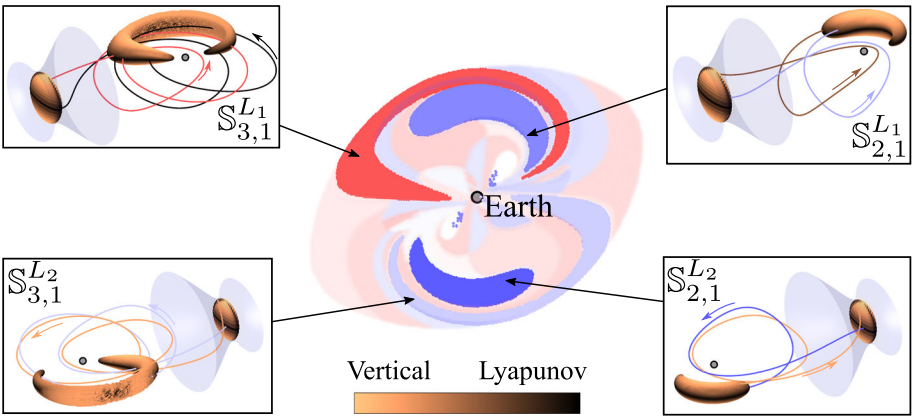

Fig. 8 Selected perigees along the global stable manifolds of tori near L_1 and L_2 in the Sun-Earth CR3BP at $C_J = 3.00088$, producing trajectories that reach the gateways before completing three revolutions around the Earth. The center of the figure projects these perigees onto the global clustering result with clusters colored in distinct shades of red and blue. The insets depict the perigees and geometry of the associated arcs, colored by the tori they asymptotically approach

additional information about each subset of perigees along the global stable manifolds of tori in either the L_1 or L_2 gateways that possess a similar geometry to each of the indicated clusters. Each inset is labeled using the convention $\mathbb{S}^{L_i}_{r,q}$: $\mathbb{S}$ indicates the perigees lie along the stable manifold, the superscript indicates that this manifold is associated with tori near the L_i equilibrium point, r records the number of returns to the perigee surface of section before reaching the torus, and q is used to differentiate the subsets of perigees for each return number r that are associated to distinct clusters. The perigees displayed in each inset are also projected onto the configuration space and colored with the copper to black color scheme used in Fig. 7 to indicate the torus that the associated trajectories approach in forward time; these tori are also plotted in the L_1 or L_2 gateways using the same color scheme. Then, one representative arc, generated by propagating a single perigee forward in time, is plotted using this copper to black color scheme for visualization of its geometry. Overlaid on these insets are representative trajectories of the associated clusters, evaluated as the medoid of trajectories that lie in the $\dot{z} = \theta = 0$ partition, and colored by the assigned cluster using distinct shades of red and blue. Finally, the Earth is plotted within this figure as a gray circle and segments of the zero velocity surfaces are displayed in light blue for dimensional perspective.

The results shown in Fig. 8 are consistent with the expected association between segments of the stable manifolds that reach the L_1 and L_2 gateways within less than three revolutions and clusters of short-term transit trajectories. The perigees displayed in the top and bottom insets of Fig. 8 correspond to the first and second returns of the stable manifolds associated with tori near L_1 and L_2, respectively, to the perigee map in backward time. Each set of perigees encompasses a similar region in the configuration space as short-term transit trajectories within the global cluster summary that possess a similar geometry, with a consistent range of values for θ. These observations are consistent with existing knowledge that trajectories within the stable manifolds of the tori near L_1 and L_2 transit through the gateways (Koon et al. 2011; Davis 2011). However, in this paper, this association is recovered in a data-driven approach.

Segments of the stable manifolds of tori in the L_1 and L_2 gateways are also examined to determine their association with trajectories that remain within the Earth vicinity for three

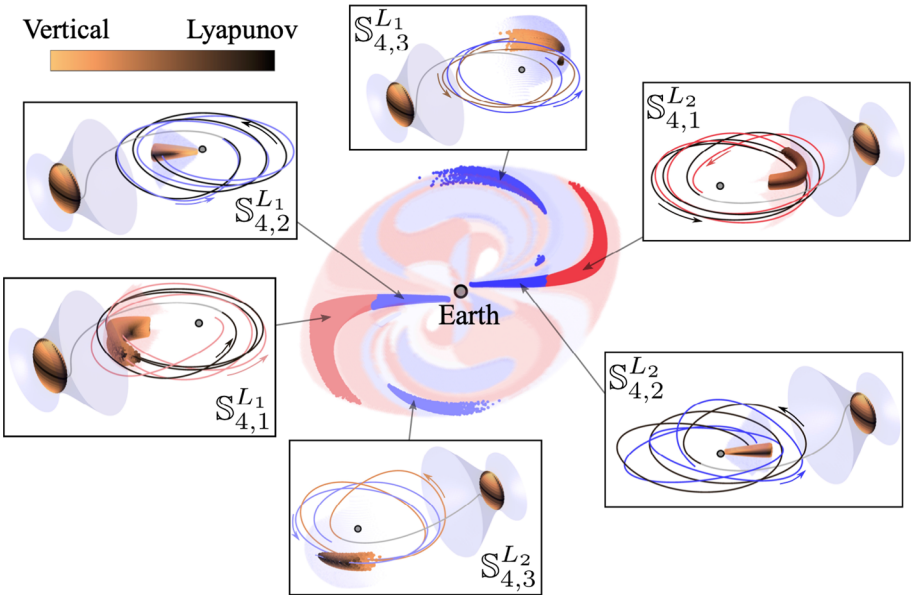

Fig. 9 Selected perigees along the global stable manifolds of tori near L_1 and L_2 in the Sun-Earth CR3BP at $C_J = 3.00088$, producing trajectories that remain within the Earth vicinity for three revolutions. The center of the figure projects these perigees onto the global clustering result with clusters colored in distinct shades of red and blue. The insets depict the perigees and geometry of the associated arcs, colored by the tori they eventually asymptotically approach

revolutions. Accordingly, Fig. 9 displays the initial perigees of these arcs along the computed stable manifolds that are associated with clusters of trajectories with a similar geometry. This figure possesses a configuration that is consistent with Fig. 8. However, the selected trajectories in the inset figures that lie along the computed stable manifolds are extended until reaching the associated torus in the L_1 and L_2 gateways; these additional arcs are colored in gray.

Trajectories that lie along the stable manifold of tori in the L_1 and L_2 gateways at the selected Jacobi constant appear to influence the geometry of a subset of temporarily captured trajectories near the Earth. Each set of perigees encompasses a subset of the region of the configuration space spanned by these clusters of temporarily captured trajectories. Only a subset of each cluster lies directly within the stable manifolds and will eventually pass through the L_1 or L_2 gateway when propagated for more subsequent perigees than used to construct the dataset in Sect. 8. However, over the short time interval that is considered in this dataset, the trajectories that are associated with additional nearby perigees inherit the geometry of the associated segments of the spatial stable manifolds of tori in the L_1 and L_2 gateways; exhibiting a distinct evolution of the location and distance of the subsequent apses across distinct clusters (Davis 2011). This example demonstrates the capability to project additional data onto the global cluster summary and, potentially, study the natural transport mechanisms that influence the characteristics of the wider solution space in a chaotic dynamical system—without requiring specification of any generalizable separation criteria or heavily burdening a human analyst. This example also supports future analyses focused on: 1) comparing a wider variety of fundamental solutions (e.g., various periodic orbits or tori and their stable or unstable manifolds) to clusters of geometrically similar trajectories with initial perigees

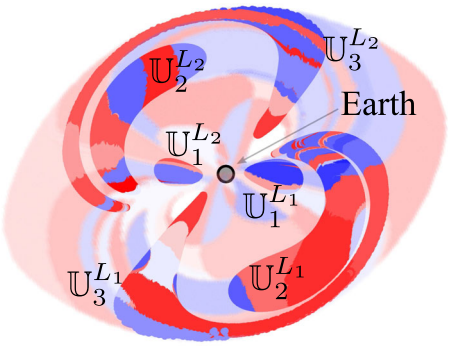

Fig. 10 Selected perigees along the global unstable manifolds of tori near L_1 and L_2 in the Sun-Earth CR3BP at $C_J = 3.00088$, producing trajectories that remain within the Earth vicinity for up to three revolutions. These perigees are projected onto the global clustering result, with each lobe labeled by the crossing number and clusters colored in distinct shades of red and blue

across the full six-dimensional phase space and 2) associating fundamental solutions that exist in low-fidelity models such as the CR3BP to the characteristics of the solution space in higher-fidelity models.

9.3 Analyzing the unstable manifolds of tori near L_1 and L_2

Unstable and stable manifolds of periodic orbits and tori are often used to design initial guesses for complex trajectories within multi-body systems. During this process, analysis of the fundamental geometries of arcs along these stable and unstable manifolds supports: construction of end-to-end trajectories with fundamentally different geometries, heuristic analysis of the region of existence of those arcs, as well as a comparison to relevant constraints (Koon et al. 2011; Davis 2011; Haapala 2014). However, when studying the four-dimensional global stable or unstable manifolds of a family of tori in the six-dimensional phase space, visualization and analysis of trajectories within this set are challenging for the human analyst; this challenge motivates the example presented in this subsection.

Segments of the unstable manifold of tori in the L_1 and L_2 gateways are projected onto the precomputed global clustering result to rapidly extract their geometries. For this example, consider only trajectories associated with the first three perigees occurring in the Earth vicinity. Figure 10 displays a projection of these perigees onto the configuration space, colored in shades of red and blue according to their cluster assignment and consistent with Figs. 3 and 4. These initial perigees are displayed and labeled using a similar convention as in Fig. 8, except with the letter $\mathbb{U}$ indicating the perigees lie along the unstable manifold. Each lobe, associated with a single crossing of one of the unstable manifolds of tori in the L_1 and L_2 gateways, is composed of initial perigees of arcs with a variety of distinct geometries, as indicated by the nonuniform coloring.

Projecting arcs along the unstable manifolds of tori in the L_1 and L_2 gateways onto the global cluster summary supplies a rapidly generated geometric grouping that facilitates visualization and analysis. To demonstrate this application of the precomputed global cluster summary, consider an alternative representation of each segment of the unstable manifolds of tori in the L_1 and L_2 gateways: one that summarizes the subsets of arcs associated with perigees within each lobe of the map in Fig. 10 based on their assignments to distinct clusters. This information is shown in Figs. 11 and 12 for the unstable manifolds of tori in the L_1 and L_2 gateways, respectively. The first rows of these figures display the medoid of each subset of perigees, colored in distinct shades of red and blue according to the cluster assignment. Each column in these figures focuses on the trajectories associated with perigees in a single lobe

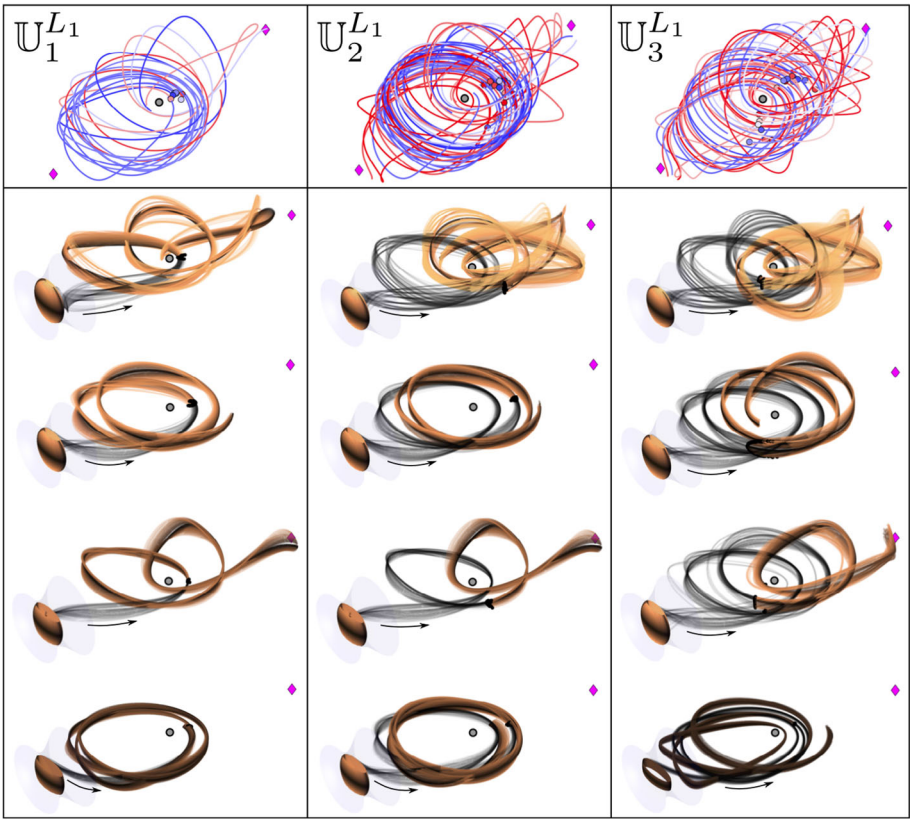

Fig. 11 Sample trajectories associated with the first three perigees of the unstable manifold of the family of tori near L_1 in the Sun-Earth CR3BP at $C_J = 3.00088$. The first row displays the distinct types of trajectories associated with each return to the perigee map colored in distinct shades of red and blue according to the cluster. The remaining rows display a subset of the trajectories with selected geometries colored by the associated tori

of the unstable manifold. The finite number of medoids displayed in the first rows of Figs. 11 and 12 supports visualization of the finite geometries admitted by arcs along the unstable manifold of tori in the L_1 and L_2 gateways. The remaining rows of Figs. 11 and 12 display a sample of the trajectories within selected subsets of each lobe. These trajectories, which are assigned to the same cluster, are colored in shades of black to copper, consistent with the color scheme in Fig. 7 that reflects the associated torus that they approach in backward time. The initial condition along each trajectory is indicated by a black circle, while the earlier arcs emanating from the associated tori are depicted in gray, and a black arrow indicates direction of motion. The subset of tori associated with trajectories of a specific geometry are also displayed using the same black to copper color scheme. Note that groups of trajectories in the same row but different columns do not necessarily correspond to the same cluster or geometry.

The data-driven grouping of perigees along the unstable manifold, as performed by projecting the associated data onto the global cluster summary, supplies valuable insight into: the array of geometries associated with arcs along the unstable manifold, their regions of existence in the configuration space, and their existence as a function of the originating tori.

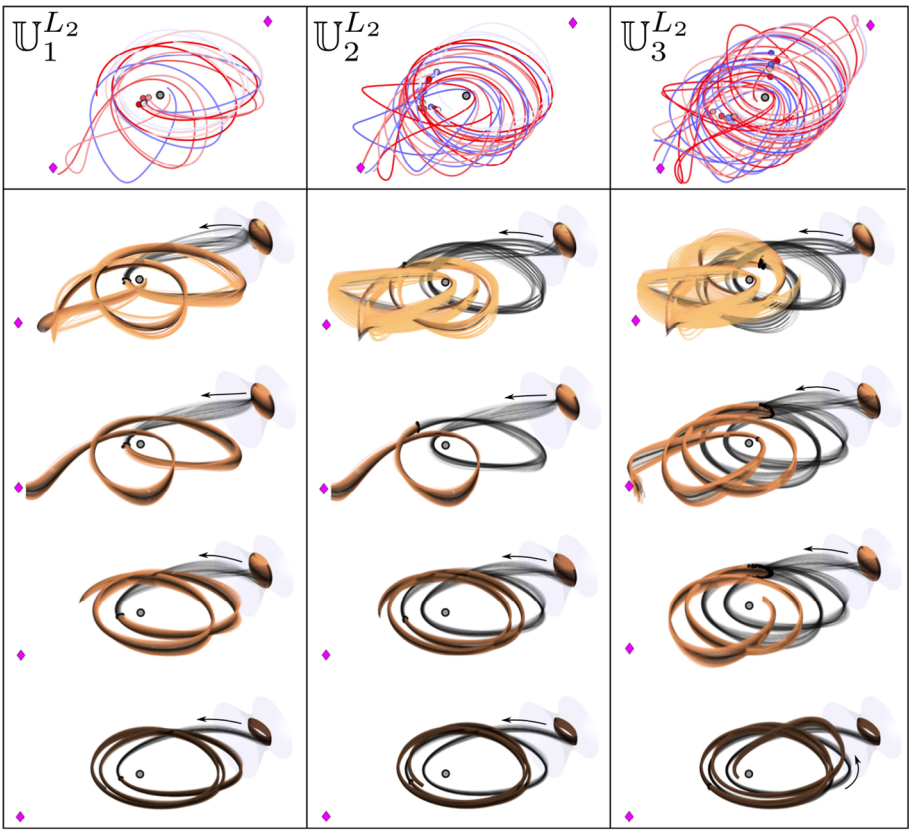

Fig. 12 Sample trajectories associated with the first three perigees of the unstable manifold of the family of tori near L_2 in the Sun-Earth CR3BP at $C_J = 3.00088$. The first row displays the distinct types of trajectories associated with each return to the perigee map colored in distinct shades of red and blue according to the cluster. The remaining rows display a subset of the trajectories with selected geometries colored by the associated tori

For instance, trajectories in the clusters in the second rows of Figs. 11 and 12 are associated with unstable manifolds emanating from the complete array of tori within each of the L_1 and L_2 gateways, respectively; this observation is a consequence of the array of copper to black trajectories that appear. However, the subsets of trajectories in the final rows of Figs. 11 and 12 are associated with the unstable manifolds of tori with a limited range of low z-amplitudes in the L_1 and L_2 gateways; this observation is a consequence of the appearance of only dark brown to black trajectories. Of course, an interactive visualization environment may further simplify visualization and analysis of these results as a human interactively examines representative trajectories and selects specific subsets of perigees and their associated trajectories for further analysis. Nevertheless, the static representations in this paper sufficiently demonstrate the value of the global cluster summary in facilitating a data-driven extraction of the fundamental geometries exhibited by members of unseen, complex datasets; in this case, arcs associated with natural transport mechanisms.

10 Conclusions

Data mining techniques are used to group the crossings on a higher-dimensional Poincaré map according to the geometry of the associated trajectories. The goal of the resulting unsupervised organization and summarization of a complex solution space is to facilitate analysis and support trajectory design tasks. This data-driven approach is developed in the context of a perigee map in the spatial CR3BP, extending previous work by Bosanac in the planar CR3BP (Bosanac 2020). First, a large dataset of initial conditions associated with spatial perigees is generated and partitioned. The trajectory associated with each initial condition is propagated for a specified number of subsequent perigees and summarized by a finite-dimensional feature vector; the difference between two feature vectors is used to assess geometric similarity between two trajectories. The map crossings within each individual partition are then clustered using HDBSCAN. Inspired by the field of distributed clustering, these local clustering results are input to a cluster aggregation procedure. Specifically, the clusters associated with each partition are successively summarized and combined in a binary tree structure using both HDBSCAN for clustering and UMAP for dimension reduction. Cluster aggregation produces a global summary of the geometries admitted by trajectories associated with perigees that exist across the entire dataset.

The presented data-driven approach to higher-dimensional Poincaré maps is demonstrated in the context of a spatial, prograde perigee map, constructed at a single value of the Jacobi constant in the Sun-Earth CR3BP. Following the approach presented in this paper, the perigees are successfully grouped according to the geometry of the associated spatial trajectories in an unsupervised manner. The outputs of this procedure include: 1) maps with selected clusters indicated by distinct colors, and each cluster sufficiently capturing solutions of similar geometry; 2) a set of representative trajectories that summarize the distinct geometries admitted by the solution space; and 3) a weighted k-nearest neighbor classifier that may be used to assign either members of the original dataset that do not appear in the final summary or unseen data to each cluster. As a result, the presented data-driven approach to higher-dimensional Poincaré maps may facilitate analysis of a complex solution space with a reduced burden on a human analyst, as well as support for the trajectory designer assembling arcs of specific geometries to construct an initial guess for a complex trajectory in the spatial CR3BP.

The constructed global cluster summary is also used to facilitate analysis and visualization of the arcs associated with hyperbolic invariant manifolds of families of tori in the L_1 and L_2 gateways. The stable and unstable manifolds associated with bounded motions near the collinear libration points serve as natural transport mechanisms within multi-body systems. Studying their characteristics and influence on the solution space is a significant analytical task that also supports trajectory design; both of these activities are typically performed by a human analyst. However, projection onto the precomputed global clustering result rapidly produces a summary of the geometries of arcs along the stable and unstable manifolds of tori in the L_1 and L_2 gateways. This example is used to demonstrate: 1) a data-driven association between the spatial and higher-dimensional stable manifolds of these families of tori and the trajectories that they influence within the solution space; and 2) visualization of the fundamental geometries of arcs along the spatial and higher-dimensional unstable manifolds of these families of tori.

Acknowledgements This work was completed at the University of Colorado Boulder, partially funded under NASA Grant 80NSSC18K1536. The authors thank the anonymous reviewers for their feedback.

Funding This work was completed at the University of Colorado Boulder, partially funded under NASA Grant 80NSSC18K1536.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

11 Appendix 1

See Figs. 13 and 14.

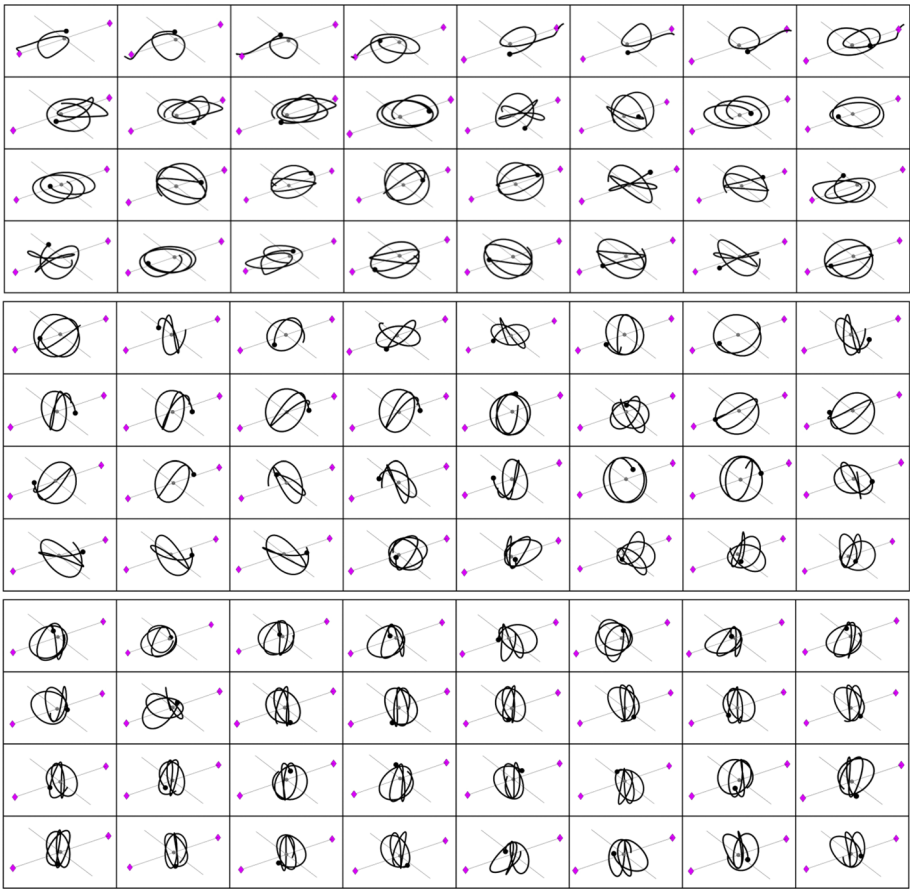

Fig. 13 Subset of representatives for global clustering result summarizing trajectories associated with prograde perigees and generated for up to three returns to a perigee map in the Sun-Earth CR3BP at $C_J = 3.00088$, with the libration points displayed as magenta diamonds

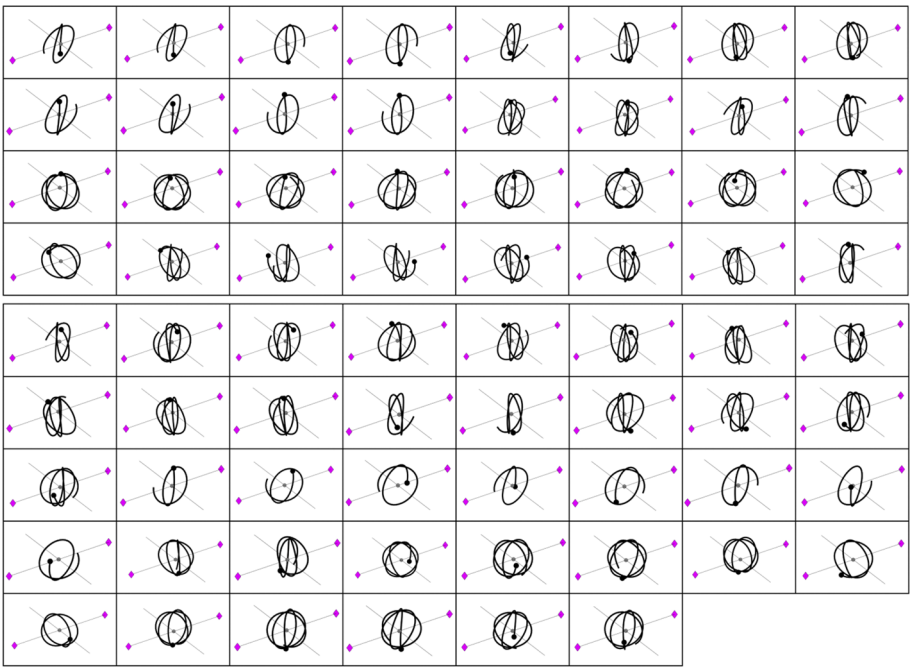

Fig. 14 Subset of representatives for global clustering result summarizing trajectories associated with prograde perigees and generated for up to three returns to a perigee map in the Sun-Earth CR3BP at $C_J = 3.00088$, with the libration points displayed as magenta diamonds

12 Appendix 2

See Table 1.

Table 1 Truncated period, initial state and stable and unstable eigenvalues, λ_S and λ_U, respectively, of the monodromy matrix for the Lyapunov and vertical orbits at L_1 and L_2 in the Sun-Earth CR3BP at $C_J = 3.00088$

	Orbit	Period [nondim]	x_0 [nondim]	Eigenvalue
L_1	Lyapunov	3.0189495	[0.9895177, 0, 0, 0, 0.0036028, 0]	$\lambda_S = 5.00 \times 10^{-4}$, $\lambda_U = 2004$
	Vertical	3.1247260	[0.9900629, 0, 0, 0, 0.0000726, 0.0032712]	$\lambda_S = 3.81 \times 10^{-4}$, $\lambda_U = 2626$
L_2	Lyapunov	3.0588881	[1.0095682, 0, 0, 0, 0.0029462, 0]	$\lambda_S = 5.14 \times 10^{-4}$, $\lambda_U = 1946$
	Vertical	3.1691039	[1.0100107, 0, 0, 0, −0.0000473, 0.0025868]	$\lambda_S = 3.90 \times 10^{-4}$, $\lambda_U = 2563$

References

Aggarwal, C., Reddy, C.: Data Clustering: Algorithms and Applications. Ch. 2.1.2, 3, 9.2, Chapman and Hall CRC (2018)

Ali, M., Jones, M., Xie, X., Williams, M.: TimeCluster: dimension reduction applied to temporal data for visual analytics. Vis. Comput. **35**, 1013–1026 (2019)

Becht, E., McInnes, L., Healy, J., Dutertre, C., Kwok, I., Ng, L., Ginhoux, F., Newell, E.: Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. **37**, 38–44 (2019)

Bendechache, M., Le-Khac, N., Kechadi, M.: Efficient large scale clustering based on data partitioning. In: 2016 IEEE International Conference on Data Science and Advanced Analytics, pp. 612–621, Montreal, QC, Canada (2016)

Bonasera, S., Bosanac, N.: Applications of clustering to higher-dimensional Poincaré maps in multi-body systems. In: 30th AIAA/AAS Space Flight Mechanics Meeting, Orlando, FL (2020a)

Bonasera, S., Bosanac, N.: Unsupervised learning to aid visualization of higher-dimensional Poincaré maps in multi-body trajectory design. In: 2020 AAS/AIAA Astrodynamics Specialist Conference, Lake Tahoe, CA (Virtual) (2020b)

Bosanac, N.: Leveraging natural dynamical structures to explore multi-body systems. Ph.D. thesis, Purdue University, West Lafayette, IN (2016)

Bosanac, N.: Data mining approach to Poincaré maps in multi-body trajectory design. J. Guid. Control Dyn. **43**(6), 1190–1200 (2020)

Bosanac, N., Cox, A., Howell, K., Folta, D.: Trajectory design for a cislunar cubesat leveraging dynamical systems techniques: the Lunar IceCube mission. Acta Astronaut. **144**, 283–296 (2018)

Campello RJGB, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G (eds) Advances in knowledge discovery and data mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D., Hill, A., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F., Trapnell, C., Shendure, J.: The single-cell transcriptional landscape of mammalian organogenesis. Nature **566**, 496–502 (2019)

Cichosz, P.: Data Mining Algorithms: Explained Using R. John Wiley and Sons, West Sussex (2015)

Conley, C.: Low energy transit orbits in the restricted three-body problem. SIAM J. Appl. Math. **16**(4), 732–746 (1968)

Contopoulos, G.: Order and Chaos in Dynamical Astronomy. Springer, Berlin (2002)

Davis, D.C.: Multi-body trajectory design strategies based on periapsis Poincaré maps. Ph.D. thesis, Purdue University, West Lafayette, IN (2011)

Davis, D., Phillips, S., McCarthy, B.: Trajectory design for saturnian ocean worlds orbiters using multidimensional Poincaré maps. Acta Astronaut. **143**, 16–28 (2018)

Delshams, A., Gidea, M., Roldan, P.: Arnold's mechanism of diffusion in the spatial circular restricted three-body problem: a semi-analytical argument. Phys. D Nonlinear Phenom. **334**, 29–48 (2016)

Gómez, G., Mondelo, J.: The dynamics around the collinear equilibrium points of the RTBP. Physica D **157**(4), 283–321 (2001)

Gómez, G., Masdemont, J., Mondelo, J.M.: Libration point orbits: a survey from the dynamical point of view. In: Proceedings of the Libration Point Orbits and Applications, Aiguablava, Spain (2003)

Gómez, G., Koon, W., Lo, M., Marsden, J., Masdemont, J., Ross, S.: Connecting orbits and invariant manifolds in the spatial restricted three-body problem. Nonlinearity **17**, 1571–1606 (2004)

Haapala, A.: Trajectory design in the spatial circular restricted three-body problem exploiting higher-dimensional Poincare maps. PhD thesis, Purdue University, West Lafayette, IN (2014)

Hadjighasem, A., Karrasch, D., Teramoto, H., Haller, G:. Spectral-clustering approach to Lagrangian vortex detection. Phys. Rev. E **93**(6) (2016)

Han, J., Kamber, M.: Data mining: concepts and techniques, Second Edition. Ch.7, Proquest EBook Central: Elsevier Science and Technology, Waltham, MA (2006)

Han, J., Kamber, M., Pei, J.: Data mining concepts and techniques, Third Edition. Ch. 9.5, Morgan Kaufmann (2014)

Ivezić, Z., Connolly, A., VanderPlas, J., Gray, A.: Statistics, data mining, and machine learning in astronomy: a practical python guide for the analysis of survey data, updated edition. Ch. 9.4, Princeton University Press (2019)

Jorba, Á.: Numerical computation of the normal behavior of invariant curves of n-dimensional maps. Nonlinearity **14**(5), 943–976 (2001)

Koon, W.S., Lo, M.W., Marsden, J.E., Ross, S.D.: Dynamical Systems, the Three Body Problem and Space Mission Design. Marsden Books, New-York (2011)

Li, X., Dyck, O., Oxley, M., Lupini, A., McInnes, L., Healy, J., Jesse, S., Kalinin, S.: Manifold Learning of four-dimensional scanning transmission electron microscopy. npj Comput. Mater., **5** (2019)
MathWorks MATLAB. Natick, MA, USA (2020)
McInnes, L., Healy, J., Astels, S.: hdbscan: hierarchical density based clustering. J. Open Source Softw., **2**(11) (2017)
McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv e-prints. **1802**, 03426 (2018)
Mommert, M., Trilling, D., Hora, J., Lejoly, C., Gustafsson, A., Knight, M., Moskovitz, N., Smith, H.: Systematic characterization of and search for activity in potentially active asteroids. Planet. Sci. J. **1**(1), 1–14 (2020)
Moulavi, D., Jaskowiak, P., Campello, R., Zimek, A., Sander, J.: Density-based cluster validation. In: Proceedings of the SIAM International Conference on Data Mining, Philadelphia, PA (2014)
Nakhjiri, N., Villac, B.F.: Automated stable region generation, detection, and representation for applications to mission design. Celest. Mech. Dyn. Astron. **123**(1), 63–83 (2015)
Olikara, Z., Scheeres, D.: Numerical methods for computing quasi-periodic orbits and their stability in the restricted three-body problem. In: IAA Conference on Dynamics and Control of Space Systems, Porto, Portugal (2012)
Paskowitz, M., Scheeres, D.: Robust capture and transfer trajectories for planetary satellite orbiters. J. Guid. Control Dyn. **29**(2), 342–353 (2006)
Perko, L.: Differential Equations and Dynamical Systems, 2nd edn. Springer, New-York (1996)
Smith, T., Bosanac, N.: Constructing a set of motion primitives in the circular restricted three-body problem via clustering. In: AAS/AIAA Astrodynamics Specialist Conference, Portland, ME (2019)
Szebehely, V.: Theory of Orbits: The Restricted Problem of Three Bodies. Academic Press, London (1967)
Verhulst, F.: Nonlinear Differential Equations and Dynamical Systems. Springer, Berlin (1996)
Villac, B., Scheeres, D.: Escaping trajectories in the Hill three-body problem and applications. J. Guid. Control Dyn. **26**(2), 224–232 (2003)
Villac, B., Scheeres, D.: On the concept of periapsis in Hill's problem. Celest. Mech. Dyn. Astronaut. **90**, 165–178 (2004)
Villac, B., Anderson, R., Pini, A.: Computer aided ballistic orbit classification around small bodies. J. Astronaut. Sci. **63**(3), 175–205 (2016)
Wenskovitch, J., Crandell, I., Ramakrishnan, N., House, L., Leman, S., North, C.: Towards a systematic combination of dimension reduction and clustering in visual analytics. IEEE Trans. Vis. Comput. Gr.aph **24**(1), 131–141 (2018)
Zheng, Y. Zhou, X.: Computing with Spatial Trajectories. Ch. 2, Springer New York (2011)