ASSESSMENT OF INFORMATION CONTENT CONTAINED IN OBSERVED PHOTOMETRIC SIGNATURES OF NON-RESOLVED SPACE DEBRIS OBJECTS

Islam Hussein, Thomas Kelecy, Kyle Miller, Matthew P. Wilkins, Chris Roscoe and Mark Bolden

Applied Defence Solutions, 10440 Little Patuxent Pkwy STE 600, Columbia, MD, USA, Emails: ihussein@applieddefense.com, tkelecy@applieddefense.com, kmiller@applieddefense.com, mwilkins@applieddefense.com, croscoe@applieddefense.com, mbolden@applieddefense.com

ABSTRACT

The information content of light curves is examined using an innovative "functional" approach, as opposed to the time and frequency domain analyses now being applied, to characterize physical and dynamic attributes (information) of orbital debris from non-resolved photometric observations. The information content is examined in a probabilistic context where a set of simulated light curves for a diverse set of object shapes, sizes and dynamics are used to demonstrate the application and value of *Functional Data Analysis*, *Clustering* and *Information Theory*. The results confirm the value of these approaches by correctly categorizing independent sets of light curve measurements and quantifying the reliability of a given light curve being associated with a specific object. These analytical tools can also be applied to better understand how well a given set of observations characterizes an object and, hence, guides the necessity of future observations.

1   INTRODUCTION

Photometric signatures are the time histories of visible light reflected from Earth orbiting space objects, such as debris, as observed by a ground or space-based optical sensor. They are a function of the relative geometry between the sun-object-observer, object size, shape, material makeup and attitude dynamics. The attitude dynamics are coupled to the orbital dynamics – e.g. through area-to-mass ratio variations which affect solar radiation pressure and drag – and hence, through dynamic mismodeling, affects our ability to accurately track and predict the orbital motion of uncontrolled debris. The challenge is that signatures from a group of detected objects may be collected in one or more frames over different time spans on objects that have not been correlated to a catalogued object. How do we know which sets of photometric signatures are associated with the same object? … or to a different object? The analysis presented here attempts to approach these questions from an information content perspective. Accurate and appropriate correlation of signatures will lead to more accurate shape and attitude retrieval, as well as more efficient tasking of observations of debris for the purpose of physical and dynamic characterization.

In space surveillance, Earth-orbiting debris objects that are too small or too far away from an optical sensor cannot be spatially resolved in the observed images. Thus, photometric time-histories of the visible radiometric signatures collected from ground and space-based sensors are now being used to extract stability, size, shape, and attitude information about objects being observed to support Space Situational Awareness (SSA). The radiometric signature is a function of the size, shape, reflective characteristics, attitude dynamics and space environment of the object, in addition to the phase angle defined by the sun-object-observer geometry. Though it is difficult to separate these physical attributes unambiguously, in particular in the absence of *a priori* information as is the case for orbital debris, one could assume that each object should have combined characteristics that are unique to the individual object being observed.

Two questions to ask are: (1) If two or more sets of signatures are collected from one or more sensor locations, and no *a priori* association between them is known, can one associate two or more of the signatures as having similar content to a quantifiable degree of certainty? (2) If one or more sets of signatures are collected from one or more sensor locations that are known to associate with a common object, how can one determine when sufficient observations have been collected to unambiguously characterize the size, shape, periodicity and attitude information content of the object data?

This work establishes a set of scenarios whereby simulated observations that encompass object size, shape attitude and reflective characteristics are generated for a variety of "typical" space debris objects and Information Theoretic algorithms are applied to assess the information content towards answering the two questions above. The results demonstrate the value of associating multiple signatures and enabling space surveillance analysts to determine when sufficient data have been collected to unambiguously determine stability, size, shape and attitude. These attributes are directly related to the attitude and orbital dynamics and so will benefit the tracking and prediction of space debris objects.

The paper is organized as follows. Section 2 describes the

problem formulation. Section 3 describes the simulation setup. Section 4 describes the functional approach where light curves are modelled as a series expansion that can be analysed for information content. Section 5 is conclusions.

2 PROBLEM STATEMENT AND ASSUMPTIONS

The photometric sensor is designed to collect reflected photons from an object orbiting the Earth. In some cases, a multi-spectral sensor is tasked to collect the reflected radiation in various wavelength bands. The signature is a result of several factors:

- In-coming light from the sun (solar flux F)

- Size of the object surfaces reflecting the light (facet area A_i)

- Direction of the object surfaces reflecting light in the body frame (facet normal n_i)

- Orientation of the object surfaces that reflect light relative to the Earth Centered Inertial (ECI) frame (attitude quaternion q)

- Phase angle geometry of the reflecting surface, i.e., the angle between the incident light from the sun and the reflected light ray going to the observer (ϕ, angle between observer vector o_i and sun vector s_i)

- Range between the object surface and the observer (range ρ)

- Atmosphere of the Earth (turbulence and refraction)

- Material makeup and reflective properties of the surfaces reflecting the light (spectral components as a function of the wavelength λ_i)

- Thermal properties of the surfaces (surface temperature component T_i as a function of input energy and material properties)

Most, if not all, of the physical attributes of debris objects are time dependent. Therefore, the photometric signature which is explicitly a function of object "features" is also an implicit function of time. For example, it is well known that aging effects of the space environment on paint can dramatically alter the reflective properties of satellites. Although the solution set contains an infinite set of parameter combinations, it is reasonable to assume that, for any single given object, the physical features are sufficiently slowly varying such that, over a reasonable observation time, the parameter solution space can be bounded. In this analysis, we will assume that space ageing affects and seasonal variations can be ignored.

Furthermore, the feature parameters are physically constrained by known material properties, sizes, shapes,

the physics associated with the dynamics. In this work, one observation pass, which may result in only a partial light curve of an object, is considered to be one observation sample. As the result of differing observation geometries exposing different surfaces at different orientations over different observing passes, a particular photometric signature observation sample can look quite different from pass to pass. Examples of this are provided in the simulated light curve cases presented in the next section. We assume that the problem as stated is observable given a sufficiently large number of samples. We have several unresolved questions: 1) How many samples are needed to uniquely characterize a single object, and with what reliability/uncertainty? 2) How reliably might we be able to distinguish unambiguously two objects, and with how many measurements? 3) Is it even possible to disambiguate between two objects? 4) How might multiple sensors viewing from different geometries (versus a single sensor) expedite characterization?

Given:

- Multiple parameter photometric signature model
- The physical/numerical constraints on these parameters
- Assumed statistical noise of the measurements is Gaussian
- Sufficiently large sample set to capture the dynamics and physical attributes

Determine:

- Parameterization models that capture the essential physical and dynamic attributes embedded in the observed light curves

- Probabilistic metrics that relate the information captured in the empirical parameterization and provide a measure for how likely a light curve signature is to be associated with an ensemble population of light curves.

Techniques are derived and applied in Section 4 for determining when a collection of observations are associated with a unique object, and which sets of observations might not be associated with the same object. But first, a set of test cases generated to test these techniques are described.

3 SIMULATED CASES

Light curve data were simulated for several cases to demonstrate the application of information theory techniques towards addressing the association questions posed in the previous section. A bi-directional reflection function (BRDF) model was used that allowed the size, shape, surface reflection properties and attitude dynamics to be specified. The visual magnitude as a function of

time was simulated for an Earth-based observer under the conditions where the object was illuminated and the sensor location was in the dark. A 5%-10% Gaussian noise was added to the simulated observations.

Four cases were simulated: 1) A simple geometric 1 *m* x 1 *m* x 1 *m* cube, 2) a simple 1 *m* x 1 *m* thin plate, 3 and 4) "box-wing" satellite models consisting of a rectangular bus, two solar arrays and two dish antennae that were tumbling or inertially fixed. For the first two cases 500 light curve realizations were generated over a 300 second span and at a 1 second sampling interval where the noise was varied. A realization of Case 1 is shown in Figure 1 where a realization of Case 2 is shown in Figure 2.
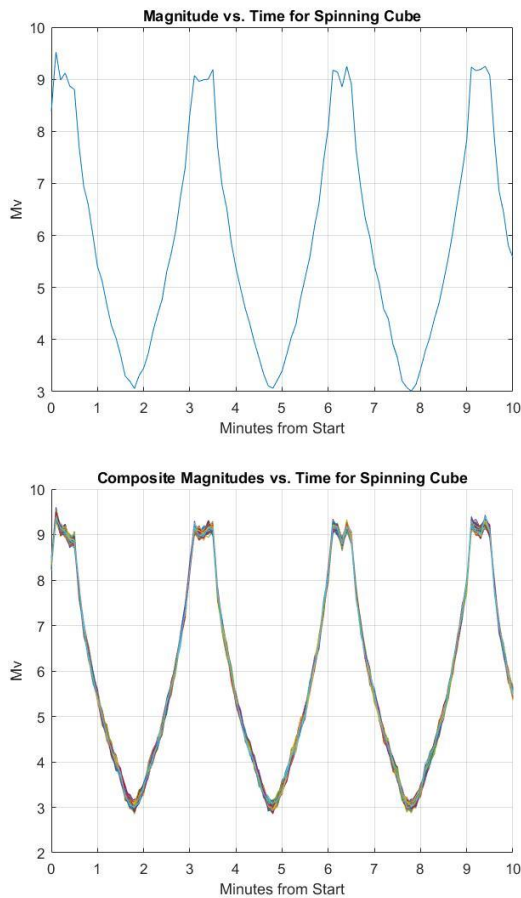

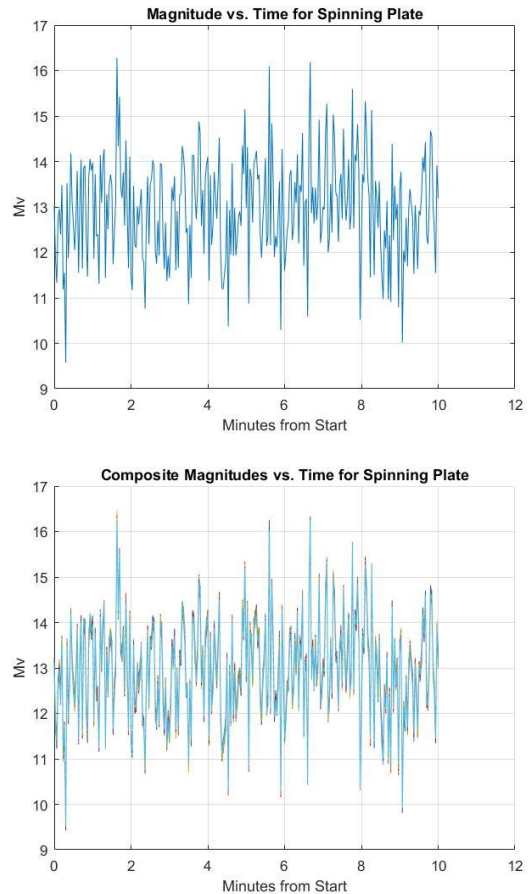*Figure 1. Spinning cube realization (upper) and composite of all realizations (lower) – Case 1*

Figure 2. Spinning plate realization (upper) and composite of all realizations (lower) – Case 2

For the third and fourth box-wing satellite model cases, 200 realizations were simulated over a span of 3 hours and at a sample interval of 30 seconds. The third case was modelled as a steady-state tumble where the initial attitude was varied for each case to model different viewing geometries for the same tumbling objects. An example realization is shown in Figure 3. The fourth case modelled the box-wing as inertially fixed, but again with the initial attitude varied for each realization. An example of the case is shown in Figure 4. In essence, Cases 3 and 4 have a more complex object shape consisting of a bus modelled as a rectangular box, two solar arrays modelled as two rectangular panels, and two dish antennae modelled as two circular plates, with each surface component having different reflective characteristics, and the light curves are "observed" from a single location but at different geometries.

Each realization within each of the cases depicts a unique object, but with noise and geometry changing from case to case, which is meant to be representative of observations of the same object taken at different times and geometries, and over different phases of the motion.
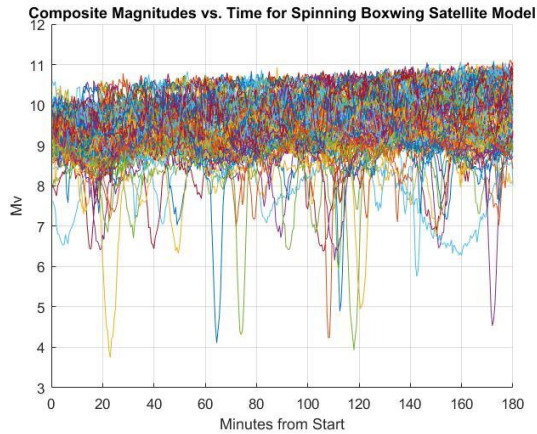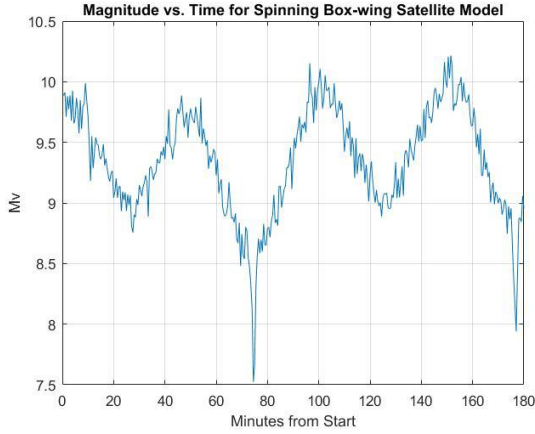
Figure 3. Spinning box-wing satellite model realization (upper) and composite of all realizations (lower) – Case 3
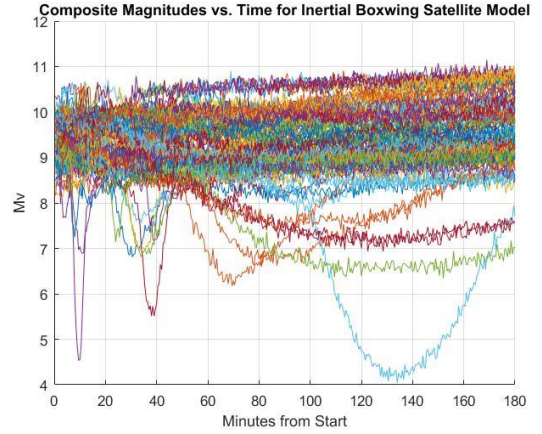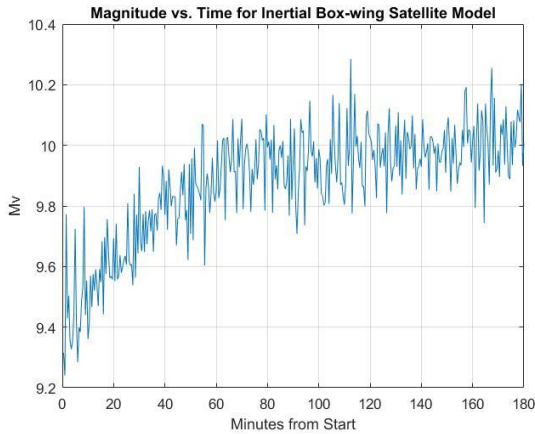

Figure 4. Inertially fixed box-wing satellite model realization (upper) and composite of all realizations (lower) – Case 4

As the four cases illustrate, light curves can be represented by a time history of observations consisting of a characteristic signature combined with both systematic and random errors resulting from imperfections intrinsic to the observing system and environment. The signature itself is of interest as its variations over time and geometry reveal the physical and dynamic attributes of the unresolved debris object. In the next section of this paper we propose an alternative representation of the light curve that is more conducive to exploring the information content and, hence, provides another analytical tool for associating independent set of observations with no *a priori* association knowledge.

4 APPLICATION AND ANALYSIS OF INFORMATION THEORY TECHNIQUES

4.1 Motivation

Towards a rigorous probabilistic analysis of light curves that would lend itself to the tools of information theory, we mathematically represent a light curve as a *function*. This approach differs from the conventional approach that treats light curves as a point-wise time series sequence of points. We begin by letting a light curve at time t over a time interval of length T be denoted by Z_t^T, object characteristic parameters by x_t, and geometric parameters by g_t. We note that only Z is treated as a curve (or model) over the interval T, and that the state and geometric parameters are assumed to be represented in a holistic way over the window of time T. For example, the state x may represent the spin state of the satellite over the interval T, including the possibility that it is executing a complex attitude motion. A joint distribution on these parameters will be denoted by $p(Z_t, x_t, g_t)$ (for ease of notation, we will drop the superscript T). Along with the joint distribution, we can use light curve models – e.g. a BRDF model – to compute the likelihood $p(Z_t|x_t, g_t)$, which was used in a pointwise (i.e., not as

a function) in the context of Multiple Adaptive Model Estimation (MMAE) [1]. The probabilistic framework developed in this paper will then enable a rigorous Bayesian analysis, where we can now compute the posterior in the object state given the light curve observation and geometry parameters $p(x_t|Z_t, g_t)$ or the joint posterior in both object state and geometric parameters given a light curve observation $p(x_t, g_t|Z_t)$.

To probabilistically analyse light curves, we propose to use tools from Functional Data Analysis (FDA) [2, 3]. Such an approach has been used in genetic analysis [4] and star and planetary classification based on an FDA analysis of light curves [5]. In FDA, a curve is converted into a finite-dimensional vector using an appropriate basis system that guarantees capturing of the main features of a light curve signal. From there on, any result that is established for finite dimensional spaces can now be applied to functional light curve data. For example, in Ref. [6], the authors use Mahalanobis distance to process functional data for classification.

In addition to a rigorous Bayesian approach to light curve analysis, one can also venture in the direction of information theory to address many SSA problems. For example, one can compute information divergence between two classes of light curve data to assess how much common information exists between the two sets of data. Such an analysis can reveal, for example, that two objects share one or more common facets. Another example is to use mutual information to solve the object-to-object correlation problem by assessing how much information overlaps between two observed objects with two light curves collected at different time instances. In other words, the proposed FDA-based probabilistic viewpoint may enable the replication of many of the information theoretic results developed for angles and/or range observations (see Ref. [7, 8]) in light curve processing.

In this section, we first describe how tools from machine learning, such as data clustering, can be used to cluster populations of light curves. We then use FDA and its associated functional Principal Component Analysis (fPCA) to obtain first and second moment statistics (the equivalent of obtaining a Gaussian model for a point cloud in particle filtering) [3]. We use this to compute the likelihood that an object that generated a test light curve belongs to a candidate class of objects.

4.2 Using Functional Data Analysis for Light Curve Data Clustering

In machine learning, clustering aims at grouping a collection of objects, where in each group the objects are similar to each other. How one defines similarity depends on the objects and a notion of a metric that measures the degree of closeness between two objects. In the present case, the objects are light curves. Representing light curves as square integrable functions on the interval T in L^2 space, one natural way to measure the distance between two curves is to first expand a light curve Z_t^T in terms of an infinite sum of basis functions such as splines or Fourier functions. The series is approximated by retaining the first N components of the sum. The N series coefficients then act as the coordinates representing the light curve in an N-dimensional vector space. The classical Euclidean N-norm can now be used to measure the distance between two light curves.

With a notion of distance defined, clustering then aims at grouping of a set of M light curves based on how close to each other the curves are. A classical algorithm is the k-means algorithm [9]. The algorithm works by first selecting a collection of k objects from the set. These k objects act as the centroids of k clusters. Each object in the set is then assigned to a cluster centroid that is closest to it in distance (cluster assignment step). The centroid of each cluster is then updated so that the new centroid is the mean of all the objects in the cluster (centroid update step). These two steps, cluster assignment and centroid update, are then repeated until no change in cluster centroids is observed. The k-means algorithm for finding the optimal clustering is NP-hard in general, but several approximations exist, such as the Lloyd algorithm, that make the solution linear in all problem parameters (the size M of the data, the dimension N of the underlying space, and the number of desired clusters k) [9]. We used a variation of the k-means algorithm, called k-means++, where the initial set of cluster centroids are probabilistically chosen in order to improve the convergence of the k-means algorithm [10], making it only $O(\log k)$.

We used the k-means++ algorithm to cluster 2400 light curves from all 4 cases. The light curves were of four main types: (1) A spinning plate, (2) a spinning cube, (3) a tumbling box-wing satellite and (4) an inertial box-wing satellite. Performance of the k-means++ algorithm is shown in Table 1. The four object types were clustered into 4 groups. We note that the cluster corresponding to the plate has all, and only, plate light curves in it. The cube cluster had 100% percent of the cube light curves, but a few inertial and tumbling light curves were also included in the cube cluster. Finally, we note that the clusters for the tumbling and inertial box-wing satellite have near 50-50 proportions of the two light curve types. This is most likely due to the identical box-wing satellite model being used for those two cases, the only difference being the attitude dynamics (one tumbling and one inertially fixed).

We note that the k-means++ algorithm requires that the user specify the number of clusters. We specified four clusters; however, future research will focus on using existing variations of the k-means++ algorithm that provide a best estimate for the number of clusters as well.

Table 1. Performance of the k-means++ clustering algorithm

Cluster #	% Plate Spinning	% Cube Spinning	% Boxwing Spinner	% Boxwing Inertial
1	100.00	0.00	0.00	0.00
2	0.00	95.00	2.27	3.03
3	0.00	0.00	51.69	48.31
4	0.00	0.00	51.00	49.00

4.3 Computing Statistics Given Clustered Light Curve Data

Once the curves are clustered, one can think of each cluster as a sample of curves drawn from the underlying object class population. Still working within the framework of the finite N-dimensional approximation of infinite dimensional functional data, one can then use the notions of sample mean and sample covariance to estimate the first two moments of each cluster. This is known in the literature as fPCA within the FDA framework These two moments can then be used to compute the likelihood that a light curve obtained at a later time belongs to one of a given set of light curve classes (i.e., clusters).

4.4 Computing Likelihood of a Light Curve

The next question we address in this paper is: How can one determine the likelihood that a light curve was generated by an object of a specific shape type? Given a candidate object's shape type, one can obtain the type's first and second moment statistics from the corresponding cluster as described in the last subsection. Results from the previous section also allow us to use Mahalanobis distance to assess whether the light curve belongs to one of the available light curve classes [6]. Such an analysis helps in answering questions such as: (1) Is the new light curve generated from an object of a give class? And short of a decisive answer, (2) how much in common does this light curve have with a given class of objects? These results also enable us to use the Kullback-Leibler (K-L) divergence, for example, to determine the degree of commonality between two classes of light curve populations. The K-L divergence between two light curve clusters C_1 and C_2 is given by:

$$D_{KL}(C_1||C_2) = \frac{1}{2}\left(tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) - N + \ln\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right)\right) \quad (1)$$

where μ_i and Σ_i are the mean and covariance of cluster C_i.

The light curve data for each case were analysed in the following way. Each data set was split into two parts where the first half of each realization was fit to a high order polynomial, as an example of a parameterized light curve model, and the mean and covariance of the polynomials was computed. A likelihood function was then generated by evaluating the second half of each realization to the sample functional mean and covariance, assuming a Gaussian distribution. This approach insured independent data sets were used for computation of the mean and covariance versus the "samples" used for evaluation of the likelihood. In this exercise, a polynomial of order 23 was used. Given that this selection was somewhat arbitrary, a more in-depth analysis should be done to evaluate appropriate orders, and even different parameterization models. One desires a model that retains the "signal" of interest (information) while minimizing noise and error related artefacts.

First, a comparison with the spinning cube samples (Case 1) from the second half of each of its realizations was compared to the mean and covariance derived from the first half of each realization. As can be seen in Figure 5, which plots the *log* of the likelihood, the large values indicate the high likelihood that the samples are related to the spinning cube statistics.
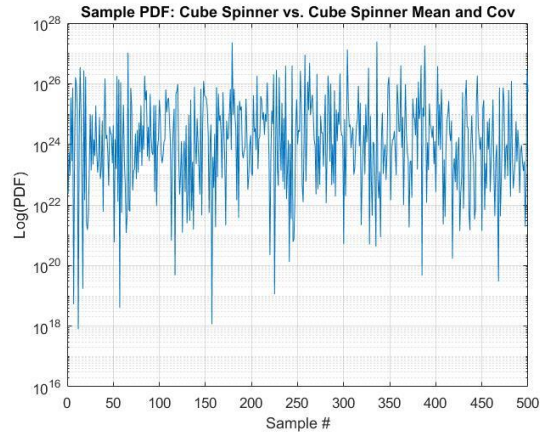

Figure 5. Log-likelihood of the spinning cube samples compared back to independent statistics sampled from the same object

Similarly, the large likelihood values presented in Figure 6 show that the independent spinning panel samples (Case 2) taken from the second half of each realization are highly likely to belong to that object. When the spinning panel samples were compared to the spinning cube mean and covariances, the resulting likelihood was very near zero, indicating the information was different. Likewise for when the likelihood was computed for spinning cube samples with respect to the spinning panel mean and covariance. These results, of course, make sense but are used as a validation and also to demonstrate how computations of likelihoods between light curves can identify similarities or differences in information content.
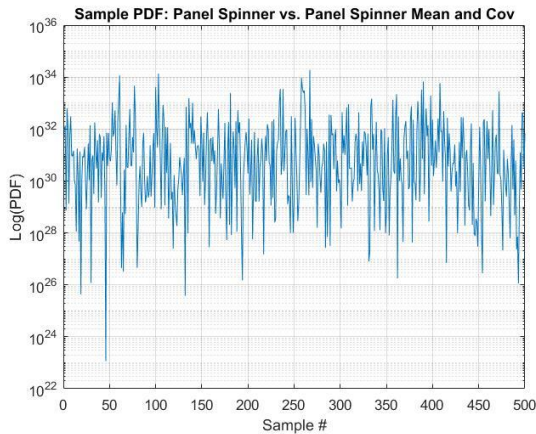
Figure 6. Log- likelihood of the spinning panel samples compared back to independent statistics sampled from the same object

The same analysis previously described was conducted for the tumbling box-wing satellite model (Case 3) and inertially fixed box-wing satellite model (Case 4). The varying geometry for each realization was an additional element that, essentially, resulted in potentially different information being represented on the same object due to differing lighting conditions, illuminated surfaces, and the dynamics observed from the given geometry.

As with the Case 1 and 2 comparisons, when each of the independent samples from the tumbling box-wing satellite model (Case 3) and inertial box-wing satellite model (Case 4) was used to evaluate a likelihood based on their respective means and covariance, the non-zero values indicate a high likelihood of the samples containing the same information. These results are shown in Figure 7 and 8, respectively.
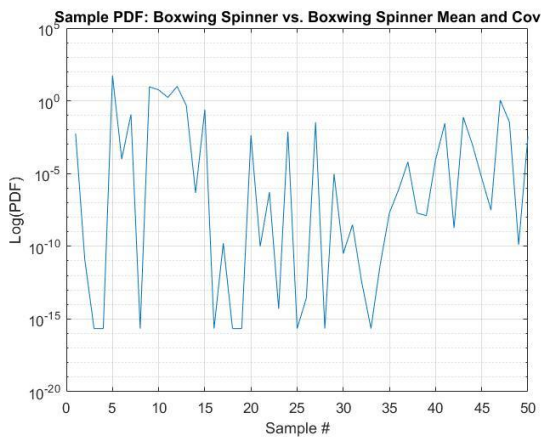

Figure 7. Log-likelihood of the spinning box-wing satellite model samples compared back to independent statistics sampled from the same object
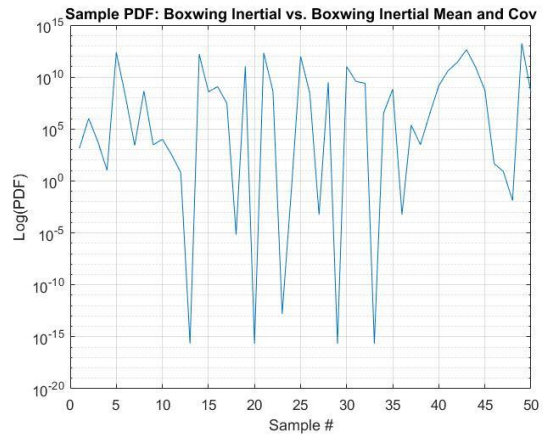

Figure 8. Log-likelihood of the inertial box-wing satellite model samples compared back to independent statistics sampled from the same object

In contrast, when the independent samples from the tumbling box-wing satellite model are used to compute the likelihood based on inertial box-wing satellite model mean and covariance, the values are again very close to zero as seen in Figure 9. The computation of the likelihood from inertial box-wing satellite model sample realizations with respect to the tumbling box-wing satellite model mean and covariance shows that, though not zero, the values are smaller than the comparisons in Figures 7 and 8. These results, shown in Figure 10, perhaps indicate some potential commonalities in the information in the samples versus independent samples used for the computation of the mean and covariance. Indeed, that same box-wing satellite model was used for both the tumbling and inertial cases, the only difference being in the dynamic characteristics.
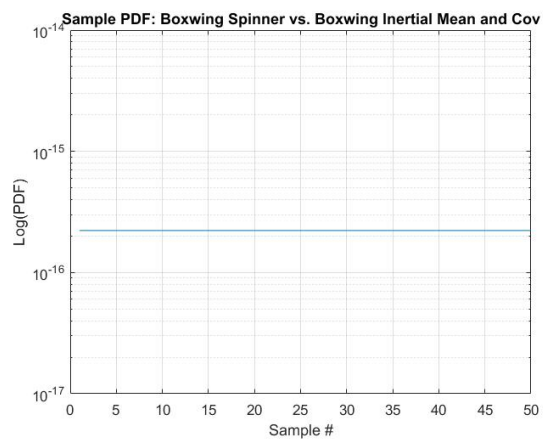

Figure 9. Log-likelihood of the spinning box-wing satellite model samples compared back to independent statistics sampled from the inertial box-wing satellite model
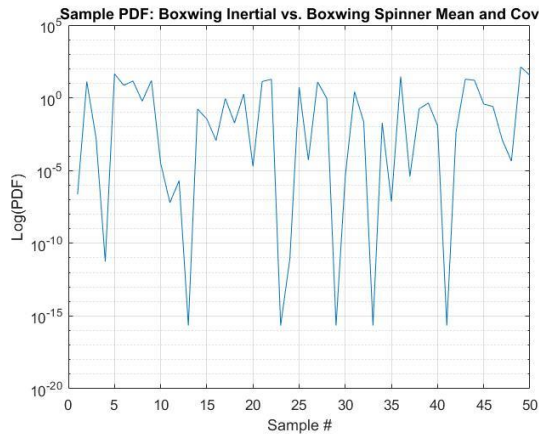
Figure 10. Log-likelihood of the inertial box-wing satellite model samples compared back to independent statistics sampled from the tumbling box-wing satellite model

Lastly, the K-L divergence (Eq. (1)) was computed for each of the combinations of cases for which the Gaussian likelihood was evaluated. Those results are given in Table 2 below. The low numerical values indicate little or no divergence, or that information between the sample cases and distribution statistics are consistent. In contrast, the large numerical values are an indication that the information is "divergent," in other words, the sample information is not consistent with the information captured in the distribution statistics.

Table 2. Kullback-Leibler divergence for the different combinations of case comparisons

Data Comarison Cases	KL-Divergence	Comments
Spinning Cube vs. Spinning Cube	0.02	Very comparable
Spinning Plate vs. Spinning Plate	-0.01	Very comparable
Spinning Cube vs. Spinning Plate	3389524.15	Not comparable (divergent)
Spinning Panel vs. Spinning Cube	1286300.08	Not comparable (divergent)
Boxwing Spinner vs. Boxwing Spinner	-0.01	Very comparable
Boxwing Inertial vs. Boxwing Inertial	0.01	Very comparable
Boxwing Spinner vs. Boxwing Inertial	110.98	Mildly comparable
Boxwing Inertial vs. Boxwing Spinner	2417.16	Moderately comparable

5 CONCLUSIONS

The analysis results for the four cases presented clearly demonstrate the viability of applying probabilistic functional data analysis to representing the information content of light curves and how it might allow a quantifiable metric for associating and correlating light curve signatures. Using Functional Data Analysis (FDA), we were able to derive a metric for the clustering of light curve data with a high degree of success. Independent realizations of light curve data for each cluster (i.e., class of object shapes) were shown to have a high likelihood of being associated with a distribution derived from an independent set of light curve data from the same object. Similarly, these preliminary results also showed a low likelihood of a set of realizations from one object being associated with a distribution derived from

a different object. The "population" of light curves for a given object retains a unique information content that allows these approaches to be successfully applied. Kullback-Leibler divergence was used to quantify the degree of similarity between different classes of light curves. The information theoretic metrics provide a measure of "how well" one can quantify association and, hence, a measure that can be used to prioritise tasking of subsequent observations to reduce ambiguity and increase the probability that a set of signatures belong to the same object.

Several assumptions were made in this analysis, and relaxing these assumptions serve as the motivation for future work. Of particular concern is the suitability of the polynomial expansion model as a representation of a light curve along with its assumed polynomial order. Though this is the parameterization chosen for modelling the light curve "attributes," no evidence currently exists that this is the best and most appropriate model. Also, the simulation setup was constructed in such a way that we insured adequate sampling of the signal. What if sampling is not adequate? Lastly, only a single sensor location was used in the simulation of the light curves. Prior analysis indicates additional sensors observing common objects should enhance the information content and, hence, require less data. Real light curve data will be analysed on known and well tracked objects to refine and validate the performance results.

6 ACKNOWLEDGEMENTS

7 REFERENCES

1. R. Linaris, M. Jah, J. Crassidis and C. Nebelecky, "Space Object Shape Characterization and Tracking Using Light Curve and Angles Data," *Journal of Guidance Control and Dynamics*, Vol. 37, No. 1, pp. 13-25, January 2014.

2. E. Gine, and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics, 2015.

3. J. L. Wang, J. M. Chiou, and H. G. Muller. Review of functional data analysis. Annual Review of Statistics and Its Application, Vol. 3, pp 257-295, 2016

4. X. Leng and H. G. Müller, "Classification using functional data analysis for temporal gene expression data," *Bioinformatics*, 22 (1): 68-76, 2006.

5. J. Faraway, A. Mahabal, J. Sun, X.-F. Wang, Y. G. Wang and L. Zhang, "Modeling lightcurves for

improved classification of astronomical objects," *Statistical Analysis and Data Mining*, Vol. 9, pp. 1–11, 2016.

6. P. Galeano, E. Joseph, and R. E. Lillo, "The Mahalanobis Distance for Functional Data with Applications to Classification," *Technometrics*, Vol. 57, No. 2, pp. 281-291, 2015.

7. I. Hussein, C. W. T. Roscoe, P. W. Schumacher, Jr., and M. P. Wilkins, "UCT Correlation using the Bhattacharyya Divergence", *Proceedings of the 26th AAS/AIAA Space Flight Mechanics Meeting*, Napa, CA, February 14–18 2016.

8. I. Hussein, M. P. Wilkins, C. W. T. Roscoe, and P. W. Schumacher, Jr., "On Mutual Information for Observation-to-Observation Association," 25th AAS/AIAA Space Flight Mechanics Meeting, Williamsburg, VA, January 11–15 2015.

9. Ian H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 4th Edition, 2016.

10. D. Arthur and S. Vassilvitskii. (2007), "k-means++: the advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia*, PA, USA. pp. 1027–1035.