Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **488,** 1377–1386 (2019)
Advance Access publication 2019 July 1

doi:10.1093/mnras/stz1795

Machine-learning identification of asteroid groups

V. Carruba [1]★ S. Aljbaae[2] and A. Lucchini[1]

[1] *São Paulo State University (UNESP), School of Natural Sciences and Engineering, Guaratinguetá, SP, 12516-410, Brazil*
[2] *National Space Research Institute (INPE), Division of Space Mechanics and Control, C.P. 515, 12227-310, São José dos Campos, SP, Brazil*

Accepted 2019 June 27. Received 2019 June 26; in original form 2019 May 7

ABSTRACT

Asteroid families are groups of asteroids that share a common origin. They can be the outcome of a collision or be the result of the rotational failure of a parent body or its satellites. Collisional asteroid families have been identified for several decades using hierarchical clustering methods (HCMs) in proper elements domains. In this method, the distance of an asteroid from a reference body is computed, and, if it is less than a critical value, the asteroid is added to the family list. The process is then repeated with the new object as a reference, until no new family members are found. Recently, new machine-learning clustering algorithms have been introduced for the purpose of cluster classification. Here, we apply supervised-learning hierarchical clustering algorithms for the purpose of asteroid families identification. The accuracy, precision, and recall values of results obtained with the new method, when compared with classical HCM, show that this approach is able to found family members with an accuracy above 89.5 per cent, and that all asteroid previously identified as family members by traditional methods are consistently retrieved. Values of the areas under the curve coefficients below Receiver Operating Characteristic curves are also optimal, with values consistently above 85 per cent. Overall, we identify 6 new families and 13 new clumps in regions where the method can be applied that appear to be consistent and homogeneous in terms of physical and taxonomic properties. Machine-learning clustering algorithms can, therefore, be very efficient and fast tools for the problem of asteroid family identification.

Key words: methods: data analysis – celestial mechanics – minor planets, asteroids: general.

1 INTRODUCTION

Asteroid families are groups of asteroids that share a common origin. They can either be the product of a collision, or they may results from the rotational fission of a parent body or their satellites (Pravec et al. 2010, 2018). Collisional asteroid families are usually identified in domains of proper elements (Hirayama 1923) that are constants of motion over time-scales of Myr (Knežević & Milani 2003). Among the methods used for identifying asteroid families, the hierarchical clustering method (HCM) is one of the most commonly used. In this method, the distance between two asteroids in domains of proper elements or frequencies is computed using a distance metric. If the second object distance from the first is less than a characteristic threshold called cut-off, the object is assigned to the first object asteroid family. The process is then repeated for the second asteroid, until no new members are found. Interested readers can found more details in Bendjoya & Zappalá (2002), or in Carruba (2010).

Recently, machine-learning clustering algorithms, available in the PYTHON programming language, have been used with great success for problems like clusters identification (Pedregosa et al. 2011). Methods such as K-means, mean-shift, and hierarchical clustering algorithms are now very commonly used among data scientists, and applied to various different fields, such as biology, palaeontology, and, as in this work, astronomy. These machine-learning algorithms have proven to be efficient, fast, and reliable in problems of supervised learning. Here, we will attempt to use machine-learning HCMs for the purpose of asteroid families identification, in domain of asteroid proper elements (a, e, $\sin i$), we will verify the results with respect to previous works, and, where possible, we will assert the validity of the new asteroid groups obtained with such methods by studying their physical and taxonomic properties.[1]

Other methods for family identification, based and not based on HCM, have been recently proposed for the same purpose. Methods

[1] Some orbital regions, like the Cybele region, are mostly populated by dark, C-complex asteroids, so that differentiating asteroid families based on their physical properties in this region is a more difficult task.

★ E-mail: valerio.carruba@unesp.br

based on cladistic classifications have been recently proposed for the identification of asteroid families among Jupiter Trojans Holt et al. (2017). Other approaches, based on the *V*-shape identification in domains of (a, 1/D), with a the semimajor axis and D the asteroid diameter, have been recently used to identify very old families not easily detectable using standard HCM (Bolin et al. 2017, 2018; Delbó et al. 2017; Delbó, Avdellidou & Morbidelli 2019).

All these methods have merits and great potential. Here, however, we will focus our attention on the use of machine-learning techniques in domains of proper (a, e, $\sin i$). Standard HCM has been used and tested with great success for several years. Since the goal of this paper is to focus on the possibility of using machine-learning techniques for the purpose of asteroid families identification, we believe that a good starting point should be the use of well-known methods first. The applications of machine-learning algorithms to new methods remain, of course, an interesting prospect for future research.

Concerning standard HCM, one problem with its application is chaining. In high asteroid number-density regions, near families may overlap and not recognizable as separate entities by the method. Milani et al. (2014) introduced new approaches to deal with this problem, by working first with a more limited sample and by extending the family identification to higher numbered objects at a later stage. Here, we will focus our attention to the low-number density regions where standard HCMs can still be used, such as the Hungaria region, inner, central and outer main belt at high inclinations ($\sin i > 0.3$), and the Cybele regions, whose boundaries will be more precisely defined later on in this paper. The problem of identifying asteroid families in higher number-density regions will be left as a challenge for future works.

The structure of the paper is as follows. In Section 2, we will revise the standard HCM and its implementation in machine learning. Methods to determine the accuracy of machine-learning routines for identifying families in old and new data bases, such as accuracy, precision, recall, and areas under ROC curves (*AUC*) will also be discussed in this section. In Section 3, we will apply the new method to data bases of asteroid proper elements in the five low-number density regions, and check for the validity of old and newly identified asteroid families, with an emphasis on newly identified dynamical groups. Finally, in Section 4, we will present our conclusions.

2 ACCURACY OF MACHINE-LEARNING HIERARCHICAL CLUSTERING ALGORITHMS

HCMs have been used to identify asteroid families for several decades and are a well-established method for the identification of dynamical asteroid groups. Several articles in the literature explained in detail the use of this approach; interested readers could find more details in Zappalá et al. (1990), Bendjoya & Zappalá (2002), and Carruba et al. (2015), among others. The basic approach follows these guidelines. First, the distance between pairs of objects involving a putative parent body and a candidate family is computed in a domain of proper elements according to a pre-defined metric. The most commonly used metric distance d is defined as (Bendjoya & Zappalá 2002)

$$d = na\sqrt{\frac{5}{4}\times\left(\frac{\Delta a}{\overline{a}}\right)^2 + 2\times(\Delta e)^2 + 2\times(\Delta \sin i)^2}, \quad (1)$$

where (a, e, i) are the proper semimajor axis, eccentricity, and inclination; the symbol Δ is associated with the difference between pairs of proper elements, and $\overline{a}$ is the mean value of the proper semimajor of a given pair of asteroids. If the distance between two objects is less than a value defined as the local cut-off, the family candidate is assigned to the parent body asteroid group. The process is then repeated with the new family member now considered as a parent body until no new members are found. Beaugé & Roig (2001) define a nominal distance cut-off d_0 as the average minimum distance between all neighbouring asteroids in the same region of the asteroid. For the 1500 sample of objects used by Carruba et al. (2015) for their families identification, the value of d_0 was of 138.45 ms^{-1}. Carruba et al. (2015) identified three reliable dynamical groups: the (87) Sylvia family, with 363 members, the (260) Huberta family, with 58 members, and the (909) Ulla family, with 30 members.

In this work, we implemented SCIKIT-learn Pedregosa et al. (2011) hierarchical clustering algorithms for the same problem of asteroid family identification. Dendrogram clusters of asteroid distances computed using equation (1) can be automatically obtained almost instantaneously using PYTHON algorithms such as *linkage*, and asteroid families obtained for different cut-off values can be easily identified. For illustrative purposes, Fig. 1 displays a dendrogram of asteroid distances for the first 50 lowest numbered objects in the 2015 proper element data base obtained with this approach. The x-axis displays the sample index of each objects, defined so that the first asteroid has an index equal to 0 and the last object has an index equal to 49. As a first step of our analysis, we evaluate how consistent are the results obtained with the new approach with respect to those obtained in 2015.

A standard practice when using unsupervised learning algorithms is to define a confusion matrix Stehman (1997). Given the actual sample and the one predicted by the hierarchical clustering approach, four quantities can be defined: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In our context, true positives are asteroids identified as family member by both methods; false positives are asteroids identified as family members by the machine-learning algorithm alone, or false alarms, or type-I errors. True negatives are asteroids not identified as family members by both approaches, and false negatives are asteroids not identified as family members only by the machine-learning algorithm, or objects with miss, or type-II errors. Using these four quantities, we can define the Accuracy, Precision, and Recall parameters for the outcome of the studied algorithm. The accuracy is a measure of the overall quality of the prediction and is defined as

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}. \quad (2)$$

The precision measures the ability of the model to avoid prediction of false data and is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}. \quad (3)$$

Finally, the recall is an indication of how many real family members are identified correctly by the machine-learning algorithm, and is given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}. \quad (4)$$

Table 1 displays values of all these parameters for the three families identified in 2015, and three other not identified in that work, but that can be retrieved using standard HCM and the 2015 proper element data base. Overall, the accuracy of the hierarchical clustering algorithm is always above 89.5 per cent, and very few false negatives

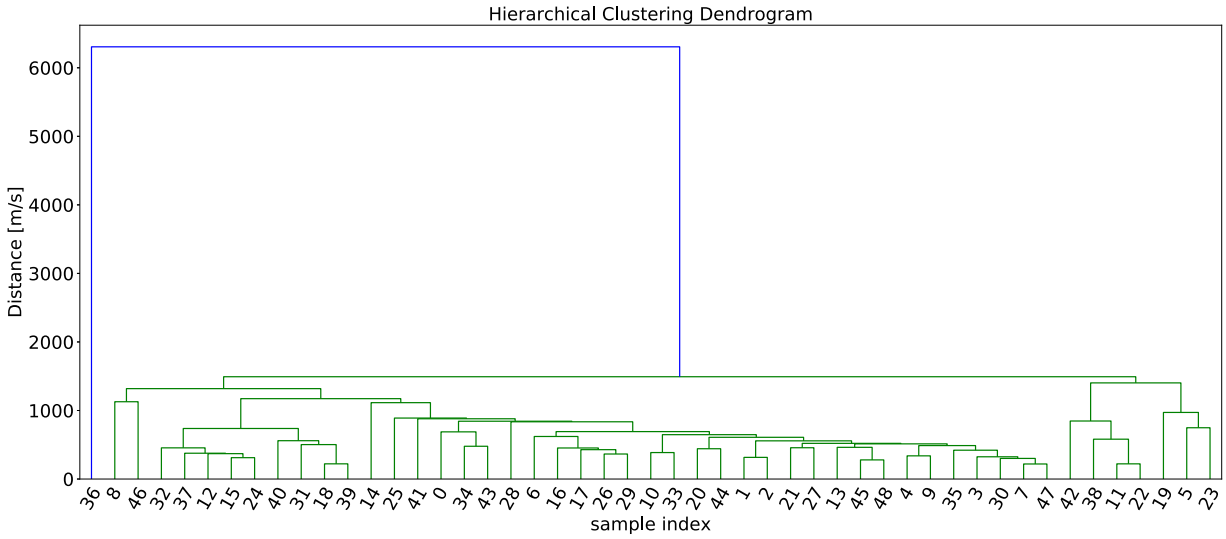

Figure 1. A dendrogram of orbital distances for 50 objects in the Cybele orbital region. The vertical axis displays the distance cut-off, while on the horizontal axis there are the sample identifications of the 50 asteroids in the Cybele region. Vertical lines identify a single cluster; horizontal lines display the merging of near clusters.

Table 1. Values of true positive (TP), false negative (FN), false positive (FP), true negative (TN), Accuracy, Precision, and Recall parameters for three studied families in the Cybele orbital region.

Family name	TP	FN	FP	TN	Accuracy (%)	Precision (%)	Recall (%)
(87) Sylvia	363	0	23	1113	98.5	94.0	100.0
(260) Huberta	58	0	52	1389	89.5	52.7	100.0
(909) Ulla	30	0	0	1469	100.0	100.0	100.0
(1028) Lydina	67	2	13	1417	99.0	83.8	97.1
(6924) Fukui	15	0	17	1467	98.9	46.8	100.0
(26607) (2000 FA33)	9	0	16	1474	98.9	36.0	100.0

are ever found by this method, with a recall coefficient mostly equal to 100 per cent, with the only exception of the Lydina family for which two false negative asteroids were found. The limitation of this method concerns, however, its precision. Machine-learning clustering algorithms tend to find more family members that the standard HCM does, which limits its Precision. This excess of asteroids assigned to a given family, especially evident for families located in more densely populated areas, such as Huberta, or (26607) (2000 FA33), needs to be considered when dealing with families obtained with machine-learning clustering algorithms. Comparison with results of standard HCM, as performed in this section, may be advisable. We will come back to this subject in later article sections.

To further study the accuracy of this method, we also use the ROC curve approach. ROC curves are a well-established tool to quantitatively define the efficiency and accuracy of machine-learning clustering algorithms. A detailed description of how ROC curves are computed and of the rationality behind them can be found in Fawcett (2006). In this work, we first computed asteroid distances with respect to the alleged parent body using the distance metric given by equation (1). The sample of family members obtained by the machine-learning HCM is then randomly split into a train sample (60 per cent of the total) and a test sample (40 per cent).[2] A logistic regression algorithm is then applied to the train sample, and the probability of the test sample to belong to the family is then assessed based on the results of the train sample.

Basically, logistic regression in its simplest form uses a logistic function to model a binary variable, with two possible values, such as pass/fail, win/lose, or alive/dead (Cox 1958). In our case, either the asteroid belongs to a given family (value =1), or it does not (value = 0), as a function of its distance in proper element domain from the possible parent body. For each object, the probability of it to belong to a given family is computed using a logistic function, and it varies between 0 and 1. For different values of this probability, the false positive rate and the true positive rate are then computed. ROC curves are shown in a the domain of true positive rates versus false positive rates and the area under the curve (AUC) coefficient, which has a value between 0.5 (random predictions) and 1.0 (perfect score), provide a quantitative measure of the goodness of the model.

Fig. 2 shows ROC curves for the six families reported in Table 1. Values of the AUC coefficients are reported in Table 2. Again, the isolated Ulla family obtains a perfect 1.00 score, with no false positive or true negative asteroids. Results for all but one of the

[2]To check how much the results depend on how the sample is divided, we used the cross-validation approach. In this method, the sample is first divided into n samples of equal size. One of them is used as a train sample, and the others as tests. The efficiency of the method is evaluated for the first sample, and the procedure is then repeated for the other $n - 1$ samples. For the case of family identification, the cross-validation score never dropped below 94 per cent, suggesting that the problem is rather robust with respect to the way the sample is divided.

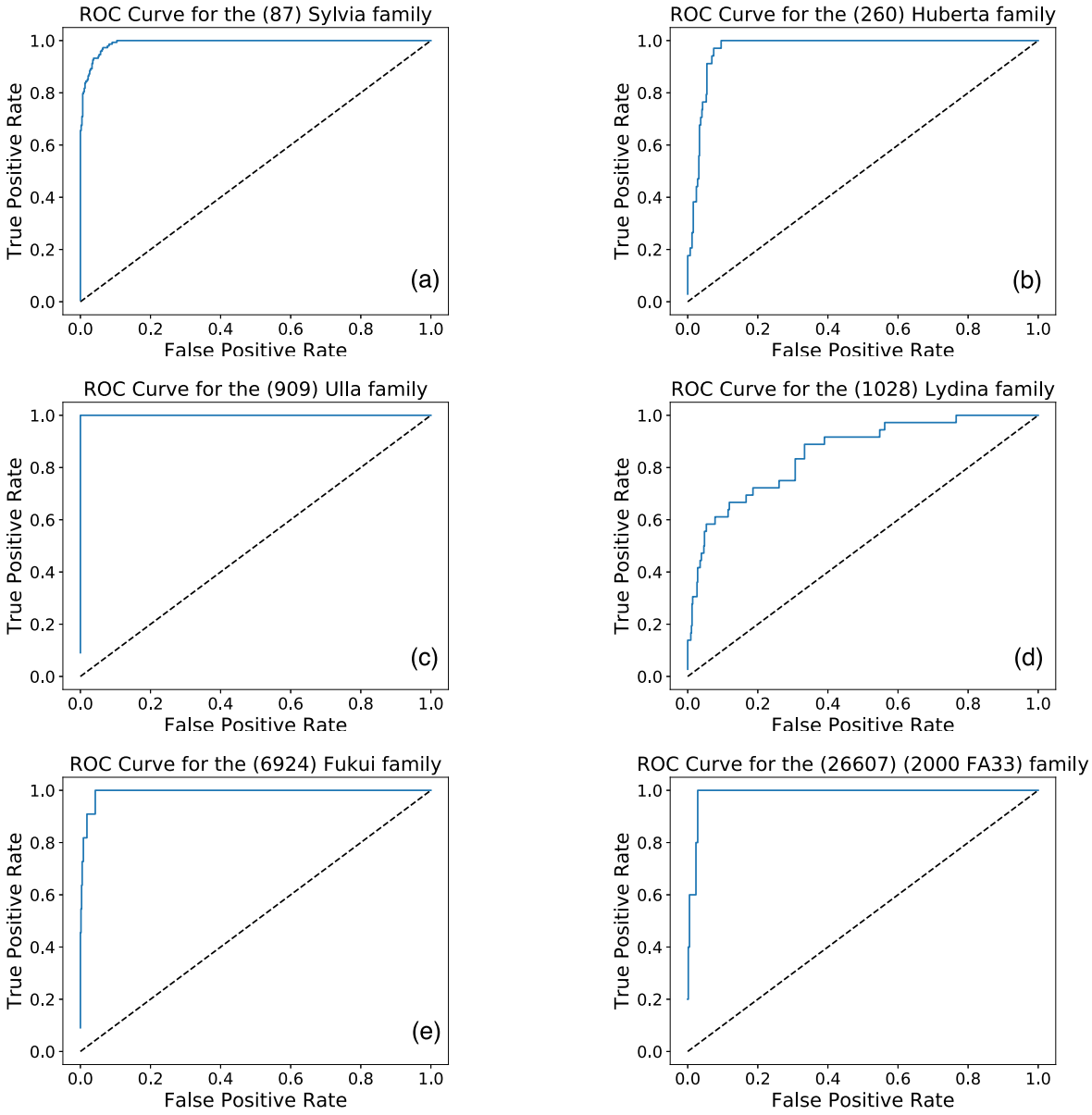

Figure 2. Area under Receiver Operating Characteristic (ROC) curves for the cases of the identification of the (87) Sylvia, (260) Huberta, (909) Ulla, (1028) Lydina, (6924) Fukui, and (26607) (2000 FA33) asteroid families obtained from logistic regression algorithms.

Table 2. Values of the AUC coefficients for the ROC curves shown in Fig. 2.

Family name	AUC value
(87) Sylvia	0.991
(260) Huberta	0.970
(909) Ulla	1.000
(1028) Lydina	0.855
(6924) Fukui	0.993
(26607) (2000 FA33)	0.988

studied families are all above 97.0 per cent. Only in relative high number density orbital regions, such as that of (1028) Lydina, AUC values drop to 0.855, which is, however, still a very good result.

Based on this analysis, we believe that machine-learning clustering algorithms appear to be an efficient tool for identifying asteroid families, with results that are characterized by a highly effi−ient performance, with AUC values all very close to 1. The Precision of the families obtained with this method may, however, vary. A comparison of results obtained with machine-learning clustering algorithms and standard HCM may, therefore, always be advisable. In the next section, we will start applying these methods to various regions of the asteroid main belt, where traditional HCM methods can be still applied.

3 DYNAMICAL FAMILIES IN LOW-DENSITY REGIONS OF THE MAIN BELT

In this section, we will implement the machine-learning clustering methods defined in the previous section to asteroid families in suitable regions of the main belt. Following the approach of Carruba (2010), families are considered robust when they are recognizable

not only at the nominal cut-off, but for a range of at least 10 ms^{-1} above and below this value. The minimal number for a group to be recognizable as a family or a clump depends on the number density of asteroids in a given region [see for instance Carruba (2010) for a discussion of this issue]. Here, for the sake of uniformity among the different orbital region studied, we used a simplified approach where we considered a group to be a family if there are at least 25 members, and a clump if there are at least 10 members. Groups with less members will be investigated only if they show signs of being of particular interests, such as harbouring fission pairs or clusters.

Once the families have been identified, we checked the physical properties of their members to see if they are consistent with a common origin. In particular, we used the approach of DeMeo & Carry (2013), based on the photo-metric data from the Sloan Digital Sky Survey-Moving Object Catalog data (SDSS-MOC4; Ivezić et al. 2001), to obtain taxonomic information for objects in that data base. We then also look for asteroids that have geometric albedo values in the *Wide-field Infrared Survey Explorer* (*WISE*) and NEOWISE, AKARI, or IRAS data bases (Ishihara et al. 2010; Ryan & Woodward 2010; Masiero et al. 2012), and that have the good signal-to-noise ratio described in Spoto, Milani & Knežević (2015). C-complex asteroids tend to have albedo lower than 0.12, while S-complex objects have higher albedos. See also Spoto et al. (2015) for a discussion of most common albedo values among asteroids classes. Finally, following the approach of Milani et al. (2019), we verified if the fraction of mass present in the largest remnant is higher or lower than 75 per cent. In the first case, families are classified as the outcome of a craterization event, while in the second case they are more likely to be the product of a fragmentation scenario.

Recently, it has been shown that asteroid families may not only form as a result of a collision, but, in some cases, they may be the outcome of a rotationally induced fission event. With respect to families formed in collisions, fission clusters tend to be more compact in proper element domains, and the mass ratio between the parent body and the other members does not exceed 0.3 (Pravec et al. 2010, 2018). Since a recent study by Carruba et al. (2019) suggests that the formation of young asteroid families may trigger a subsequent chain of formation of fission clusters; in this work, we also checked for the presence of possible fission clusters. Following the approach of Carruba et al. (2019), we checked for fission cluster fission pair candidates whose distance in proper element domain computed using equation (1) is less than 5 ms^{-1}, and whose mass ratio is less than 0.3. The validity of these candidate clusters is then checked using methods that integrate asteroids orbits in the past, such as backward integration method (BIM) and close encounter method (CEM).

The BIM obtains estimates of the ages of young asteroid families by integrating the orbits of possible members in the past and by checking for the convergence of the longitudes of nodes and pericenters with respect to those of the alleged parent body. At the moment of family formation, those angle differences should all converge to within a few degrees. The CEM checks for the time of close encounters between pairs of objects in the past. Several clones of asteroid pairs accounting for the orbital uncertainties and for the Yarkovsky effect are integrated into the past, and the time of approach to within a given distance is registered. The median value of the distribution of close encounters times provides an estimate of the time of formation of the pair, while the 5th and 95th percentile of the distribution allow estimating the uncertainty on the age. Interested readers can found more details on these two methods on Carruba et al. (2019) and reference therein. Here, we first apply BIM to all the new families, clumps and clusters identified in this work. To confirm the possible young age of the group, CEM is then used.

Table 3. Boundaries of the orbital regions of the main belt studied in this work. The table reports the zone identification name (where H. I. stands for highly inclined), the value of a_{min}, the name of the Jupiter mean-motion resonance associated with that value, the value of a_{max} and its associated mean-motion Jupiter resonance name, and the minimum value of sin (i_{min}).

Region name	a_{min} (au)	Res. name	a_{max} (au)	Res. name	sin (i_{min})
Hungaria region	1.780	–	2.253	7J:2A	0.000
H. I. inner main belt	2.257	7J:2A	2.465	3J:1A	0.300
H. I. central main belt	2.520	3J:1A	2.818	5J:2A	0.300
H. I. outer main belt	2.832	5J:2A	3.240	2J:1A	0.300
Cybele region	3.290	2J:1A	3.800	5J:3A	0.000

Finally, for the families with a number of members large enough such that dating methods based on the slopes of *V*-shapes in (a, 1/D) domains, where D is the asteroid diameter, we will also attempt to obtain estimates of the family age based on Spoto et al. (2015) approach.

As discussed in the introduction, five orbital regions are suitable for the applications of the new methods of asteroid family determination: the regions of the Hungaria asteroids, the inner, central and outer main belt at high inclinations, and the region of the Cybele asteroids. Table 3 displays the name and the limits in proper a and sin (i) of the regions studied in this work. We will start our analysis by analysing the Hungaria orbital region.

3.1 Dynamical families in the Hungaria orbital region

The region of the Hungaria asteroids is delimited by several mean-motion resonances with Mars at lower a, and by the 7J:-2A mean-motion resonance with Jupiter at higher a. Only the dynamical family of 434 Hungaria itself is currently well established (Milani et al. 2010), while S-complex possible groups at higher inclinations that were proposed in the past have not yet been confirmed [see Carruba et al. (2013) and references within for a discussion on the possible existence of such groups]. The current value of d_0 for asteroids in this region is 68.59 ms^{-1}. Therefore, we also obtained families for 58.59 and 78.59 ms^{-1} cut-offs. Our results are shown in Table A1 in the Appendix.

No new families or clumps were identified in the Hungaria region, where we only confirm the existence of the namesake family. Three asteroid pair candidates were confirmed by BIM in the Hungaria region, those of 84203 (2002 RD133) and 285637 (2000 SS4), 88259 (2001 HJ7) and 337181 (1999 VA117), and that of 145046 (2005 GD2) and 453039 (2007 RB325). All pairs are members of the Hungaria dynamical family. CEM found ages of $0.069^{+0.776}_{-0.049}$, $0.248^{+1.319}_{-0.176}$, and $1.107^{+1.530}_{-0.967}$ Myr, respectively.

In the next sub-section, we will analyse the dynamical groups in the high-inclination inner main belt.

3.2 Dynamical families in the highly inclined inner main belt orbital region

The inner main belt is a dynamically stable island delimited by the 7J:-2A mean-motion resonance with Jupiter at low a, by the 3J:-1A resonance at high a, and, at low i, by the ν_6 secular resonance (Knežević & Milani 2003; Carruba 2009). Most of the asteroids in the island are members of the (25) Phocaea family, with just a few

other possible small groups that have been proposed in the past. Dynamical groups in domain of proper elements and frequencies were obtained for this region by Carruba (2009), which, apart from the Phocaea family itself, identified four other small dynamical groups. Novaković et al. (2017) studied the dark population of objects in the Phocaea island, and identified in this restricted domain the (326) Tamara family, which is not easily retrievable in larger domains, and will not further discuss in this work. The value of the cut-off d_0 for the H. I. inner main belt current data base is $d_0 = 110.58$ ms^{-1}, so we also obtained groups for $d_0 = 100.58$ and $d_0 = 120.58$ ms^{-1}. Our results are shown in Table A2 in the Appendix.

The families of (6246) Komurotoru, (19536) (1994 JM4), and (26142) (1992 PL1) found in Carruba (2009) join the larger Phocaea family and are not confirmed in this work. The Carruba (2009) clump (17628) (1996 FB5) has only six members and will not be further discussed. Among the new groups currently detected, the only clump that satisfied our selection criteria was that of (7758) Poulanderson, whose age is not obtainable with BIM. It is a 15 members clump, with a two-sided shape, and taxonomy and albedo most likely to belong to the C-complex. Consistently with the analysis of Section 2, Accuracy, Recall, and AUC coefficients are all very good, with values all above 95 per cent. The Precision of this group is, however, 73.3 per cent. Four false positive asteroids and one false negative were found by the machine-learning HCM for the Poulanderson clump. The properties of this and other clumps and families identified in this work are listed in Table 4. In the last column, we report the Precision of the newly identified groups.

There were two asteroid pairs candidates in the inner main belt that were confirmed by BIM: the pair 51866 (2001 PH3) and 326894 (2003 WV25) that are not members of any dynamical group, and the pairs (1999 RX33) and (2013 GZ99) that are inside the Phocaea family. CEM found an age of $0.110^{+0.305}_{-0.072}$ and of $0.055^{+0.109}_{-0.032}$ Myr for these two pairs, respectively. The next sub-section will deal with groups identified in the high-inclination central main belt.

3.3 Dynamical families in the highly inclined central main belt orbital region

The central main belt is separated in proper a from the inner and outer belts by the 3J:-1A and 5J:-2A mean-motion resonances with Jupiter. The boundary at lower i is given by the ν_6 secular resonance. The ν_5 and ν_{16} secular resonances cross the regions, dividing it into several islands, the so-called stable archipelago in the central main belt (Carruba 2010).

The current value of the cut-off d_0 is 106.68 ms^{-1}, and families were also obtained for cut-offs of 96.68 and 116.68 ms^{-1}. Our results are displayed in Table A3 in the Appendix. We confirm and improve the membership of families found in previous works, such as Carruba (2010), and we identified two new families and six new clumps not previously known: the families of (694) Ekard, and (75779) (2000 AC201), and the clumps of (46542) (1987 AD), (2134) Dennispalm, (4404) Enirac, (59511) (1999 JP14), (21785) Mechain, and (5438) Lorre. Results are summarized in Table 4. As for the case of the new clump in the inner main belt, Accuracy, Recall, and AUC coefficients were all above 90 per cent and, for brevity, were not reported.

Among the already known families, of particular interest there were those of (2) Pallas, (1222) Tina, and (5438) Lorre. Pallas is the third most massive body in the asteroid belt, and its orbit is one of the most inclined. Understanding the origin of Pallas and its family has been of the main challenges of asteroid dynamics since their discovery [see also Carruba (2010) for a more in-depth discussion of this family]. The family of (1222) Tina is special, since it is located in a stable island of the ν_6 secular resonance, on anti-aligned librating orbits, that prevent them from experiencing planetary close encounters (Carruba & Morbidelli 2011). (5438) Lorre, first identified by Novaković, Hsieh & Cellino (2012) also using multi-opposition and single opposition asteroids, is one of the youngest family in the main belt for which an age estimate with BIM can be obtained. It contains several internal sub-clusters, most likely originating from rotational fission, as recently found in Carruba et al. (2019). Other families in the region, according to the definition given in this work, were detected by previous works, such as for the case of (10000) Myriostos, identified in Novaković, Cellino & Knežević (2011).

Concerning the new families listed in Table 4, our results confirm the trend observed in Section 2: families retrieved by the machine-learning algorithm tend to be larger than the analogous families identified by the traditional HCM. Two clumps, those of (46542) (1987 AD) and (2134) Dennispalm, have values of the Precision coefficient below 60 per cent and should be considered as dubious. The largest number of false negative case, 5, was observed for the case of the (75779) (2000 AC201) family. In general, machine-learning HCM did a good job in retrieving family members that were also retrieved by the traditional approach. Finally, for all the groups, we considered false positive asteroids as candidate members to be confirmed by later studies.

BIM confirmed the already known case of angles convergence of (5438) Lorre. The only other possible case of angle convergence with BIM was, possibly, observed for the 64785 (2001 XW197) clump. The young age for this cluster, however, is not confirmed by CEM. The only fission pair candidate that we found in the central main belt was that of (5438) Lorre and 208099 (2000 AO201). Since fission clusters inside the Lorre family were already studied in detail Carruba et al. (2019), we will not further discuss this issue here. Interested readers can find more information in the cited paper.

Concerning the physical properties of the newly identified groups, most of the groups are compatible with a C-complex taxonomy, with just three groups possibly belonging to the S-complex, those of (2134) Dennispalm, (4404) Enirac, and (59511) (1999 JP14). Most of the groups may be the outcome of fragmentation events, with just two families, those of (694) Ekard and (5438) Lorre the possible outcome of cratering events. There were five two-sided families in (a, 1/D) domains, four incomplete left shapes, and two incomplete right shapes. The next sub-section will be dedicated to the groups identifiable in the outer main belt region.

3.4 Dynamical families in the highly inclined outer main belt orbital region

The highly inclined outer main belt is the region between the 5J:-2A and the 2J:-1A mean-motion resonances with Jupiter in proper a and inclinations above those of the ν_6, or, roughly speaking $\sin(i) > 0.3$. For the outer main belt, current d_0 is 77.81 ms^{-1}, and dynamical groups were also identified for $d = 67.81$ and $d = 87.81$ ms^{-1}. Our results are shown in Table A4 in the Appendix.

The method does not identify new families or clumps in this region, but confirms all the already known families of 31 Euphrosyne, 781 Kartvelia, 350 Ornamenta, 780 Armenia, 702 Alauda, and 704 Interamnia, extending their membership. Among this groups, of particular interest are the Euphrosyne family, characterized by its interaction with the ν_6 secular resonance and a possible sources of NEAs (Carruba, Aljbaae & Souami 2014; Masiero et al. 2015), and the Alauda family, which is possibly among the oldest asteroid

Table 4. The main belt region, cluster identification, number of members, the mean albedo values with its uncertainty, defined as one standard deviation, and the number of objects with albedo values, the number of albedo interlopers, if any, the taxonomic class and the number of objects with such information, the number of taxonomic interlopers, if any, the qualitative shape in the (a, $1/D$) plane (T for families with two sides, L for those with only the left side, and R for those with only the right side), and if the family is either the outcome of a fragmentation (F, $\frac{M_{PB}}{M_{Tot}} < 0.75$) or of a cratering (C, $\frac{M_{PB}}{M_{Tot}} > 0.75$) event, and the Precision coefficient given by equation (3), for the new asteroid families and clumps identified in this work.

Main belt region	Cluster id.	# of members	Mean Albedo	# of albedo int.	Tax. Class.	# of tax. int.	*V*-shape and fam. type	Precision (%)
Inner M. B.	7758 Poulanderson	15	0.06 ± 0.03-2	0	C-1	0	T-F	73.3
Central M. B.	694 Ekard	118	0.10 ± 0.05-28	9	C-3	1	L-C	61.0
Central M. B.	75779 (2000 AC201)	50	0.06 ± 0.01-15	0	C-2	0	T-F	62.0
Central M. B.	46542 (1987 AD)	23	0.06 ± 0.02-8	0	X-1	0	T-F	52.2
Central M. B.	2134 Dennispalm	19	0.30 ± 0.07-2	0	K-1	0	T-F	42.1%
Central M. B.	4404 Enirac	17	0.24 ± 0.23-1	0	−0	0	L-F	88.2
Central M. B.	59511 (1999 JP14)	17	0.27 ± 0.08-4	0	−0	0	R-F	58.8
Central M. B.	21785 Mechain	13	0.12 ± 0.06-6	2	X-2	1	T-F	76.9
Central M. B.	5438 Lorre	12	0.06 ± 0.01-2	0	−0	0	L-C	91.7
Cybele R.	1028 Lydina	93	0.05 ± 0.01-31	0	C-8	0	T-F	74.2
Cybele R.	6924 Fukui	35	0.05 ± 0.01-11	0	X-7	1	T-F	85.7
Cybele R.	19513 (1998 QN7)	35	0.08 ± 0.02-19	1	C-3	0	T-F	37.1
Cybele R.	26607 (2000 FA33)	31	0.06 ± 0.01-17	0	X-2	0	T-F	71.0
Cybele R.	3622 Ilinsky	21	0.09 ± 0.02-7	0	D-1	0	T-F	66.7
Cybele R.	1390 Abastumani	12	0.06 ± 0.04-2	0	D-1	0	L-C	100.0
Cybele R.	522 Helga	11	0.04 ± 0.04-1	0	−0	0	T-C	100.0
Cybele R.	3092 Herodotus	11	0.07 ± 0.02-8	0	D-1	0	R-F	100.0
Cybele R.	102167 (1999 RC223)	11	0.06 ± 0.06-1	0	−0	0	T-F	90.9
Cybele R.	46305 (2001 OW71)	10	0.08 ± 0.08-1	0	−0	0	R-F	100.0

families in the main belt (Carruba et al. 2016). There are two candidate pairs in the region, those of (62128) (2000 SO1) and 84085 (2002 QU24) in the Kartvelia family, and those of 243446 (2009 FX56) and 368051 (2012 HK36) in the Euphrosyne family, but neither is confirmed by BIM. The next sub-section will deal with the dynamical families in the Cybele orbital region.

3.5 Dynamical families in the Cybele orbital region

Following the approach used for the other regions of the asteroid main belt, we first apply the HCM to objects in the Cybele orbital region. The cut-off value d_0 for this region is 130.3 ms^{-1}, so we also obtained families at cut-offs of 120.3 and 140.3 ms^{-1}. Our results are shown in Table A5 in the Appendix. Apart from the previously identified Sylvia, Huberta, and Ulla groups (Carruba et al. 2015), we identified five new families and five new clumps in the region that attend our selection criteria: the families of (1028) Lydina, (6924) Fukui, (19513) (1998 QN7), (26607) (2000 FA33), and the clumps of (3622) Ilinsky, 1390 Abastumani, (522) Helga, (3092) Herodotus, (102167) (1999 RC223) and (46305) (2001 OW71). The clump of (522) Helga does not strictly satisfy our selection criteria, since it only has eight members at the lower cut-off. However, because of its dynamical importance (Helga could be the farthest family in the main belt, not considering the Hilda asteroids, that are in the Jupiter mean-motion resonance 3J:-2A, and the Thule asteroids, in the resonance 4J:-3A), we will keep it in our list none the less. As expected, the number of members of known families also significantly increase with respect to previous determinations.

Among the new families, the Lydina and Helga families were previously proposed by Vinogradova (2014) and Carruba et al. (2015) for the case of the Helga family, and appear to be confirmed by this work. We computed values of Accuracy, Recall, AUC, and precision coefficients for all the new families identified in this work. With respect to the results shown in Section 2, the values of these coefficients for the families of (1028) Lydina, (6924) Fukui, and (26607) (2000 FA33) are different because of the larger data base of proper elements used with respect to the work of Carruba et al. (2015). Overall, values of the Precision coefficient were above 66.6 per cent for all but one case, the family of (19513) (1998 QN7) whose existence needs to be confirmed with independent methods. The number of false negative cases in no cases exceeded 1, consistently with results from Section 2. Again, further studies are needed to confirm false positive asteroids as family members.

All the new groups and possible fission clusters were studied with BIM. We only found one possible fission pair that passed our selection criteria: the pair 16918 (1998 FF32)-411534 (2011 BM107) in the 3622 Ilinsky family has a possible age of $\simeq 3$ Myr, but this result is not confirmed by CEM. Overall, we did not find young clusters in the Cybele region.

We then studied the newly found groups and checked their physical and orbital properties with the methods described in Section 3. Our results are shown in Table 4. Most new clusters are dark, C-complex groups, with very limited numbers of albedo or taxonomic interlopers. Most of the groups are most likely to be the outcome of fragmentation's and show two-sides of their V-shape in (a, $1/D$) diagrams. There were, however, two cratering groups [those of (1390) Abastumani and (522) Helga] and three groups with incomplete V-shapes, those of 1390 Abastumani, 3092 Herodotus, and 46305 (2001 OW71). In the next section, we will try to obtain age estimates and determination accuracies for the largest families (80 members) or more that we determined so far.

4 AGE ESTIMATES OF NEWLY IDENTIFIED LARGE ASTEROID FAMILIES

In this section, we will focus our attention on the largest asteroid families identified in this work, the ones with a population large enough (80 members or more) that dating methods such as those of

Table 5. The AUC coefficient, the IN and OUT values of the slopes, and the estimated ages, with their errors, for the new large families (80 members or more) identified in this work.

Family Id.	AUC value	S_{IN}	S_{OUT}	Age_{IN} (Myr)	Age_{OUT} (Myr)
694 Ekard	0.914	-6.07 ± 3.00	–	$284.5^{+278.0}_{-94.1}$	–
1028 Lydina	0.855	-4.48 ± 9.84	5.39 ± 3.95	$509.3^{+1266.4}_{-350.0}$	$423.3^{+1161.1}_{-188.0}$

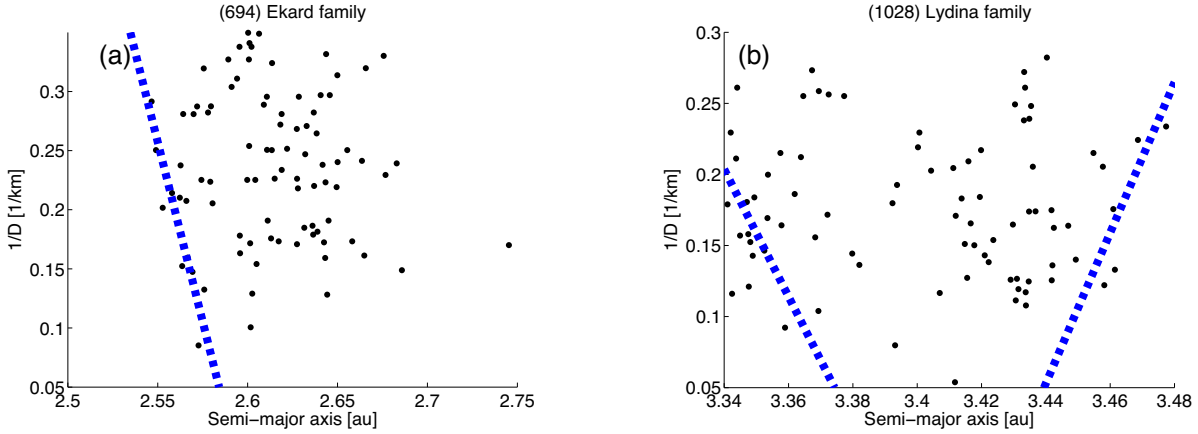

Figure 3. *V*-shape slopes in the (a,1/D) plane for the families listed in Table 5.

Spoto et al. (2015) can be used, i.e. the Ekard and Lydina families. This approach computes the slopes S of the inner and outer parts of the V-shaped distribution of asteroids in the (a, $1/D$) domain. The age of the family is given by the inverse of the slope divided by the da/dt drift rate caused by the Yarkovsky effect, which is tuned to the physical properties of a given family, according to the method described in section 4.1 of Spoto et al. (2015). Values of da/dt for the Ekard and Lydina families obtained with this approach are 1.59×10^{-12} and 1.20×10^{-12} au/d, respectively. Interested readers can find further information on that paper, or in Carruba et al. (2018), for an implementation of that method by our group.

Table 5 displays the *AUC* coefficient, the *IN* and *OUT* values of the slopes approach, and the estimated ages, with their errors, obtained with Spoto et al. (2015) for the new large families (80 members or more) identified in this work, while the V-shape slopes in the (a,1/D) domains for these families are shown in Fig. 3. It is possible to find only a solution for the IN part of the slope of the Ekard family, while both solutions can be found for the Lydina group. However, because of the limited number of members in the Lydina family (93), errors are quite large. To within these large errors, the IN and OUT solutions for the age of the Lydina group appear to be in agreement. The fact that it is possible to date these two families suggest that results of machine-learning clustering algorithms appear to produce reliable new dynamical groups.

5 CONCLUSIONS

The main purpose of this work was to explore the possibility of using machine-learning hierarchical clustering algorithms for the purpose of asteroid families identification. We applied this method to five regions of the asteroid main belt that are not affected by the problem of chaining, i.e. the regions of the Hungaria asteroids, the inner, central and outer main belts at high inclinations, and the region of the Cybele asteroids. We compared the outcome of this new approach with the results of standard HCM for the region of the Cybele asteroids, and we found that the new method is able to retrieve family members with an accuracy, defined according to equation (2), above 89.5 per cent. All asteroids previously identified as members with standard HCM are also retrieved by the new approach. Values of the areas under the curve (AUC) coefficients below Receiver Operating Characteristic (ROC) curves for dynamical groups identified with this approach are also very good, and consistently higher than 85 per cent. The methods was able to retrieve all known asteroid families in the five regions, including new members. However, the new approach also consistently tended to identify more objects as family members than standard HCM, with a Precision coefficient defined using equation (3) as low as 36.0 per cent. False positive group members obtained by this method are the main limitation of this approach. Not necessarily all these false positive members are to be discarded, and the lower number of family members identified by standard HCM may be an indication of limitations of the standard method itself. But a cautious study and comparison of results obtained with the two methods is generally advisable.

Six new families and 13 clumps were found, for the first time, using this approach. Fig. 4 displays a (a, $\sin(i)$) projection of the families (shown as coloured full dots) and clumps (shown as points) in the high-inclination central main belt and in the Cybele region that were newly identified in this work. For brevity, we do not show the location of the (7758) Poulanderson clump identified in the inner main belt, and we do not print the names of the (4404) Enirac and (59511) (1999 JP14) clumps in the central main belt, to avoid overlapping of text. All clumps names in the central main belt are available in Table 4. The newly identified groups appear to be relatively uniform in terms of physical and taxonomic properties. For the two cases for which standard dating methods based on V-shapes in the (a, $1/D$) domain like those of Spoto et al. (2015) can be applied, those of the Ekard and Lydina families, we found that ages for these new families can also be obtained, which confirm

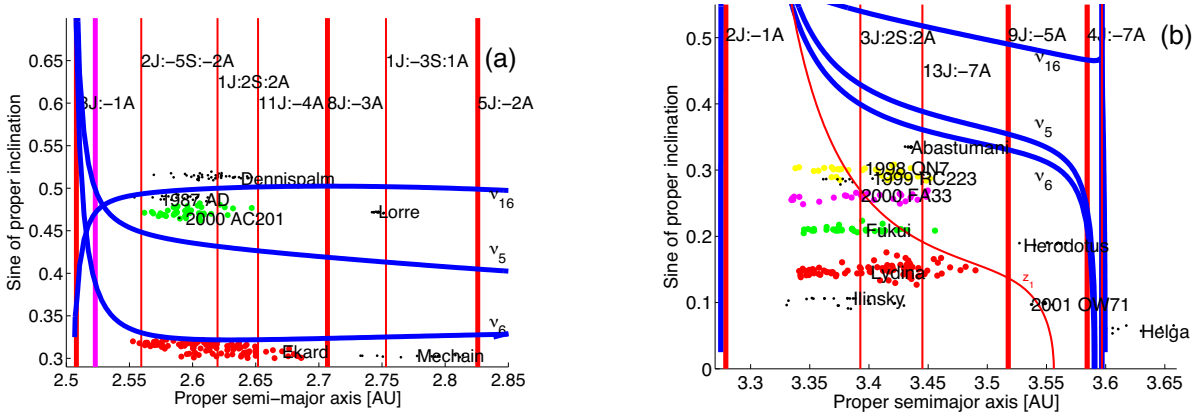

Figure 4. An (a, sin (i)) plot of the newly identified families (full dots) and clumps (points) in the high-inclination central main belt (panel a) and in the Cybele region (panel b).

the reliability of the outcome of machine-learning hierarchical clustering algorithms for this task.

Most of the newly discovered groups, with just one exception in the inner main belt, are found in the high-inclined central main belt and in the Cybele region. For the case of the central main belt, this can be explainable in terms of probabilities of collisions. Central main belt asteroids orbits can interact with both inner and outer main belt objects more easily than the 'fringe' regions, and, as a consequence, they are more likely to experience collisions than other asteroid populations. Most asteroid families are indeed found in the central main belt, and it is only natural to expect a higher rate of findings of new groups in this region. Regarding the Cybele asteroids, selection effects may have been at play in the past. Cybele asteroids tend to be dark, low-albedo objects. They are also further away from Earth than objects in other regions of the main belt, and therefore, they are more difficult to be found. Since new surveys since the work of Carruba et al. (2015) have significantly increased the population of Cybele asteroids, it is likely that the increased sample allowed for the identification of families that were not visible in the past.

Overall, we believe that this preliminary study showed that machine-learning clustering algorithms have great potential for the problem of asteroid families identification. While, computationally, standard and machine learning HCM are both quite efficient, one advantage of machine learning HCM is the facility with which new asteroid groups can be identified. Dendrogram clusters of asteroid distances, such those shown in Fig. 1, are produced almost instantaneously by machine learning HCM, and obtaining a list of clusters for a given orbital region is also a straightforward and rapid procedure. The opposite is not true for standard HCM, at least in the version implemented by our group. Stalactite diagrams used to identify families, like that used by Carruba et al. (2015) for the Cybele region, are computationally demanding and may miss some asteroid groups. The ease of use of machine learning HCM is, in our opinion, one of the big advantages of this method. Extending the use of these approaches to higher number-density regions remains a challenge for future works.

ACKNOWLEDGEMENTS

We are grateful to the reviewer of this paper, Dr. Bojan Novaković, for useful comments and suggestions. We would like to thank the São Paulo State Science Foundation (FAPESP, grant 2018/20999-6) and the Brazilian National Research Council (CNPq, grant 301577/2017-0). We acknowledge the use of data from the Asteroid Dynamics Site (AstDys; http://hamilton.dm.unipi.it/astdys, Knežević & Milani 2003). We are grateful to Edmilson Roma de Oliveira for discussions that motivated this work. This publication makes use of data products from the Wide-field Infrared Survey Explorer (WISE) and Near-Earth Objects WISE (NEOWISE), which are a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

REFERENCES

Beaugé C., Roig F., 2001, Icarus, 153, 391
Bendjoya P., Zappalà V., 2002, Asteroids III. Univ. Arizona Press, Tucson, p. 613
Bolin B. T., Delbo M., Morbidelli A., Walsh K. J., 2017, Icarus, 282, 290
Bolin B. T., Walsh K. J., Morbidelli A., Delbó M., 2018, MNRAS, 473, 3949
Carruba V., 2009, MNRAS, 398, 1512
Carruba V., 2010, MNRAS, 408, 580
Carruba V., Morbidelli A., 2011, MNRAS, 412, 2040
Carruba V., Domingos R. C., Nesvorný D., Roig F., Huaman M., Souami D., 2013, MNRAS, 433, 2075
Carruba V., Aljbaae S., Souami D., 2014, ApJ, 792, 46
Carruba V., Nesvorný D., Aljbaae S., Huaman M. E., 2015, MNRAS, 451, 4763
Carruba V., Nesvorný D., Aljbaae S., Domingos R. C., Huaman M., 2016, MNRAS, 458, 3731
Carruba V., Vokrouhlický D., Nesvorný D., Aljbaae S., 2018, MNRAS, 477, 1308
Carruba V., Spoto F., Barletta W., Aljbaae S., Fazenda A., Martins B., 2019, Nat. Astron, in press
Cox D., 1958, J. Roy. Stat. Soc. B., 20, 215
Delbó M., Walsh K., Bolin B., Avdellidou C., Morbidelli A., 2017, Science, 357, 1026
Delbó M., Avdellidou C., Morbidelli A., 2019, A&A, 624, A69
DeMeo F. E., Carry B., 2013, Icarus, 226, 723
Fawcett T. 2006, Pattern Recognit. Lett., 27, 861
Hirayama K., 1923. Annales de l'Observatoire Astronomique de Tokyo, 11, 55
Holt T. et al., 2017, AAS/Division for Planetary Sciences Meeting Abstracts #49, 49, 511.03
Ishihara D. et al., 2010, A&A, 514, A1

Ivezić Ž. et al., 2001, AJ, 122, 2749
Knežević Z., Milani A., 2003, A&A, 403, 1165
Masiero J. R., Mainzer A. K., Grav T., Bauer J. M., Jedicke R., 2012, ApJ, 759, 14
Masiero J. R., Carruba V., Mainzer A., Bauer J. M., Nugent C., 2015, ApJ, 809, 179
Milani A., Knežević Z., Novaković B., Cellino A., 2010, Icarus, 207, 769
Milani A., Cellino A., Knežević Z., Novaković B., Spoto F., Paolicchi P., 2014, Icarus, 239, 46
Milani A., Knežević Z., Spoto F., Paolicchi P., 2019, A&A, 622, A47
Novaković B., Cellino A., Knežević Z., 2011, Icarus, 216, 69
Novaković B., Hsieh H. H., Cellino A., 2012, MNRAS, 424, 1432
Novaković B., Tsirvoulis G., Granvik M., Todović A., 2017, AJ, 153, 266
Pedregosa F. et al., 2011, JMLR, 12, 2825
Pravec P. et al., 2010, Nature, 466, 1085
Pravec P. et al., 2018, Icarus, 304, 110
Ryan E. L., Woodward C. E., 2010, AJ, 140, 933
Spoto F., Milani A., Knežević Z., 2015, Icarus, 257, 275
Stehman S. V., 1997, Remote Sens. Environ., 62, 77
Vinogradova T., Shor V., 2014, Proceedings of Asteroids, Comets and Meteors 2014. Helsinki, Finland, p. 567
Zappala V., Cellino A., Farinella P., Knezevic Z., 1990, AJ, 100, 2030

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Appendix 1

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.