Asteroids Comets Meteors 2005
Proceedings IAU Symposium No. 229, 2005
D. Lazzaro et al. eds.

DOI: 00.0000/X000000000000000X

Unbiased orbit determination for the next generation asteroid/comet surveys

A. Milani[1], G. F. Gronchi[1], Z. Knežević[2], M. E. Sansaturio[3], O. Arratia[3], L. Denneau[4], T. Grav[4], J. Heasley[4], R. Jedicke[4], J. Kubica[5]

[1] Department of Mathematics, University of Pisa, Via Buonarroti 2, 56127 Pisa, Italy
email: milani@dm.unipi.it
[2] Astronomical Observatory, Volgina 7, 11160 Belgrade 74, Serbia and Montenegro
[3] E.T.S. de Ingenieros Industriales, University of Valladolid Paseo del Cauce s/n 47011 Valladolid, Spain
[4] Institute for Astronomy, University of Hawaii, Honolulu, HI, 96822
[5] Carnegie Mellon University, Robotics Institute, Pittsburgh, PA, 15213

Abstract. In the next generation surveys, the discovery of moving objects can be successful only if an observation strategy and the identification/orbit determination procedure are appropriate for the diverse apparent motions of the target sub-populations. The observations must accurately measure the displacement over a short interval of time; observations believed to belong to the same object have to be connected into *tracklets.* Information contained in tracklets is in most cases not sufficient to compute an orbit: two or more of them must be *identified* to provide an orbit. We have developed a method for recursive identification of tracklets allowing an unbiased orbit determination for all sub-populations and efficient enough to cope with the data flow expected from the next generation surveys. The success of the new algorithms can be easily measured only in a simulation, by consulting a posteriori some "ground truth".

We present here the results of a simulation of the orbit determination for one month of operations of the future Pan-STARRS survey, based upon a Solar System Model with a downsized population of Main Belt asteroids and a full size populations of Trojans, NEO, Centaurs, Comets and TNO. The results indicate that the method already developed and tested to find identifications of NEO and Main Belt asteroids are directly applicable to Trojans. The more distant objects often require modified algorithms, fitting orbits with only 4 parameters in a coordinate system specially adapted to handle very short arcs of observations. These orbits are mostly used as intermediate results, allowing to find full solutions as more tracklets are identified.

When the number density of detections is as large as expected from the next generation surveys, both joining observations into tracklets and identifying tracklets can produce some false results. The only reliable way to remove them is a procedure of tracklet/identification management. It compares the tracklets and the identifications with a complex logic, allowing to discard almost all the false tracklets and all the false identifications. However, the distant objects still present a challenge for orbit determination: they require three tracklets in separate nights. If this requirement is met we have found no problem in achieving an unbiased orbit determination for all populations. Further work will lead to more advanced simulations, in particular by introducing a realistic model for astrometric and photometric errors.

Keywords. surveys, orbit determination, identification, population models

1. Introduction

Most of the next generation astronomical surveys are going to be "general surveys", with the goal to discover and catalog as many astronomically significant objects as possible with the available resources, rather than "specialized surveys" dedicated to one specific population. That is, with the same telescope and camera, if possible on the same images, they will try to discover and measure whatever can be detected, without restricting the discoveries to the interests of a specific astronomical sub-community. This approach can have great advantages with respect to restricted interest surveys, which are too often forced to use marginal resources. However, is it really possible to share the resources with satisfactory efficiency for all the subprojects?

In this paper we will address the problems arising in the discovery of moving solar system objects†. The target populations could be asteroids (Near Earth, Main Belt, Jupiter Trojans), comets (short and long period), Trans Neptunian Objects (regular, Plutinos, scattered disk), Centaurs, Trojans of other planets, satellites of the outer planets, and even classes of objects not discovered yet. Can all these populations be the target of the same survey? For this there are two main critical issues: the observation strategy and the identification/orbit determination procedure.

To observe an object moving with respect to the fixed stars the survey needs to take a sequence of images of the same field: *detection* can take place by having just $m = 2$ images with a time interval Δt long enough to be able to measure the displacement, and short enough to be able to connect into *tracklets*‡ the individual observations believed to belong to the same object; the procedure to form tracklets from individual observations is discussed in Section 4. The optimal value for Δt depends upon the *number density* of the detectable moving objects per unit area on the images and upon the *proper motion* (angular velocity on the celestial sphere), with typical values between 15 minutes and 1 hour. The first critical issue can thus be summarized as follows: can we select m, Δt and the scanned portion of the sky in a way optimal for all moving objects populations? Of course not, but it appears to be possible to find a reasonable compromise solution. We shall show in Section 3 that an interesting mix of populations can be observed on the same images, and discuss in Section 6 the problems arising from this compromise in the orbit determination of the fastest moving and the slowest moving objects.

The second critical issue arises from the fact that a tracklet in most cases does not allow to determine an orbit and to establish to which population the detected object belongs. Even for $m > 2$, if Δt is short, a tracklet contains only information which can be summarized in an *attributable*, a 4-dimensional vector containing two angles and two angular rates (a vector tangent to the celestial sphere): orbital elements containing 6 independent parameters cannot be uniquely determined. A tracklet with this property is called a *Too Short Arc* (TSA). The only way to compute a unique orbit is to *identify* two TSAs, separated by a time interval much longer than Δt, as belonging to one and the same object. Then there are more equations than unknowns, and a least squares solution is well defined. The classical methods for identification and orbit determination are not suitable to cope with the data flow expected from the next generation surveys. Thus it has been necessary to develop entirely new algorithms, to code them in efficient software and to perform extensive tests without waiting for the real data. So far, the interest has been concentrated on the specific problems of orbit determination for NEO, both because it can be more difficult and because of the immediate interest in impact monitoring. However, distant objects are a serious challenge for orbit determination because the angular rate relative accuracy is intrinsically very poor. The specific problems arising for distant objects and the new algorithms required are discussed in Section 5.

The success of the algorithms to find identifications can be assessed in terms of completeness and reliability. These can be easily measured only in a simulation (such as the

† The problem can be even more general, e.g., sharing resources with searches for supernovae and candidate gamma ray burst sources.

‡ In Milani *et al.* 2004 we use Very Short Arc instead of tracklet, for the same concept.

one presented in Section 3), by consulting a posteriori some "ground truth" specifying which observation belongs to which object. In this case, completeness is the ratio between the objects for which identifications and orbits (of a specified quality) have been obtained and the total number observed, (lack of) reliability is the ratio between *false identifications* and *true* ones. An identification is false if it contains observations belonging to different objects; also a tracklet can be false. As the survey goes deeper in apparent magnitude the number density increases and the problem of false tracklets and identifications becomes more serious. Section 6 discusses how high levels of completeness and reliability can be achieved, and Section 7 assesses the quality of the resulting orbits.

The studies described in this paper are intended as preparation for one of the next generation survey, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), an all-sky survey telescope under development at the University of Hawaii. Pan-STARRS is composed of 4 individual 1.8 m telescopes observing the same region of sky simultaneously. Each telescope will have a $\simeq 7$ deg^2 field of view and will be equipped with $\simeq 1$ billion pixel CCDs in the focal plane, resulting in a spatial sampling of $\simeq 0.3$ arcseconds per pixel. With exposure times of ~ 30 s it is estimated that the system will reach a limiting magnitude of $V \sim 24.5$. The design of Pan-STARRS is weighted toward its primary purpose, which is to detect potentially hazardous Solar System objects. However, the wide-field, repetitive nature of the system make it ideal to detect a host of other astronomical phenomena, ranging from Solar System to cosmology.

A single 1.8 meter telescope prototype, essentially a one quarter part of the full system, is currently being built on Haleakala in Maui, Hawaii. This prototype will allow for testing all the technology that is being developed for Pan-STARRS and will be used to make a full-sky survey to provide an astrometric and photometric calibration data set that will be used for the full system. First light on this prototype is scheduled for early 2006.

One of the features of this project is that the survey will be fully simulated to develop and test the observing strategy and the data processing chain, from raw images to orbits, before real data are processed. This has the purpose of achieving optimum performance from the very beginning of the operational life rather than having to solve the problems a posteriori, when it might be late for some corrections.

The simulation requires an assumed solar system model. If the goal is to show that the survey can achieve a satisfactory discovery rate on different populations, then the model must include a representative sample of each target population. Although population models were available for some populations, a global model of all the objects observable (in a survey much deeper than the current ones) had to be build specifically for this purpose, see Section 2. This model was used in an *observation simulation* to show that even less numerous and dimmer populations of moving objects can nevertheless be observed and successfully pass through the procedure ending in a reliable orbit catalog.

2. The Pan-STARRS Solar System model

After ten years of operations Pan-STARRS should have detected roughly 10 million small objects of the Solar System. The Moving Object Processing System (MOPS) churns through these detections and identifies those corresponding to known and new objects resulting in a final database of orbits for all of them. To test and compare old and new algorithms for these tasks we are developing a realistic model of the small body populations of the Solar System that could be observed during the lifetime of Pan-STARRS. The use of a realistic model will ensure that the MOPS efficiency at finding tracklets/identifications and at computing orbital parameters is tested well before the system becomes operational.

It will eventually also allow the MOPS to monitor its efficiency while it is operating, by processing synthetic data in parallel with the real data.

To our knowledge this is the first attempt to create a model of the solar system for all populations of small bodies, including those recently discovered. As an example, the Statistical Asteroid Model (Tedesco *et al.* 2005) provides only a model of the Main Belt. Unfortunately, we found it impossible to create models for each of the solar system's small body populations using a single algorithm. Our knowledge and the distribution of the members for each population varies greatly and this required a different technique for generating each synthetic sub-population. Furthermore, the model will incorporate objects for which very few members are known (e.g. objects interior to the Earth's orbit) or populations that are entirely unknown (e.g. interstellar objects, distant major planets). It will be released on the Internet and described in detail in another paper soon.

Since there exist no directly measured size distributions of small bodies we chose to create the SSM based on measured absolute (H) or apparent magnitude distributions.

• **Near Earth Objects (NEO) and objects Interior to Earth's Orbit (IEO).** The NEO model is based on the orbit element and absolute magnitude distributions of Bottke *et al.* 2002 (for $H < 24$) and the H distribution of Rabinowitz *et al.* (2000) for $H \geqslant 24$. IEOs are included as a consequence of the transport model generating the known NEO population. Very small objects could be detected by Pan-STARRS when passing very close to Earth, the lower limit being set by the *trailing loss*†. We chose to arbitrarily set the limit at $H = 25$, corresponding to asteroids of diameter 35m/70m (depending upon the composition), unlikely to result in damage if they impact on Earth. This is below the recommended diameter limit of 100m for the next series of NEO surveys as promoted by Stokes *et al.* (2003). According to the size distributions cited above, this choice results in a model population of about $250,000$ NEOs.

• **Main Belt Objects (MBO).** We expect that there are $\sim 10^7$ small objects within reach of Pan-STARRS. It is believed (Jedicke *et al.* 2002) that the known MB sample is nearly complete for $H < 14.5$ (diameters greater than about 6.5 km) based on the lack of new discoveries of asteroids in that range. Our synthetic model was generated by randomly selecting known MB asteroids with $H < 14.5$ and cloning them until we reached 10 million objects. The cloning process 'smears' each orbital element and creates a 'fuzzy' MB. This synthetic model relies on the assumption that there is no or little dependence of the orbit distribution of MB objects with their size so that the distribution of orbital elements for objects with $H < 14.5$ should be representative of the unbiased MB orbit population. The H distribution for these objects is as specified in Jedicke *et al.* (2002) who recommend that the known H-distribution be used for objects larger than the completeness limit and the Sloan Digital Sky Survey H-distribution (Ivezić *et al.* 2001) for smaller objects. To reduce the number of objects in the model it includes only those objects that can achieve $V < 24.5$ at opposition.

• **Trojans (TRO) of outer planets.** The Pan-STARRS SSM currently contains synthetic Jupiter Trojans whose orbit elements were generated without correcting for observational selection effects in the known population. Briefly, the observed population distribution in semi-major axis, eccentricity, inclination, and mean longitude difference with Jupiter's were fit to reasonably shaped analytic functions. The synthetic population was then generated randomly from those functional forms, using the size distribution from Jewitt and Luu (2000)‡, resulting in $320,000$ objects. We have generated $20,000$

† A very close object has a large proper motion, thus its image spreads on many pixels.

‡ Jewitt and Luu (2000) provided results only for L4. We assumed a L5 cloud identical to the L4 cloud. This implies the L5 population might be overestimated with respect to the L4 one.

synthetic Trojans of Mars, 40,000 of Saturn, 20,000 of Uranus and 20,000 of Neptune in the same manner as for the Jupiter Trojans but then rotating and scaling the orbital elements to the appropriate position and distance of each outer planet.

- **Centaurs (CEN).** The Centaur model is based on the observational analysis of Jedicke and Herron (1997) who fit the dynamical 'residence-time' distribution of the Centaurs in (a, e, i) space from Duncan *et al.* (1995) to reasonable analytic functions. The differential H distribution was modelled as $n(H) \propto 10^{0.61\times H}$ (Jedicke & Herron 1997, Sheppard *et al.* 2000). Over 60,000 synthetic Centaurs were generated according to the analytic distributions with minimum opposition magnitude $V \leqslant 24.5$.
- **Trans-Neptunian Objects (TNO).** In the Pan-STARRS SSM the TNOs encompass both the classical and resonant types. Following Levison & Morbidelli (2003) and Gomes *et al.* (2004), the synthetic population was generated using a symplectic integrator (Wisdom & Holman 1991) to model the migration of the outer solar system into a near planar disk of small bodies. The migration causes a large number of objects to be scattered inwards and later to be ejected completely due to close encounters with Jupiter and Saturn. Other objects get caught in resonances with Neptune that sweep through the disk creating a large population of resonant objects. The synthetic objects for our SSM were derived from the 7,159 stable orbits from the simulation which were sampled at thousand year intervals. The position and velocity of the objects were rotated and scaled such that Neptune at the time the orbital elements were extracted was close to the position of Neptune at some chosen epoch. The new position and velocity were used to compute the osculating orbital elements: 72,000 simulated objects have been generated. The absolute magnitude was then chosen from an apparent magnitude distribution consistent with Bernstein *et al.* (2004).
- **Scattered Disk Objects (SDO).** With only ~80 SDO currently known it is exceedingly difficult to determine the unbiased orbit distribution of these objects. We assumed that the a, e and i distributions were independent and applied a reasonable but *ad hoc* bias correction to each and then fit a common functional form to the corrected distribution. The synthetic orbits for the SDOs were randomly selected from those fits. The apparent magnitude distribution of the SDOs was taken from Elliot *et al.* (2005). Requiring that each synthetic object be brighter than $V = 24.5$ at some time during Pan-STARR's operational lifetime we generated over 20,000 SDOs.
- **Short Period Comets (SPC) and Long Period Comets (LPC).** As was done for the SDO, the synthetic SPC and LPC populations for this preliminary SSM is based only on the observed orbit distribution for each class of object. Analytical functional forms were fit to the observed distributions and synthetic orbit elements were then generated randomly according to those distributions. The apparent magnitude distribution of objects at perihelion was chosen to be $N(> R) \sim e^{0.04(R-24.5)}$. Only objects that reached $V \leqslant 24.5$ during the ten years of Pan-STARRS operation were included in the model, for a total of 10,000 SPC and 10,000 LPC.

3. Survey Simulation

The Pan-STARRS Solar System Survey Simulation (S^4) used in this work is a single lunation from a multi-year S^4; the details will be described in a future paper.

In an ecliptic reference frame centered on opposition (λ', β) the S^4 is defined by two near-quadrature areas with $|\beta| < 10°$, $-120° < \lambda' < -90°$ or $+90° < \lambda' < +120°$ (~550 deg^2 each) and also the opposition region with $|\lambda'| < 30°$ and $|\beta| < 40°$ (~4550 deg^2). A single Pan-STARRS field covers about 7 deg^2 so that each near-quadrature area needs at least 84 fields to be covered, while the opposition region

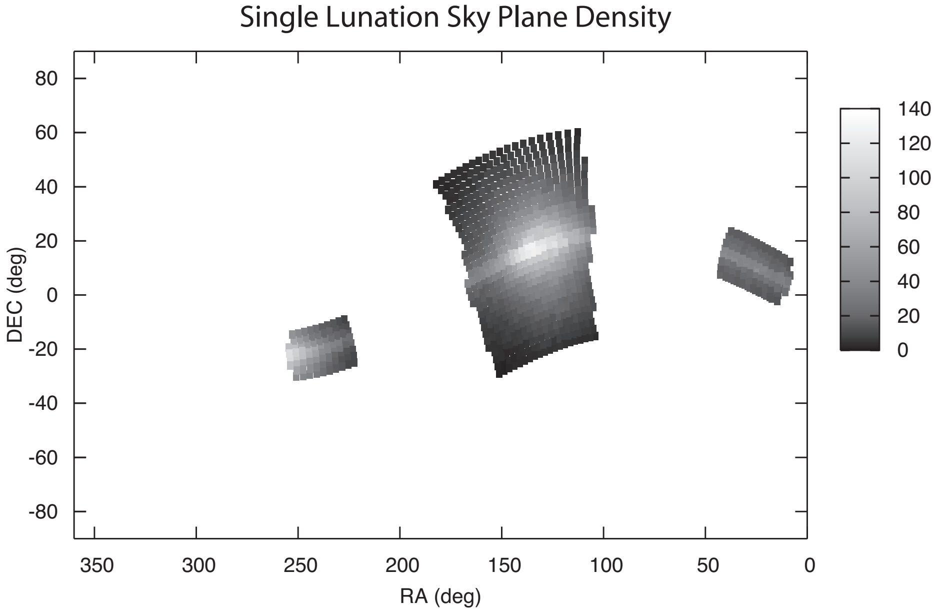

Figure 1. Sky plane density (number per square degree) of all types of objects in the survey simulation used in this work. Note the peak density corresponding to the center of the L5 Trojan cloud in the opposition region, and the signature of the L4 cloud in the quadrature area on the left.

requires at least 660 pointings. The nominal survey requires about 5 nights of clear sky per lunation and provides 3 visits to each field within that time† Each field is visited twice with about a 15 minute time separation and we required that the minimum altitude for observations be 20°. We used Paulo Holvorcem's TAO (http://pan-starrs.ifa.hawaii.edu/project/MOPS/tao.html) field scheduler to handle the scheduling within a night and wrote our own wrapper routine to handle multi-night observing including a crude approximation of the weather. The efficiency of scheduling the selected fields in all the regions was close to 100%.

To generate synthetic data for this work we determined the astrometric position and apparent magnitude for those objects with $V \leqslant 24.5$ in the SSM that appear in the simulated survey fields. This simulation used only $1/100^{th}$ of the MBO but the full density of all other types of objects. The synthetic detections have no astrometric or photometric error: we have assumed, to weight the observations to be fitted, a standard deviation of 0.1 arcsec in each angular coordinate. There were no false detections generated within each simulated Pan-STARRS field. Still, the sky-plane density of objects tops at $122/\mathrm{deg}^2$ in the center of the Jupiter Trojan cloud as shown in Fig. 1.

4. Tracklets

We define a *tracklet* as a set of observations which *could* belong to the same moving object. In the simulation described in the previous Section, most tracklets are composed of only 2 observations (with about a 15 minute separation); tracklets with up to 4 observations can result from the superposition near the edge of two fields. *A posteriori*,

† To maximize the number of discoveries with a given telescope time and performance, exactly the same windows should be visited in three separate nights. Thus the windows should be defined with respect to the opposition at some reference epoch during the lunation.

after the inventory of detections in one frame has been extracted, the only information available to suggest that 2 observations belong to the same object is their proximity on the sky at slightly different times.

At the sky-plane densities anticipated for the final Pan-STARRS system of $\sim 250/\deg^2$, with a comparable density of false detections at the 5 σ level, we will need to join observations using some *a priori* knowledge or other information supplementing the positional one (e.g. trail orientation, trail length, apparent brightness).

Even at the reduced sky-plane densities of this work the task of identifying all possible pairs of detections that are closely spaced on the sky within the 7 deg^2 Pan-STARRS field-of-view is not trivial. Calculating the distance between every possible pair has a computational complexity $O(N^2)$ for N observations, thus it is inefficient. We have implemented an approach using kd-trees (Kubica *et al.* 2005a, 2005b) to search for the close pairs. This technique uses a tree structure (like a binary search generalized into multiple dimensions) to quickly eliminate large groups of detections in the search process: thus the search time for close sets of detections is $O(N \log N)$.

For instances where tracklets of more than 2 observations are possible these are created by fitting observations to a linear function of time within some error, with the distance between the first and last observation limited by a maximum proper motion. Thus if four observations are within the allowable error off a linear fit, but only combinations of three satisfy the maximum proper motion cut, two tracklets will be created with two observations in common. When the tracklet determination is done, tracklets that are entirely included in another tracklet are removed.

In fact, a tracklet composed as discussed above may be *false*, that is it may contain observations belonging to different moving objects†. A tracklet quality control can be performed by fitting degree 2 polynomials of time to the individual observations, separately for right ascension and declination (Milani et al. 2005b, Section 2). If there are 4 (or more) observations, high residuals with respect to the quadratic fits is a good diagnostic for false tracklets. By using RMS > 0.3 arcsec as a control, 47% of the false tracklets with 4 observations have been discarded. For the tracklets with 3 or more observations the second derivatives of right ascension and declination as estimated in the quadratic fits could be used as diagnostics of false tracklets. However, significant second derivatives can be interpreted as an indication that the observed object is very near. To discard a tracklet on the basis of these metrics would introduce the risk of discarding the discoveries of some NEO. For the tracklets with 2 observations there is no way to remove the false ones at this stage. This implies that the task of deleting the 0.6% of the tracklets which are false needs to be included in the identification process; see Section 6.

5. Identification

Given a tracklet, in most cases the information contained allows to compute only a 4-dimensional attributable (that is the tracklet is a TSA). An attributable does not contain enough information to produce a full set of orbital elements: in fact the range r and the range rate $\dot{r}$ of the corresponding moving object are left completely undetermined. In Milani et al. (2004) we have introduced some dynamical and physical constraints to these undetermined variables, confining them to a bounded region of the $(r, \dot{r})$ plane, the *Admissible Region*: each point of this 2-dimensional region corresponds to a Virtual Asteroid (VA), with a full orbit. In the same paper we have also discussed a method for

† It may also contain some observation not belonging to any moving object, but to a variable star etc.; this does not occur in the present simulation.

sampling the Admissible Region by triangulation, to obtain a finite (and not too large) set of VAs that can be propagated in time till the epoch of another attributable.

In Milani et al. (2005b) we have considered the problem of joining two attributables, referring to different times, to produce preliminary orbits to be used as first guess in a differential corrections procedure fitting the whole set of observations generating the two attributables. We call this kind of identification a *linkage*. The VAs defined by the triangulation of the Admissible Region of one attributable are propagated to the epoch of the other one, then the *identification penalty* technique described in Milani et al. (2000) is used to assess whether the second attributable can belong to the same object.

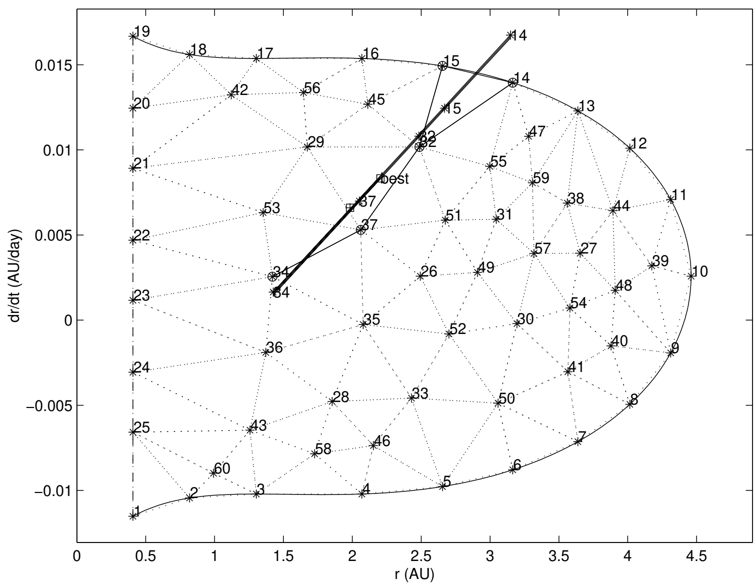

Figure 2. The triangulated Admissible Region for $2003BH_{84}$ corresponding to the attributable from the discovery night and the identification with an attributable of 12 days later. The nodes of the triangulation with moderate identification penalty are joined with solid lines; the bold line shows a part of the LOV. The nominal solution is marked "best" and the true solution by a crossed square sign.

Even if in this way we can obtain some preliminary orbits, the differential corrections algorithm may not converge: to avoid this problem, we use a *constrained differential correction* algorithm, providing solutions along the Line Of Variation (LOV solutions with only 5 free parameters, see Milani *et al.* 2005a). These solutions are meant as a sampling of the part of the LOV (a 1-dimensional set) lying inside the Admissible Region, that form a *second generation* set of VAs. The next step is testing these orbits against the available attributables in a third night of observations to see if they can belong to the same object: this kind of identification is called *attribution*. In most cases this allows to fit a full least squares orbit of good quality (see Section 7). This procedure is illustrated by Figure 2 in which the triangulated Admissible Region is shown for the asteroid $2003BH_{84}$; the VAs of the triangulation that have moderate identification penalty served as starting points for constrained differential corrections, which resulted in convergent LOV solutions. The true solution, computed using additional observations, is close to the LOV.

The procedure above is very effective in obtaining identifications and good orbits for NEO and MB (Milani *et al.* 2005b). In the present simulation we have found that the same algorithms can straightforwardly be applied to Jupiter and Mars Trojans and to

Centaurs, but for distant objects (including TNO, SDO, Trojans of the other outer planets) we found additional problems to be solved. A distant object has a proper motion less than 100 arcsec per day near opposition, even less in the near-quadrature areas. Within $\Delta t = 15$ minutes, the motion is less than 1 arcsec, thus the angular rates are determined with very poor relative accuracy.

Another related problem occurs when two tracklets, belonging to distant objects, from different nights are proposed for identification. The two sets of observations together might still be a TSA: the projection of the orbit on the celestial sphere might be a great circle within the observational errors: because the observations of the same night are very close, the great circle through the nightly average points fits all the observations in a satisfactory way. Then the attempt to compute an orbit might fail even if the identification is true. After this failure, the recursive chain of identifications is interrupted and even if the same object has been observed in a third night, this will not be found.

The solution we adopted is to accept a lower quality orbit when it is found that the joined observations from 2 tracklets are still a TSA. The procedure has three steps. First, we test whether the set of all observations from both nights is a TSA or not: this is done by computing the two components of curvature tangent to the celestial sphere, namely *acceleration* (along track component) and *geodetic curvature* (cross track), together with their uncertainty as deduced from the covariance matrix. If the curvature components are in absolute value less than their standard deviations then the set of observations is still a TSA, and differential corrections starting from a rough preliminary orbit are likely to diverge. In this case the second step is to compute a 4-parameter orbit, with values of range and range-rate fixed at the preliminary orbit values and the 4-dimensional attributable coordinates fitted: essentially, we compute a single attributable for the two nights together. The third step is to use the output of this fit, converted to some other coordinates (e.g., cartesian coordinates) as the first guess for constrained differential correction. This succeeded for $\simeq 82\%$ of the distant objects; however, for the remaining 18% we kept the 4-parameter orbits as a VA to be used in the next recursive step.

These 4-parameter orbits, even when it is not possible to upgrade them to 5-parameter LOV orbits, are used to compute predictions for a third observing night: by using the identification penalty we select the candidate 3-night identifications. When the data from three distinct nights (with an interval of at least 4 days between each couple of nights) are joined together, they have significant curvature, even for distant objects: the angular rates are estimated very roughly, but the average angles for each night are determined to sub-arcsec accuracy, and the three couples of average angles are very unlikely to be found on a great circle with such a precision. Thus, if the proposed identification is true the differential corrections always converge to a LOV orbit, independently from the quality of the 2-night orbit used to provide a first guess. Starting from the LOV orbit a full least squares orbit can be computed in $\simeq 87\%$ of the cases. With this method, the distant objects are not more difficult than the other ones from the point of view of finding 3-night identifications confirmed by a least squares orbit (with either 5 or 6 parameters).

The whole procedure is somewhat more complicated than the description above, because to achieve top completeness in a computationally efficient way the procedure must be organized in a sequence of iterations. The first iteration is based on a "smart triangulation", that is the Admissible Region is triangulated only for those tracklets which are likely to belong either to a NEO or to a distant object (this choice is based on the value of the proper motion and on the overall size and number of disconnected components of the Admissible Region). For most tracklets, likely to belong to MB or Jupiter Trojans, only 2 VAs are computed (one in the MB and one in the Trojan region).

The second iteration operates on all tracklets left unidentified in the first, triangulating the admissible region for all with a metric optimized to find NEO; indeed, all NEO identifications are found. The third iteration solves the few remaining cases, mostly Jupiter Trojans and distant objects: one especially difficult object was the only long period comet observed over 3 nights in this simulation, which required loosening the controls on maximum eccentricity in the differential correction iterations: the eccentricity of the final orbit is 0.988 ± 0.02, but an eccentricity > 1.5 was reached during the iterations.

In conclusion, all the identifications, both for objects observed in 2 nights and for those observed in 3 and more nights, can be found. The problem, as dicussed in the next Section, are the false identifications found along with the true ones.

6. Identification management

The properties of our identification and orbit determination procedure we want to measure are *completeness* and *reliability*. Completeness is measured by the ratio between the true identifications found and the ones hidden in the data (known by means of the "ground truth"). Reliability is measured by the fraction of false identifications among those proposed. The overall completeness depends upon the fraction of simulated objects for which all tracklets have been used in least square orbital fit and the fraction of objects which have been "lost", that is all their tracklets remained unlinked. Moreover, we need to take into account the fraction of true but incomplete identifications, in which not all the tracklets belonging to the same object have been identified. To compute these metrics we need first to join all the identifications with 2, 3 (and possibly more) tracklets.

The final stage of this identification management procedure is the *normalization* of the identification database (Milani et al. 2005b, Section 7.1). The purpose is to remove all duplications and contradictions accumulated in the identification process. We first sort all the identifications found by "quality", that is, an identification is *superior* if either it contains more nights or has the same number of nights and lower RMS of residuals.

The normalization procedure uses the following binary relations among identifications: *compatible* (all the tracklets belonging to the first are among the tracklets of the second), *independent* (none of the tracklets belonging to the first are among the ones of the second) and *discordant* (neither compatible nor independent).

Then we scan this sorted list from the top to reduce it to a *normalized* list of identifications. The first one is kept in the normalized list. Each of the others is compared with all the ones already included and it is kept in the normalized list if it is independent from all the others. It is removed if it is compatible with some of the previous ones. It is also removed if it is discordant with a previously included one with more tracklets.

Discordant identifications appear as contradictions, unless they have been removed by an identification containing all the tracklets of both: e.g., A=B=C and B=C=D are discordant unless A=B=C=D is in the list. Thus the discordant identifications with the same number of nights are both removed from the normalized list at the end of the procedure. This improves the reliability, because it is very likely that the tracklets used in a false identification belong to an object which has been observed in other nights: thus, if the true identifications are found, the false identifications will be discordant with them and will not appear in the normalized list. However, the price to be paid for this is the loss of some true identifications, resulting in lower completeness.

In conclusion, there is a trade-off between reliability and completeness: the normalization procedure defined above ensures top reliability. Moreover, this procedure needs to be applied in batch, working on all the observations of several observing nights, not one by one, not even night by night. After completing the search for all possible identifications, we can normalize the list and then remove the tracklets belonging to normalized identifications, leaving the unidentified ones for another iteration of the procedure.

The special attention we pay to reliability is due to the quadratic growth of the false tracklets/identifications number with the number density of observable objects (per unit area on the sky, see Figure 1). In the area corresponding, in the simulation we are using, to the center of the L5 Trojan cloud, the number density is already such that both false tracklets and false identifications are common. In a full simulation with all the MB objects the number of false tracklets/identifications should grow by an order of magnitude.

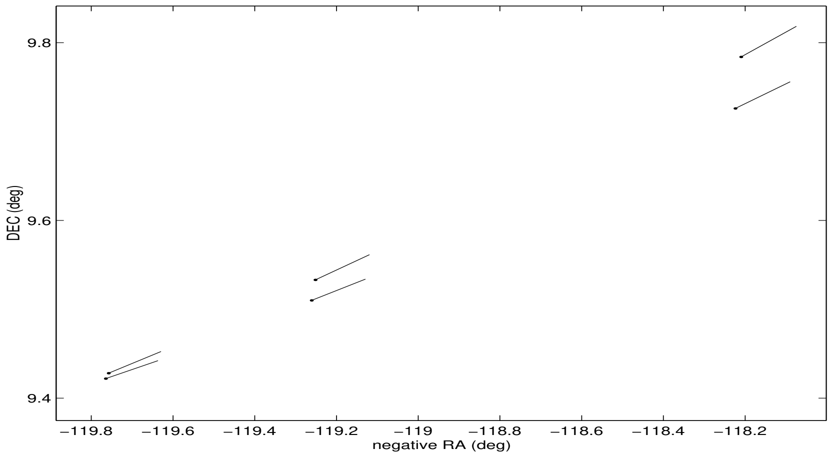

Figure 3. The attributables of 6 tracklets from the peak number density region at the center of the L5 Trojan cloud; the points represent the average positions, the segment the angular motion in one day. The corresponding tracklets can be identified in 4 ways, two of which are false.

A surprising result of this simulation is that false identifications with good least squares fits of all the observations can be found even without simulating astrometric noise. As an example, we have found $\simeq 50$ false identifications with 3 nights of data: the exact number depends upon the value of the controls. All these false identifications, without exceptions, are removed by the normalization procedure, but each one of them can "kill" one or two true identifications which are discordant. In Figure 3 we show one example of "Trojan twins", so close on the celestial sphere and with so similar angular rates that the tracklets may be identified in 4 different ways, all with RMS below 0.06 arcsec.

Thus what matters for false identifications is the number density of detected objects with similar angular rates: e.g., either Jupiter Trojans or distant objects. If either a NEO or a MB happened to be close in position to one of these, it would not be confused.

Table 1. Results for three-nighters

Class	no.tot	complete	incomp	lost	comp%	inc%	lost%
All	74904	74855	0	49	99.9	0.0	0.1
Tro J	55147	55098	0	49	99.9	0.0	0.1
Main B	6326	6326	0	0	100.0	0.0	0.0
Neo	1086	1086	0	0	100.0	0.0	0.0
Com/cent	711	711	0	0	100.0	0.0	0.0
Distant	11634	11634	0	0	100.0	0.0	0.0

The results of the identification procedure applied to this simulation are summarized in Table 1 for 3-nighters and Table 2 for 2-nighters. There were also some objects observed

Table 2. Results for two-nighters

Class	no.tot	complete	lost	comp%	lost%
All	24191	24171	20	99.9	0.1
Tro J	15989	15980	9	99.9	0.1
Main B	2202	2200	2	99.9	0.1
Neo	596	596	0	100.0	0.0
Com/cent	245	245	0	100.0	0.0
Distant	5159	5150	9	99.8	0.2

in more than 3 nights ($1,013$ in 4, 295 in 5 and 60 in 6 nights), and for all of them complete identifications were found in 100% of the cases. Note that of the very few lost objects not even one is due to the difficulty of finding the identification: they are all due to discordance with false identifications.

Of course without astrometric and photometric errors the results should be "perfect", and as shown in the Tables essentially they are, regardless of the orbit class. This means the algorithms are effective, it does not mean the same results can be obtained under more realistic conditions, e.g., with astrometric and photometric noise.

The identification management described in this Section turns out to be also effective as "tracklet management", to remove the false tracklets and join tracklets of the same night belonging to the same object. In fact, we have found not a single case of a false tracklet included in an identification (confirmed by a least squares solution). We found no incomplete identifications (see the third column in Table 1), implying that all the tracklets of the same night belonging to the same object have been included. Thus we can discard all the tracklets discordant (with some but not all observations in common) with the tracklets included in identifications: no true tracklet is wrongly removed.

The result is that 94.6% of the false tracklets are discarded. Moreover, most cases of multiple tracklets in the same night for the same objects are "solved" by merging in an identification together with others. Thus the leftover database of unidentified tracklets contains few false and few couples of true tracklets belonging to the same object in the same night. That is, tracklet management is mostly done. The question is whether the result would be that good in a more realistic simulation (with noise) and with real data.

7. Orbit determination

The definition of identification we use requires that an orbit can be fit to all the observations included in the identified tracklets. However, the quality of the identification is not the same as the quality of the orbit obtained with it. Indeed, one orbit could fit very well all the available observations of a given object, and still be quite different from the "true" orbit (in this context, the orbit used for generating the simulated observations).

Three metrics have to be considered in the quality of the orbit determination resulting from a true identification:

(1) The number of parameters which have been fit. An orbit can be a full least squares solution, with 6 parameters, a constrained LOV solution with 5 fit parameters, a 4-parameter fit with two parameters (range and range rate) kept fixed.

(2) The distance between the "true" orbit and the one determined, measured taking into account the covariance of the latter. If X_0 is the true orbit, X the one determined with normal matrix C_X, we use the 6-dimensional norm of the orbit error

$$D_6 = ||X - X_0||_6 = \sqrt{(X - X_0)^T\ C_X\ (X - X_0)/6}\ .$$

(3) The distance between the attributable predicted for an epoch one month later based on the true orbit, and the one based on the determined orbit, taking into account the expected uncertainty of the prediction. If A_0 is the attributable predicted one month later based on X_0, and C_{obs} is the normal matrix for such an attributable resulting from observations at that later date†, A is the attributable predicted with X and $C_A = \Gamma_A^{-1}$ is its normal matrix obtained by propagating the covariance Γ_X, we use the 4-dimensional norm of the prediction error

$$K_4 = ||A - A_0||_4 = \sqrt{(A - A_0)^T \; C \; (A - A_0)/4} \text{ where } C = C_A - C_A \; (C_A + C_{obs})^{-1} \; C_A \; .$$

The value of these metrics depends upon the purpose, that is how we are going to use these orbits. The norm D_6 is useful in the hypothesis that the same object is rediscovered under similar conditions (especially with the same number of observed nights) after some comparatively long time, years later. Then, if D_6 is small, the orbit identification between the two independent discoveries is easy. The norm K_4 is useful in the hypothesis that the same object is detected one month later, even in a single night. K_4 is the same metric used in the search for attributions of single tracklets: if it is small, the attribution is easy.

The results obtained for the identifications found, subdivided by number of nights of observations, are as follows. For 3-nights identifications, the norm D_6 is more than 3 in 0.4% of the cases with 6-parameter full orbits, in $\simeq$ 10% of the cases with 5-parameter orbits (which are only 2% of the total); there are no 4-parameter orbits. The K_4 norm is less than 1 in all cases. In conclusion, almost all the 3-nights orbits are good enough both for next month attributions and for next apparition orbit identifications; this confirms the results by Spahr et al. (2004). 4-nights orbits are of course even better.

The results for 2-nighters are very different: the norm D_6 is more than 10 in 0.4% and more than 3 in 4% of the 6-parameter orbits. For 5-parameter orbits, 13% of the cases have $D_6 > 10$ and 58% have $D_6 > 3$. For 4-parameter orbits, 0.2% of the cases have $D_6 > 10$, 17% have $D_6 > 3$; however, the normal matrix of such orbits has rank 4, and the norm D_6 may not be a good diagnostic. From this test we can conclude not only that 2-nights identifications provide only inaccurate orbits (in agreement with Spahr et al. 2004), but also that the linear approximation to estimate the uncertainty (by means of the normal and covariance matrix) can fail, especially for 5-parameter LOV orbits. This indicates that orbit identification with rediscoveries in another apparition would be difficult, but not necessarily impossible, given the extreme robustness of the methods to identify 2-nights LOV solutions discussed in Milani *et al.* (2005a).

The K_4 norm for 2-nighters tells a different story. For 5- and 6-parameter orbits the value is < 1 for almost all cases (exceptions are only 0.02%). For the 4-parameter orbits, which are needed only for distant objects (see Section 5), K_4 is larger than 10 in 91.5% of the cases, even > 100 in 1.6% of the cases. This implies that, if a distant object is observed in only two nights in one lunation, and in only 1 night in the next one, it could be difficult to find a 3-nights identification with the algorithms used in this work‡.

Further progress in the algorithms for identification is possible (and we are working on them). Moreover, it is necessary to run a multi-lunation simulation to try to achieve good orbit determination for distant objects. However, the conclusion for now is that for the distant objects (TNO, scattered disk, Trojans of S-U-N) the survey observation planning should guarantee three tracklets in three separate nights in each lunation, otherwise a

† The uncertainty is computed under the same conditions of the present simulation, e.g. assuming two observations with an interval of 15 minutes.

‡ It is possible that other methods would work in such a case, e.g., a procedure using Gauss' preliminary orbit method with some smart trick to avoid the $O(N^3)$ computational complexity.

non negligible fraction of the discoveries for this population could be lost. If this requirement is satisfied, we see no obstacle from the identification and orbit determination to a *unbiased* survey discovering all sub-populations of solar system objects. The problem will have to be reassessed with more realistic simulations, e.g., introducing astrometric and photometric errors and a probabilistic detection model, before this conclusion can be reliably applied to the real data.

ACKNOWLEDGMENTS

This research has been funded by: the Italian *Ministero dell'Università e della Ricerca Scientifica e Tecnologica*, PRIN 2004 project "The Near Earth Objects as an opportunity to understand physical and dynamical properties of all the solar system small bodies", *Ministry of Science and Environmental Protection of Serbia* through project 1238 "Positions and motion of small Solar System bodies", the *Observatorio de Mallorca* (OAM), the Spanish *Ministerio de Ciencia y Tecnología* and the European funds *FEDER* through the grant AYA2001-1784. The design and construction of the Panoramic Survey Telescope and Rapid Response System by the University of Hawaii Institute for Astronomy are funded by the United States Air Force Research Laboratory (AFRL, Albuquerque, NM) through grant number F29601-02-1-0268.

References

Bernstein, G. M., Trilling, D. E., Allen, R. L., Brown, M. E., Holman, M. & Malhotra, R. 2004, *AJ* 128, 1364

Bottke, W. F., Morbidelli, A., Jedicke, R., Petit, J., Levison, H. F., Michel, P., & Metcalfe, T. S. 2002, *Icarus*, 156, 399

Duncan, M. J., Levison, H. F., and Budd, S. M. 1995, *AJ* 110, 3073

Elliot, J.L., Kern, S.D., Clancy, K.B., Gulbis, A.A.S., Millis, R.L., Buie, M.W., Wasserman, L.H., Chiang, E.I., Jordan, A.B., Trilling, D.E. and Meech, K.J. 2005, *AJ* 129, 1117

Gomes, R.S., Morbidelli, A. and Levison, H.F. 2004, *Icarus* 170, 492

Ivezić, Ž., and 32 colleagues 2001, *AJ* 122, 2749

Jedicke, R. and Herron, J. D. 1997, *Icarus* 127, 494

Jedicke, R., Larsen, J., and Spahr, T. 2002, *Asteroids III* 71.

Jewitt, D. C. and Luu, J. X. 2000, *AJ* 120, 1140

Kubica, J., Moore, A., Connolly, A. and Jedicke, R., 2005 to appear in *Signal and Data Processing of Small Targets*

Kubica, J., Moore, A., Connolly, A. and Jedicke, R. 2005, to appear in *The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Levison, H.F. and Morbidelli, A. 2003, *Nature* 426, 419

Milani, A., La Spina, A., Sansaturio, M. E. and Chesley, S. R. 2000, *Icarus*, 144, 39

Milani, A., Gronchi, G.F., de' Michieli Vitturi, M. and Knežević, Z. 2004, *CMDA* 90, 59

Milani, A., Sansaturio, M.E., Tommei, G., Arratia, O. and Chesley, S.R. 2005a, *A & A* 431, 729

Milani, A., Gronchi, G.F., Knežević, Z., Sansaturio, M.E. and Arratia, O. 2005b, *Icarus*, in press

Rabinowitz, D. L., Helin, E., Lawrence, K. and Pravdo, S. 2000, *Nature* 403, 165

Sheppard, S. S., Jewitt, D. C., Trujillo, C. A., Brown, M. J. I., and Ashley, M. C. B. 2000, *AJ* 120, 2687

Spahr, T., Chesley, S., Heasley, J., Jedicke, R. 2004, *AAS, DPS meeting* #36, #32.19

Stokes, G. H. and 11 co-authors. 2003, Report of the Near-Earth Object Science Definition Team. August 22, 2003 (available at `http://neo.jpl.nasa.gov/neo/report.html`).

Tedesco, E. F., Cellino, A., and Zappalá, V. 2005, *AJ* 129, 2869

Wisdom, J. and Holman, M. 1991, *AJ* 102, 1528