DATA MINING OF LARGE ASTRONOMICAL SURVEYS FOR PRECOVERY AND DISCOVERY OF NEAR-EARTH OBJECTS

T. Saifollahi[(1)], G. Verdoes Kleijn[(1)], O.R. Williams[(1)], E. Helmich[(1)], M. Micheli[(2)], L. Conversi[(2)], T. Santana-Ros[(3)], and D. Koschny[(4)]

[(1)]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands, Email: saifollahi@astro.rug.nl*
[(2)]*ESA NEO Coordination Centre, Largo Galileo Galilei, 1, 00044 Frascati, Italy Email: marco.micheli@esa.int*
[(3)]*Departamento de Fisica, Ingeniería de Sistemas y Teoría de la Señal, Universidad de Alicante, Carr. de San Vicente del Raspeig, s/n, 03690 San Vicente del Raspeig, Alicante, Spain Email: tonisantanaros@gmail.com*
[(4)]*ESA/ESTEC, Keplerlaan 1, 2200 XH Noordwijk, The Netherlands Email: detlef.koschny@tum.de*

ABSTRACT

The serendipitous appearances of Near-Earth Objects (NEOs) in a subset of the millions of archival exposures of optical and near-infrared astronomical imaging surveys can improve our knowledge of orbits and compositions of NEOs. We show how the data processing and data mining of such imaging archives can be exploited to identify new and known NEOs, leading to a re-assessment of the impact probability of hazardous NEOs for the latter. We describe our automatic pipelines that precovers[1] and discovers NEO appearances making use of the Astronomical Wide-field Imaging System for Europe (ASTROWISE). ASTROWISE is an information system that contains a 10 Terabyte-scale database for data mining that is connected to a Petabyte scale imaging storage archive from various telescopes. As a pilot study, we performed a systematic search for the ESA risk-list NEOs appearances in a decade of archival observations of the OmegaCAM optical wide-field imager at ESO's VLT Survey Telescope. The observatory has been used for several multi-year large surveys (including the Kilo-degree Survey, VST-ATLAS, Fornax Deep Survey and VPHAS+) plus many smaller programs. None of these surveys and programs are dedicated to the detection and surveillance of NEOs. Our current NEO precovery pipeline detects 196 NEO appearances and the discovery pipeline is expected to discover order 50 new NEOs in the Kilo-Degree Survey. The NEO precovery and discovery can be expanded to other archives of wide-field imaging instruments/surveys thanks to the homogenized metadata interface offered by ASTROWISE to such archives.

Keywords: Near-Earth Objects, Wide Field Surveys, Observations, Astrometry, Photometry, Data Analyses.

1. INTRODUCTION

Astronomical surveys not specifically designed for detecting and monitoring Near-Earth Objects (NEOs) may still be valuable in precovery, discovering, detecting, and assessing the risk of these objects. These surveys have the potential to complement dedicated NEO surveys in two ways: they typically use larger telescopes and longer exposure times, allowing for the detection of smaller and more distant NEOs, and they often survey areas away from the ecliptic plane, allowing for the detection of NEOs on highly inclined orbits. However, these surveys may not be optimized for detecting moving objects, and their primary focus is typically on galactic and extragalactic science. If it can be demonstrated that these surveys can make a significant contribution to our understanding of NEOs, they could have both scientific and societal benefits. This paper describes an exploratory pilot project to investigate these benefits, which involves reusing astrophysical imaging data for NEO precovery and discovery and adapting an existing information system called ASTROWISE to work with a range of astronomical archives in a single data flow environment using a single pipeline.

2. DATA AND SOFTWARE SYSTEMS

Archival imaging data from OmegaCAM, a wide-field camera on the VLT Survey Telescope (VST) at ESO's Cerro Paranal Observatory, are used for the search for previously undiscovered NEOs reported here. OmegaCAM has 32 science CCDs with a field of view of approximately 1 degree by 1 degree. Over the first decade of its operation, OmegaCAM has covered a significant portion of the southern hemisphere (Fig. 1) and is ideal

[1]Throughout this paper, the term precovery includes both precovery and recovery. The term "precovery" refers to extracting an NEO appearance from an exposure taken at a time predating the time of the observation sequence in which the NEO was discovered, while "recovery" is extracting an NEO appearance from an exposure taken at a time later than the discovery sequence.

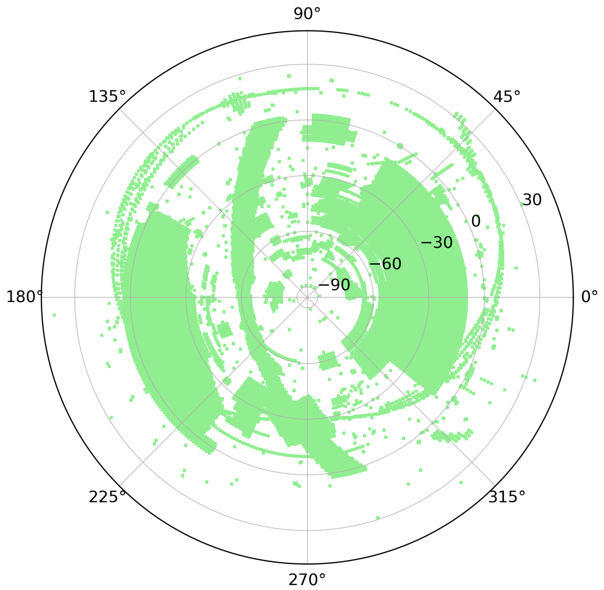

Figure 1. The sky coverage in (RA, Dec) of the OmegaCAM/VST observations over a decade of observations is shown for declinations between +45^0 and -90^0. The ASTROWISE *archive contained 361977 individual exposures at the creation date of this plot (mid-2022).*

for NEO precovery. OmegaCAM has been used for several wide-field surveys, including the Kilo-degree Survey (KiDS) [1] which is used is this paper.

The precovery effort described here is focused on two datasets of known NEOs: the "risk-list" provided by the European Space Agency's Near-Earth Objects Coordination Centre (NEOCC), which includes about 1 350 NEOs with a non-negligible chance of impact in the next 100 years, and a "full-list" of all known NEOs, approximately 30 000 sources.

The pipelines that have been developed for this study use the Astronomical Wide-field Imaging System for Europe (ASTROWISE) system ([2]) for data management, image processing, and calibration. Several public web-tools were used to collect the predicted properties of NEOs. In addition, several dedicated software applications were used for streak detection, astrometry and photometry.

3. NEO PRECOVERY PIPELINE

The precovery pipeline has four main steps. These steps are briefly described here. For more details on the pipeline, we refer the reader to [3].

Step 1: A preliminary spatial and temporal cross-match of NEOs with the KiDS archival images. The (SSOIS)[2] web service [4] was used to generate an ephemerides database, which was then cross-matched with the OmegeCam observation date, RA, and Dec.

[2]https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/ssois/

Table 1. Summary of the filters applied to the NEO candidates in step 3 of the precovery pipeline

Precovery Step	value for risk-list/ full-list
Initial SNR cut (in step 2)	1 / 1
Upper limit on trail length	— / 3”
Upper Limit on angular separation	1° / 1°
Upper limit on 1σ errors in position	1° / 1°
Lower limit on 1σ errors in position	— / 1”
Limit on SNR	3 / 3

Step 2: A final spatial and temporal cross-match of NEOS with the KiDS archival images. The pipeline uses the JPL Horizons[3] web-service to obtain more accurate predictions of the NEO positions over the duration of exposures and to obtain the predicted angular motion and visual magnitude.

Step 3: Data Processing and Astrometry. Several filters were applied to the results from step 2 to narrow the search to NEOs which are likely to be both detectable and interesting (see Table 1). Once a frame passed all the filters, the pipeline used ASTROWISE to produce astrometrically and photometrically calibrated pixel images. SCAMP 2.10 [5] was used to astrometrically recalibrate the detector images using stars in the *Gaia* EDR3 [6] catalogue as astrometric reference objects.

Step 4: Streak Detection and Association. the pipeline deploys STREAKDET [7] on the photometrically and astrometrically calibrated images to extract candidate streaks. STREAKDET is a package that was developed for the purpose of identifying streaks in ground-based and space-based imaging data. Once streak detection is done, then, the pipeline searches for the best match between the streaks found within the 3σ uncertainty ellipse of NEOs and their predicted properties. The predicted properties used include the NEOs length, the direction of motion and the magnitude. This procedure is illustrated in Fig. 2. Finally, those frames with a matched streak are visually inspected to check if it is a true precovery.

4. NEO DISCOVERY PIPELINE

The NEO discovery pipeline is implemented to be able to detect new solar system objects (SSOs) in the KiDS Survey observations [1]. Here we chose to identify asteroids in the KiDS images using the method described in

[3]https://ssd.jpl.nasa.gov/horizons/app.html

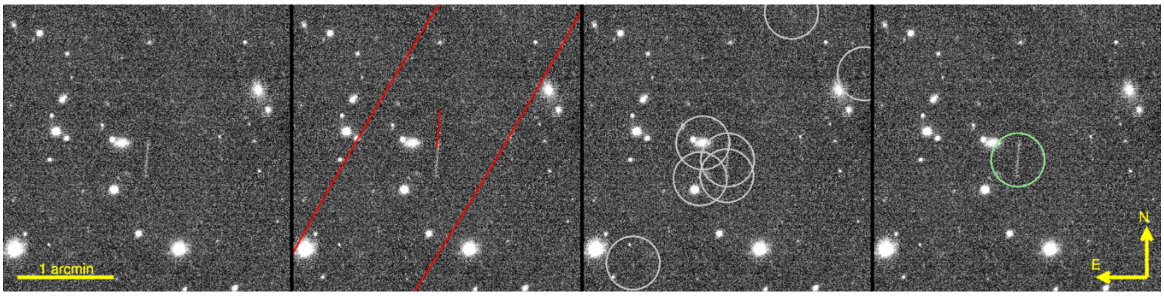

Figure 2. The result of the precovery pipeline for one NEO is shown (as an example). The first column (from left to right) displays the calibrated image cutouts around the predicted position of the NEO. The second column displays the 3σ uncertainty ellipse, as well as the predicted appearance of the NEO taking into account its proper motion and direction of motion (red ellipse and red line). The third column shows the streaks identified by STREAKDET *(grey circles). The fourth column shows the streak that matches the predicted properties of the NEO (green circles).*

[8]. The method takes advantage of the dithering strategy implemented by KiDS. Many SSOs display a motion large enough (from a few to many arcseconds per hour) to make their identification possible by looking for objects that move across the dithers. The theoretical lower limit to the proper motion detectable in KiDS is derived using the sequence duration and observed residual dispersion in the astrometric repeatability. The upper limit is set by the chosen cross-match radius and sequence duration. The proper motion range encompasses the anticipated proper motions for most Near-Earth Asteroids. Furthermore, the KiDS data products that were used are photometrically and astrometrically calibrated frames along with their respective weight maps. These are all produced with ASTROWISE.

In each individual image of a dithered sequence, all the sources are detected and their centroid position is measured using SEXTRACTOR [9]. Using this data as input, the pipeline deploys SEXTRACTOR (on each of the 32 detector exposures independently) to detect sources on the images. Once all the source lists are made, the pipeline combines the SEXTRACTOR output files in a given field and a given filter and spatially associates the detected sources across the 4 or 5 dithered exposures in a field. This results in a table file that contains sets of extracted parameters on sequences of 4 or 5 detections across dithers that are from potentially a single physical object. Subsequently, the pipeline selects from these sequences those that are likely detections of a real single physical object with detectable proper motion consistent with being an SSO. We use the term "candidate SSO sequence" to denote these. The candidate SSO sequences are distilled from the list of sequences by requiring the information in the sequences to satisfy a series of criteria which were developed to remove sequences due to spurious source detections and/or without detectable proper motion. The criteria require an appearance in a sufficient number of dithers, consistency with constant linear motion across the dithers, appearance not too close to bright stars (to avoid contamination by bright star artifacts) and proper motion values consistent with SSOs. We refer to [8] for the detailed rationale for these criteria.

To optimize the discovery of Near-Earth Objects we configured the pipeline with a set of values for parameters used in the criteria to optimize it for NEO discovery. These parameters are slightly different from the choices by [8]. as they optimized the criteria to find SSOs in general, in particular ones with a small proper motion. Additionally, we included a script that submits the NEO observations to the NEO rating web-tool[4] of the Minor Planet Center (MPC) that obtains the NEO rating and the root-mean-square (RMS) values estimated by this service. The latter is the RMS of the observed positions of an object around a linear line. Additionally, we estimated the RMS of magnitudes and ellipticities of detected moving objects to further filter spurious detections. Here we evaluate the efficiency of this NEO discovery pipeline by searching for the NEOs that are precovered in the KiDS dataset (earlier using the precovery pipeline).

5. RESULTS PRECOVERY PIPELINE

Astrometric and photometric accuracies: The astrometric accuracy is estimated to be at most 0.12" (2D). The photometric accuracy is estimated to be about 0.1 mag. The astrometric accuracy will be improved in the future by propagating the proper motions in the Gaia astrometric reference catalogue to the observation date of the science image.

NEO precovery rate: We define NEO detectability as the fraction of detectable appearances among the total occurrences that a NEO is predicted to be located within the FoV of the images. Table 2 summarizes the down selection of precovery candidates during the deployment of the pipeline and illustrates that the dominant reason for assuming non-detectability is an SNR lower than 3. The NEO detectability in the OmegaCAM archive is 0.005. We expect no significant improvement can be made in detectability given the low 3σ threshold in predicted SNR and the fact that detected NEOs tend to be often a few

[4] `https://www.minorplanetcenter.net/iau/NEO/PossNEO.html`

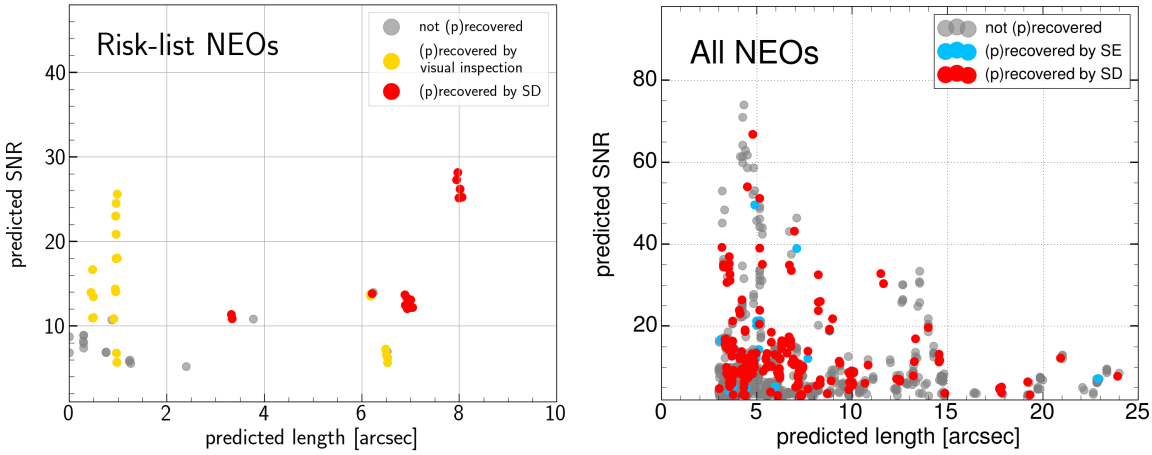

Figure 3. SNR and length for precovered NEOs of the risk-list (left) and the full-list (right).

tenths of magnitude fainter than predicted. The fraction of NEOs predicted to be detectable that is actually precovered is 40% for NEOs on the risk-list and 20% for the full-list of NEOs. Fig. 3 shows the predicted properties of detected and non-detected cases. STREAKDET is effective for detecting high SNR streaks of sizes between 5-20", but its performance decreases for faint and long streaks. We propose that utilizing deep learning techniques may improve streak detection and NEO precovery. In conclusion, a factor of up to 5 more NEOs might be precovered from the OmegaCAM archive. Furthermore, the current detection of risk-list NEOs led to the removal of three of them from this list. This could be increased by a factor of up to 2 by improving the recovery process. Alternatively, if failed recoveries turn out to be truly non-recoveries after improving the precovery processes, this is a significant result as well. It would suggest that the actual orbital accuracy for those objects on the risk-list is significantly worse than predicted. For an in-depth discussion about the precovery results for OmegaCAM and the feasibility of deploying it to other surveys next, we refer the reader to [3].

6. RESULTS DISCOVERY PIPELINE

NEO detection rate: the Discovery Pipeline was deployed on $\sim$650 deg^2 of KiDS area in its four filters. This resulted in the detection of 30621 SSO sequences in total of which 1101, 11493, 9628 and 8399 were in u, g, r and i respectively. About 34% of these are real discoveries, i.e., unknown SSOs. There are 61, 4183, 4410 and 1667 discoveries in u, g, r and i respectively. Within the already known detections 0.1% are NEOs and 98% are Main Belt Asteroids [8]. Simple extrapolation by area and assuming the same fraction of discoveries for all SSO classes suggest the Discovery Pipeline will detect about 110 NEO sequences in the full KiDS area. Of these about 41 NEOs would be true discoveries and NEOs and 69 already known NEOs. The previous rough estimate appears conservative by factor 3 compared to the 40 unique NEOs already precovered in KiDS and given the 20% precovery rate. So the total set of NEO sequences in the full KiDS area might be as large as 300. The performance of the discovery pipeline for detecting NEOs and for classifying them as such was assessed using 40 unique NEOs precovered in the KiDS data set (each NEO is detected 4-5 times, therefore the total number of NEO precoveries is a factor of 4-5 more). Out of 40 NEOs, 13 cases (about 32%) did not end up in the final output of the pipeline: the SSO detection catalogue. did not pass the initial detection filters (numbers of detections and linearity). The majority of these cases were caused by issues with the software used for merging catalogues and sources, SCAMP, which did not match objects correctly. The majority of these cases are caused by the fact that the pipeline uses SEXTRACTOR which is not intended for the detection and characterisation of streaks. It often segments NEO streaks into multiple sources and/or does not identify the middle of the streak as its centroid. The subsequent grouping of detections belonging to the NEO into sequences then fails and/or is contaminated by non-NEO detections. This results in discarding these sequences as SSO candidates in subsequent steps. In conclusion, the current pipeline detects about 67% of the NEOs.

NEO classification rate: For the remaining 27 NEOs (about 67%), a NEO rating of 32 and larger was found, with 75% of them having a NEO rating larger than 75. Additionally, the values of RMS for linearity, magnitudes and ellipticities of detections were within 0.01-0.5, 0.0-0.3, and 0.0-0.1. Using a NEO threshold of 32 and requiring the three RMS values to lie within the ranges found for the 27 NEOs, 122 additional sequences are classified as NEO candidates. If all of these sources were false-positives (FP), it implies that the classification criteria select 100% of the NEOs (27 NEOs) but lead to a NEO to FP ratio of 1 to 5. To improve this ratio we can set the classification criteria to stricter values of 0.01-0.25, 0.0-0.1, and 0.0-0.1, respectively. Then 85% of the known NEOs (23 NEOs) are classified as such and an additional 27 sequences are classified as NEO candidates. These roughly equal numbers of known and new NEOs

Table 2. Summary of ESO OmegaCAM Precovery for the risk-list and the full-list of NEOs.

Precovery Step	Number of precovery candidates selected in this step	Number of precovery candidates removed in this step	Percentage of precovery candidates removed in this step
Risk-list			
Query SSOIS	10 345	-	-
Query NEODyS and an initial SNR cut	441	9 904	∼96%
Upper Limit on angular separation limit	295	146	∼33%
Upper limit on 1σ errors in position	251	44	∼15%
No raw data available on the AW database	236	15	∼6%
Not covered by the camera (OmegaCAM/VST)	170	66	∼28%
Limit on SNR	68	102	∼60%
Precovered	27	41	∼60%
Full-list			
Query SSOIS	186 476	-	-
Query Horizons and an initial SNR cut	55 692	130 514	∼70%
Lower limit on the predicted length	8 440	47 252	∼85%
Upper limit on angular separation limit	7 875	565	∼7%
Upper/lower limits on 1σ errors in position	2 683	5 192	∼66%
No raw data available on the AW database	2 463	220	∼9%
Not covered by the camera (OmegaCAM/VST)	2 251	212	∼9%
No calibrated data available (calibration failed)	2 231	20	<1%
Limit on SNR	968	1 263	∼57%
Precovered	196	772	∼80%

are similar to the roughly equal numbers of known and new SSO discoveries. Visual inspection of the 27 sequences can weed out non-SSOs among them. For the remainder follow-up observations on the same or the next night would be required to verify which of those are indeed NEOs instead of other classes of SSO. In conclusion, for a KiDS-like observing strategy, it appears feasible that the discovery pipeline can deliver a list of candidate NEOs for follow-up observations that have 50% purity and 85% completeness which could lead to up to 150 new NEOs. For Planetary Defence it would be good to assure no hazardous NEOs are among such serendipitous discoveries in astronomical surveys by complementing such surveys with a fast NEO follow-up campaign. The same telescope or other telescopes can be considered for such a follow-up campaign "piggybacking" on astronomical surveys.

REFERENCES

1. Kuijken, K., Heymans, C., Dvornik, A., et al. The fourth data release of the Kilo-Degree Survey: ugri imaging and nine-band optical-IR photometry over 1000 square degrees. *A&A*. **625** pp. eA2 (2019,5)
2. McFarland, J., Verdoes-Kleijn, G., Sikkema, G., et al. The Astro-WISE optical image pipeline. Development and implementation. *Experimental Astronomy*. **35**, 45-78 (2013,1)
3. Saifollahi, T., Verdoes-Kleijn, G., Williams, O., et al. Mining archival data from wide-field astronomical surveys in search of near-Earth objects. *submitted to the Astronomy and Astrophysics (A&A)*
4. Gwyn, S., Hill, N. & Kavelaars, J. SSOS: A Moving-Object Image Search Tool for Asteroid Precovery. *PASP*. **124**, 579 (2012,6)
5. Bertin, E. Automatic Astrometric and Photometric Calibration with SCAMP. *Astronomical Data Analysis Software And Systems XV*. **351** pp. 112 (2006,7)
6. Gaia Collaboration, Gaia Early Data Release 3. Summary of the contents and survey properties. *A&A*. **649** pp. eA1 (2021,5)
7. Virtanen, J., Poikonen, J., Säntti, T., et al. Streak detection and analysis pipeline for space-debris optical images. *Advances In Space Research*. **57**, 1607-1623 (2016,4)
8. Mahlke, M., Carry, B. & Denneau, L. Asteroid phase curves from ATLAS dual-band photometry. *Icarus*. **354** pp. e114094 (2021,1)
9. Bertin, E. & Arnouts, S. SExtractor: Software for source extraction.. *A&AS*. **117** pp. 393-404 (1996,6)