INSTITUTE FOR DEFENSE ANALYSES

Statistical Approach to the Operational Testing of Space Fence

Daniel L. Pechkis
Nelson S. Pacheco
Tye W. Botting

July 2015

IDA Document NS D-5541
Log: H 15-000654

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2015		**00-00-2015 to 00-00-2015**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Statistical Approach to the Operational Testing of Space Fence	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Institute for Defense Analyses,4850 Mark Center Drive,Alexandria,VA,22311-1882**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Space Fence will be a terrestrial-based radar designed to perform surveillance on earth-orbiting objects. Its capabilities will increase the number of objects tracked from approximately 17,000 to over 100,000. Testing a system whose complete set of observations cannot be verified in a timely manner by existing systems presents challenges for gathering detection and accuracy truth data while ensuring a reasonable test duration. We propose a rigorous statistical test design with candidate on-orbit test targets that span orbital limits defined by Space Fence operational requirements. We characterize system performance across the entire operational envelope by using relatively small subsets (containing no more than 1530 satellites) of the public Satellite Catalog (SATCAT) grouped by altitude, inclination, and size. We identify the type and number of on-orbit test targets needed for evaluating metric accuracy, probability of track, object correlation, small object sensitivity, and data latency. Our method quantifies the probabilities of meeting requirements, determines how performance varies as a function of an object???s altitude, inclination and/or size, estimates a 25-day test duration, and determines that modeling and simulation methods may be needed to represent 125 additional satellites. These results provide testers and users a mathematical basis of evaluation for Space Fence employment decisions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**12**	

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This publication proposes a new methodology for operationally testing Space Fence. Space Fence will be a terrestrial space-directed surveillance radar system that is expected to increase the number of tracked satellites from approximately 17,000 to more than 100,000. This dramatic improvement will be difficult to test because less capable legacy systems are needed as verification instruments. We propose a test design, based on rigorous statistical principles, that characterizes system performance across the entire operational envelope by using a relatively small set of objects.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-5541

Statistical Approach to the Operational Testing of Space Fence

Daniel L. Pechkis
Nelson S. Pacheco
Tye W. Botting

Statistical Approach to the Operational Testing of Space Fence

Daniel L. Pechkis, Nelson S. Pacheco, Tye W. Botting

Abstract*—Space Fence will be a terrestrial-based radar designed to perform surveillance on earth-orbiting objects. Its capabilities will increase the number of objects tracked from approximately 17,000 to over 100,000. Testing a system whose complete set of observations cannot be verified in a timely manner by existing systems presents challenges for gathering detection and accuracy truth data while ensuring a reasonable test duration. We propose a rigorous statistical test design with candidate on-orbit test targets that span orbital limits defined by Space Fence operational requirements. We characterize system performance across the entire operational envelope by using relatively small subsets (containing no more than 1530 satellites) of the public Satellite Catalog (SATCAT) grouped by altitude, inclination, and size. We identify the type and number of on-orbit test targets needed for evaluating metric accuracy, probability of track, object correlation, small object sensitivity, and data latency. Our method quantifies the probabilities of meeting requirements, determines how performance varies as a function of an object's altitude, inclination and/or size, estimates a 25-day test duration, and determines that modeling and simulation methods may be needed to represent 125 additional satellites. These results provide testers and users a mathematical basis of evaluation for Space Fence employment decisions.

***Index Terms*— Analysis of variance, Operational testing, Least squares methods, Phased array radar, Satellite tracking, SATCAT, Space Fence, Statistical test design, Metric accuracy**

I. Introduction

The United States is acquiring a new terrestrial space-directed radar system to detect, track, and catalog space objects, including the growing population of space debris ("space junk"). The new system will consist of two S-Band (2-4 GHz) phased-array radar sites, located at Kwajalein in the Marshall Islands and one other location (to be determined). The radars will autonomously perform cued and uncued surveillance as well as cued searches for objects in low- and medium-earth orbits (LEO, MEO), and higher. Space Fence will provide object tracking and radar characterization data to the U.S. Air Force Joint Space Operations Center (JSpOC) to support Satellite Catalog (SATCAT) maintenance and other space situational awareness needs [1].

Submitted for review July 2015. This work was supported by the Institute of Defense Analyses (IDA), Alexandria, VA, under Task BD-9-2299(82), sponsored by the OSD Director of Operational Test and Evaluation (DOT&E). D. Pechkis and T. Botting are with the IDA Operational Evaluation Division (OED). N. Pacheco was with IDA/OED but is now an independent consultant.

Space Fence's large surveillance field of view (FOV) and improved small object detection sensitivity are expected to increase the number of orbiting objects routinely tracked from approximately 17,000 to more than 100,000 [1,2].

For the majority of objects tracked by current radars and optical telescopes, we lack position, velocity, and time truth data of sufficient accuracy to test the higher accuracy expected from Space Fence. Testing a space surveillance system whose complete set of observations cannot be verified in a timely manner by existing radars and optical telescopes presents some challenges: (1) How do we know Space Fence is "seeing" all the objects it is intended to "see"? (2) Are the radar measurements on all the objects "seen" by Space Fence of sufficient accuracy and precision to meet its requirements and support orbital prediction and catalog maintenance? And (3) can adequate testing of an operationally representative sample population, covering all intended object sizes, altitudes and inclinations, be performed in a timely manner?

Historically, space surveillance radar performance has been tested by comparing radar observations against truth data for a small number of well-understood objects with accurately known positions (measured within 1 meter), determined either from laser ranging or onboard beacons [3-6]. Although valuable for initial calibration, sole reliance on such objects may not extrapolate to radar performance against an operationally representative population that includes different orbit types, inclinations, altitudes, sizes, shapes, or rotational motions that are expected to be observed by Space Fence.

To address the issue of radar performance across the full operational space, we propose extending initial calibration tests into a broader rigorous statistical test design, using on-orbit test targets that span orbital limits defined by Space Fence operational requirements. Through this new approach, we characterize system performance across the entire operational envelope by using a relatively small subset (containing no more than 1530 satellites) of the nearly 17,000 objects contained in the publically-available SATCAT [7], grouped by altitude, inclination, and size.

We propose using the SATCAT because it is the largest database of verified earth orbiting objects, representing a wide range of orbit types, inclinations, altitudes, sizes, shapes, and rotational motions that Space Fence will track in their real-world population frequencies.

Building on recent experimental design work for assessment of naval surface radar performance [8], we use the target altitude, size, and inclination as predictor variables (also referred to as factors) in statistical tests of radar performance

requirements (e.g., range accuracy) as dependent variables. Our approach quantifies the probabilities of Space Fence meeting its requirements, determines if and how satisfaction of individual requirements depend on an object's orbit and/or size, and estimates the sample sizes needed. Comparing the resulting sample sizes with the number of currently known targets, we determine the areas where augmentation by modeling and simulation (M&S) may be needed. Assuming a Kwajalein-based radar coverage and a conservative number of radar passes per object per day, we also estimate the test duration.

This paper is organized as follows. Section II motivates designing Space Fence testing in terms of the inherent properties of an object and its orbital characteristics. Sections III and IV present test designs to evaluate metric accuracy, and probability of track, and object correlation. Section V proposes a test approach for characterizing Space Fence's detection threshold. Section VI develops a statistical design to evaluate data latency. Section VII summarizes our results.

II. EVALUATING SPACE FENCE IN TERMS OF OPERATIONAL REQUIREMENTS

We evaluate Space Fence in terms of radar performance parameters against the inherent orbital properties of an object (which determines object-to-radar geometry on specific passes) to characterize the system relative to its operational requirements. For this paper, we illustrate our statistical approach by focusing on five common radar requirements, shown in Table I, that illustrate our statistical approach.

Space Fence's performance is ultimately determined by the radar receiver's Signal to Noise (S/N) ratio produced when acquiring a target. While S/N is a measurable quantity, it is not an inherent property of the object. Instead, S/N can fluctuate over the same pass or over different passes based on factors such as the object-to-radar geometry, the object's rotational motion, atmospheric and solar conditions, and others. Therefore, S/N, although a key design measure, is difficult to use as an operational measure of radar performance.

Because of this, Space Fence operational requirements are written terms of performance parameters that can be averaged over many object types and orbital/size conditions, such as probability of track, radar cross section accuracy, and TEARR accuracies.

TABLE I
SELECTED REQUIREMENTS

Requirement	**Specifics**
Metric Accuracy	Time, Elevation, Azimuth, Range, and Range Rate errors
Probability of Track	Percent of fence penetrations that are detected and tracked
Object Correlation	Correct matching of detected object with catalogued object
Small Object Sensitivity	Smallest object detected and tracked
Data Latency	Time from observation generation to receipt at forward user

Our test approach evaluates operational performance parameters, both overall and as a function of satellite orbital and size aggregations identified in the requirements. These aggregations include size, altitude, and inclination. Altitude is important because higher altitudes generally make for longer range, and range appears as an inverse fourth power relationship in the radar equation, thus lowering S/N ratio (although the higher altitude is somewhat offset by longer track windows). Size also affects the S/N ratio, making small targets more difficult to track than larger ones. Inclination might be a less direct influence than altitude and size, but it affects the number and duration of passes across the radar's FOV, indirectly affecting observation accuracy. Additionally, the high object flux density in certain inclination bands might stress the radar's energy management and/or data processing.

III. TESTING METRIC ACCURACY

Space Fence metric accuracy is stated in terms of Time, Elevation, Azimuth, Range, and Range Rate (TEARR) error variance in each of its five components: time, elevation, azimuth, range, and range rate. Each component's error sequence over time can be assumed to be composed of independent, identically distributed pulse-to-pulse errors, so that their distributions are approximately Gaussian, and requirements are specified in terms of 68% one sigma intervals on each component.

A. Approach

To set up our approach to metric accuracy testing, consider the vector TEARR sequence $\mathbf{x}(k,S)$ of radar-centered time, elevation, azimuth, range, and range rate at discrete times k, relative to a fixed location, from the track of a satellite that belongs to a class of satellites S. Classes in S are defined as aggregations of satellites with altitudes, inclinations, or sizes within certain ranges to be specified below. We can express radar observations of $\mathbf{x}(k,S)$ as

$$\mathbf{z}(k,S) = \mathbf{x}(k,S) + \boldsymbol{\varepsilon}(k,S) \quad (1)$$

where $\boldsymbol{\varepsilon}(k,S)$ represents Gaussian metric accuracy errors distributed as $N(\boldsymbol{\upsilon}(k,S),\boldsymbol{\Sigma}(k,S))$, so that $\boldsymbol{\upsilon}(k,S)$ and $\boldsymbol{\Sigma}(k,S)$ represent respectively, the bias and variance/ covariance of TEARR errors.

Space Fence is required to provide estimates of $\mathbf{z}(k,S)$ along with $\boldsymbol{\upsilon}(k,S)$ and $\boldsymbol{\Sigma}(k,S)$ [9]. At initial acquisition, $\boldsymbol{\upsilon}$ and $\boldsymbol{\Sigma}$ will have initial values $\boldsymbol{\upsilon}(0,S)$ and $\boldsymbol{\Sigma}(0,S)$ that, for stable orbits (e.g., non-maneuvering) should converge to steady state values $\boldsymbol{\upsilon}(S)$ and $\boldsymbol{\Sigma}(S)$ as the tracking filter converges. Space Fence will use a systematic error model to determine and correct for bias error [9] so that, under steady state operations, it can be assumed that $\boldsymbol{\upsilon}(S) = \mathbf{0}$. The requirements assume that $\boldsymbol{\Sigma}$ is constant over classes of satellites and, although the radar is required to provide the full $\boldsymbol{\Sigma}$, metric accuracy requirements are only specified on the variance of individual TEARR elements σ_{qq}, q = 1,..,5 (the diagonal elements of $\boldsymbol{\Sigma}$ and not on their covariance σ_{qr} for $q \neq r$).

We propose a two-tier approach in which metric accuracy is

first tested against satellites with accurately known ephemeris (as in the traditional approach), then against the broader set of objects as described earlier. By using the traditional approach we provide for estimation and removal of bias errors, and an estimate of σ_{qq} for addressing metric accuracy in the narrow manner in which requirements are stated.

Following this, we expand our estimate of metric accuracy and test the dependence of $\boldsymbol{\Sigma}$ on the satellite aggregations in S by considering if and how $\boldsymbol{\Sigma}$ may vary over classes of satellites with different orbital and size characteristics.

B. Evaluation of TEARR Requirements against Satellites with Accurate Ephemeris

1) Statistical Approach

We first consider the sample size of satellite tracks needed to estimate metric accuracy for individual TEARR parameters, which we calculate via chi-square hypothesis tests on their error variance, σ_{qq}, using the sample variance reported by the radar as a test statistic. The null hypotheses would be that each TEARR parameter meets the threshold requirement (H_o: $\sigma_{qq} \leq \sigma^*_{qq}$), versus the alternate hypothesis that the parameter exceeds the threshold by a given amount, or effect size, (H_1: $\sigma_{qq} > \sigma^*_{qq} + \delta$). The resulting sample size depends on the effect size δ, the desired statistical power (the probability of correctly rejecting the null hypothesis when it is false), and the significance level, referred to as α error (the probability of incorrectly rejecting a null hypothesis whose thresholds are met) [10]. At a $\delta = 10$ percent effect size and an α error of 5 percent, sample sizes of 300, 400, and 600 yield power levels of 76, 86, and 95 percent, respectively. Because Space Fence will operate in a "target rich" environment for cataloged objects where the number of targets is generally not constraining, we choose the higher level of 95 percent statistical power with 600 tracks.

2) Effect Size Justification

The effect size should be sufficiently large to detect significant improvement above that of existing radars. While the effect sizes could be adjusted for each requirement, we choose a uniform 10 percent effect size, unless otherwise noted. The choice of effect size is driven by comparing Space Fence's required capabilities with other phased array radars. For a theoretical example, if a legacy radar and Space Fence had tracking bandwidths around 5 mega Hertz and radar S/N of 10 and 25 decibels, respectively, their range one-sigma errors would be approximately 7 and 4 meters. A 10 percent effect size would detect Space Fence range one-sigma errors of 4.4 meters, larger than the 4 meters, but less than the legacy 7 meters.

3) Candidate Test Targets

Two candidate accurate ephemeris subsets are the International Laser Ranging Service (ILRS) satellites (for which less than one-meter accuracies are possible) and the High Accuracy Satellite Drag Model (HASDM) satellites [11,12][1]. Assuming that only half of the HASDM/ILRS satellites are available (60 such satellites) with a conservative number of two acceptable passes[2] per day over a Kwajalein-based radar, 600 data points (tracks) could be obtained in as few as 5 test days.

C. Evaluation of TEARR Requirements against satellites that span the operational envelope

We now extend tests of $\boldsymbol{\Sigma}$(S) to explore the effect of satellite grouping factors, S, of altitude, inclination, and size, on metric accuracy. Using the methodology described in section IIIA, and omitting subscripts for simplicity, we entertain the analysis of variance (ANOVA) model [13]

$$\sigma(i,j) = \sigma + S(i) + e(i,j) \qquad (2)$$

where $\sigma(i,j)$ is the observed variance for the j-th track of the i-th satellite grouping for each of the five TEARR parameters, σ is the overall metric accuracy error for that TEARR parameter, S (i) is the effect of the i-th satellite grouping, and e(i,j) represents random error.

Consistent with groupings found in the requirements, for altitude we choose four levels: 250 to 600 kilometers, 600 to 2,000 kilometers, 2,000 to 6,000 kilometers, and 6,000 to 22,000 kilometers. For inclination we choose three levels: high (80-171 degrees, representing near-polar and retrograde orbits), mid (45-80 degrees, centered on the highly populated mid-60 degree inclination band), and low (9-45 degrees). For size we choose two levels: $\geq$ 10 centimeters (approximate tracking limit of current space radars) and <10 centimeters (to capture sensitivity improvements from Space Fence). While results for alternative levels might differ, the approach is valid generally for any aggregation structure.

1) Statistical Approach

We approach the problem as a 4×2×3 full factorial ANOVA design in which the three continuous factors of altitude, size, and inclination are grouped into levels [14]. This approach preserves the operational significance of the altitude, inclination, and size bands stated in the requirements while exploring how the radar's metric accuracy is affected from any of the three factors or their interactions.

In approaching this problem as an ANOVA, it is important to verify that the key statistical assumptions for this model are justified: normality of observations, homoscedasticity, and randomization. The normality assumption was previously justified by radar precision errors being the result of independent, identically distributed, pulse to pulse variation. Homoscedasticity, or equality of variance within groups, is considered sufficiently met if the largest variance within a group is no larger than twice that of the group with the smallest variance within a group. A difference in variance this large would be unlikely, but it could be identified as part of the test on accurate ephemeris objects and, if observed, could be addressed through methods such as logarithmic data transformations. Concerning randomization, although objects to be tracked cannot be randomly assigned to altitude, inclination, or size factors and levels, the total number of

[1] HASDM satellites are tracked multiple times per day to improve orbital predictions in high drag regions. Accurate orbital ephemerides could be generated by fusing the data from their frequent tracking, making them useful as calibration targets for Space Fence metric accuracy.

[2] By an acceptable pass we mean a pass of sufficient elevation and length of track to allow the radar to gather sufficient data to generate observations.

objects needed for testing (under 1,000) is a small fraction (1/20 or less) of the cataloged orbital population. This allows random selection of objects for the radar to track within each factor level combination and provides for a test that "elucidates cause-effect relationships," as referred to by Cochran [15], between the factors and radar performance.

2) Candidate Test Targets

For this ANOVA approach, we have calculated the tracks likely to be available from all cataloged objects passing through Space Fence Kwajalein coverage in the same way as we did with the HASDM/ISLR satellites. For this calculation we assume a conservatively low number of one acceptable pass per day for altitudes less than 600 kilometers, and two acceptable radar passes per day for all targets above 600 kilometers. In the ANOVA design, the 600 tracks needed to test the radar calibration can be evenly divided across all factor-level combinations to ensure that all combinations are tested.

TABLE II[3].
NUMBER OF AVAILABLE SATCAT OBJECTS ORDERED BY INCLINATION, ALTITUDE, AND SIZE.

Incli-nation (deg)	Altitude (km)	SATCAT Objects of Size (cm)		Real Tracks/ Min Test Days		M&S Tracks Needed	
		<10	≥10	<10	≥10	<10	≥10
9 ≤ I ≤ 45	250-600	1	32	25/25	25/1	0	0
	600-2,000	4	101	25/4	25/1	0	0
	2,000-6,000	0	6	0	25/3	25	0
	6,000-22,000	0	2	0	25/7	25	0
45 < I ≤ 80	250-600	16	85	25/2	25/1	0	0
	600-2,000	534	2498	25/1	25/1	0	0
	2,000-6,000	0	10	0	25/2	25	0
	6,000-22,000	1	246	25/13	25/1	0	0
80 < I ≤ 171	250-600	28	276	25/1	25/1	0	0
	600-2,000	1,372	5728	25/1	25/1	0	0
	2,000-6,000	0	89	0	25/1	25	0
	6,000-22,000	0	2	0	25/7	25	0
Total		1,956	9,075	175/25	300/7	125	0

[3] The columns labeled "Real tracks/Min test days" are shaded green or red depending on whether 25 tracks can be obtained within one month, with "25/nn" indicating that 25 tracks can be obtained in a minimum of nn days. The columns labeled "M&S Tracks Needed" represent the number of M&S tracks that would be needed to augment the real track to meet the 25-track limit.

As shown in Table II, there are a total of 24 combinations, so that each combination requires 25 data points. Table II also contains the number of tracks expected to be available from the SATCAT[4] over an approximate one-month test period[5] for each factor-level combination, compared with the 25 tracks needed. Tracks from objects in the SATCAT would be available in all inclination, altitude, and size regimes, except for objects smaller than 10 centimeters at altitudes between 2,000 and 22,000 kilometers.

Of the 600 tracks required, 475 could be obtained within 25 days from real objects, leaving 125 to be met through M&S, should the entire trade space be explored. This is intended to illustrate the approach for identification of M&S needs, not to infer that Space Fence could track objects that small at that range.

If M&S demonstrates that such targets are beyond Space Fence range, the balanced ANOVA we propose with equal sample per factor-level combination would become unbalanced. In this case, the ANOVA could still proceed with unbalanced methods, for example, through merging of some factor-level combinations [16].

By choosing a factorial ANOVA design, we can estimate the statistical power to differentiate between levels of a factor, or in other words, estimate the probability of detecting whether differences in satellite orbits or sizes affect the observation accuracy.

Using the JMP program [17] to calculate power at a 5 percent α error, Table III shows the statistical power achieved in differentiating between levels of the main factors of Inclination (I), Altitude (A), and Size (S) or the interactions of I × A, I × S, and A × S, at various levels of statistical signal to noise ratio[6] (S-SNR), using 24 replicates.

For example, if the difference in mean predicted range accuracy going from level to level is one-quarter of the radar's precision in measuring range accuracy (S-SNR = 0.25), the power to differentiate I, A, S, I × A, I × S, and A × S will be at least 70.4%, 62.5%, 86.4%, 46.0%, 70.4%, and 62.5%, respectively.

Given the multiple factor levels, the statistical power cited is a conservative estimate of the actual power achieved, estimated at the worst possible S-SNR from the various combinations.

TABLE III.
POWER AT SPECIFIC S-SNR, ORGANIZED BY INCLINATION (I), ALTITUDE (A), SIZE (S) AND THEIR INTERACTIONS

Factor	Power at Statistical SNR			
	0.25	0.3	0.375	0.5
Inclination	70.4 %	85.0%	96.3%	99.9%
Altitude	62.5%	78.1%	92.8%	99.5%
Size	86.4%	95.6%	99.6%	99.9%
I × A	46.0%	60.7%	79.7%	96.1%
I × S	70.4%	85.0%	96.3%	99.9%
A × S	62.5%	78.0%	92.8%	99.5%

[4] The publicly available SATCAT, as of June 2013, contains 16,845 objects, of which 15,842 are in Earth orbit and have complete data.

[5] A one-month test period is consistent with historical cost-effective operational test periods that allows for schedule flexibility.

[6] In JMP, the full factorial design SNR is the ratio of the difference in mean predicted response going from level to level to variation in the response variable due to random events.

Similar results can be obtained on other radar requirements specified in terms of continuous variables, such as radar cross-section accuracy. For the sake of brevity, those results are not presented here.

IV. TESTING PROBABILITY OF TRACK AND OBJECT CORRELATION

Unlike metric accuracy, which is expressed in measurement errors, probability of track and object correlation requirements are stated in terms of binomial responses. As such, we propose statistical hypothesis tests on binomial outcomes to assess the system against the letter of the system requirements. We then apply logistic regression methods to determine if system performance varies with an object's altitude and inclination.

A. Probability of Track

Probability of track, p_t, is defined as the probability of keeping track of the position and velocity of a given object that penetrate the radar's surveillance volume. Assuming that the conditions under which objects are tracked remain constant, p_t can also be considered constant over random fence penetrations. Under these assumptions, consider a Bernoulli random variable, *T*, which takes values *T*=1 (object tracked) with probability p_t or *T*=0 (object not tracked) with probability 1- p_t. Assuming independence between tracking attempts, p_t can be estimated by the average $\hat{p}_t(n) = (\Sigma^n_{i=1} T_i / n)$ of a series of n tracking attempts. We develop our approach for $p_t = 0.5$ because the associated probability distribution leads to the largest possible sample size.

We define a null hypothesis H_o: $p_t \geq 0.5$ versus the alternative H_1: $p_t < 0.5 - \delta$ for an effect size δ. Using exact methods as implemented in the R programming language using the binom.power function, Table IV lists sample sizes for 5 percent α, 3 values of statistical power, and 2 effect sizes.

Choosing 95 percent statistical power, for example, we conclude that we need 268 data points (candidate objects to be tracked) to meet the desired power for a 10 percent effect size at the 5 percent α error.

TABLE IV.
POWER AND SAMPLE SIZE FOR PROBABILITY OF TRACK

Probability to Track	**Effect Size**	**Power**	**Alpha**	**Sample Size**
p_t =0.5	10% ($p_t \leq 0.40$)	95%	5%	268
		80%	5%	153
		70%	5%	116
	5% ($p_t \leq 0.45$)	95%	5%	1081
		80%	5%	617
		70%	5%	469

Because *T* is a binary random variable with parameter p_t, the effects of altitude, and inclination on p_t cannot be conducted in the same way as for the continuous TEARR variables [19]. To analyze this binary response design, we use the logit transformation $\ln (p_t / (1- p_t))$. As an illustration, for the Altitude (A) and Inclination (I) factors, the regression model follows the form

$$\ln (p_t / (1- p_t)) = \beta 0 + \beta 1\ A + \beta 2\ I \quad (3)$$

where βi are the regression coefficients. Once the regression coefficients are estimated, the probability of an object being tracked for a given altitude and inclination, and the statistical significance of those factors, can be calculated from the reverse logit transformation as:

$$p_t = \exp (\beta_0 + \beta_1\ A + \beta_2\ I) / (1 + \exp (\beta_0 + \beta_1\ A + \beta_2\ I)). \quad (4)$$

We estimated the sample sized needed to determine whether altitude or inclination are statistically signification factors by running a Monte Carlo simulation of the logistic regression model with 1,000 iterations. The null hypothesis assumed that the factors did not significantly affect p_t, versus the alternative that the factors were significant for various simulated values of p_t. The power of the test, or the probability of identifying a statistically significant effect for altitude or inclination given that there really is an effect, was calculated on a post-hoc basis as the percent of iterations where the significance level was less than α. Unlike for metric accuracy, where the sample size from the hypothesis test was sufficient to examine the effects of altitude and inclination in the full factorial ANOVA design, the logistic regression model requires larger sample sizes than the hypothesis test for the same values of α, power, and effect sizes.

A total of 1,530 data points, or 170 evenly divided over 9 factor/level combinations, are needed to achieve a post-hoc power of 90 percent at the 5 percent α error using a simulated probability of track of 0.5, 0.45, and 0.4 for the high, mid, and low levels of altitude or inclination, respectively.

Table V contains the number of objects expected to be available from the SATCAT over an approximate one- month period, compared with 170 tracks needed per factor-level combination. The 1,530 samples can be collected in 8 days with real tracks assuming one acceptable pass per day for altitudes less than 550 kilometers, and two acceptable radar passes per day for all targets above 550 kilometers.

TABLE V.
THE NUMBER OF AVAILABLE SATCAT OBJECTS TO TEST PROBABILITY OF TRACK, ORDERED BY INCLINATION AND ALTITUDE.[7]

Inclination (degrees)	Altitude (km)	Quantity	Real Tracks/ Min Test Days
9 < I ≤ 45	250-550	22	> 170/8
	550-800	60	> 170/2
	800-3,000	37	> 170/3
45 < I ≤ 80	250-550	67	> 170/3
	550-800	1,094	> 170/1
	800-3,000	1,536	> 170/1
80 < I ≤ 171	250-550	156	> 170/2
	550-800	1,356	> 170/1
	800-3,000	4,039	> 170/1
Total		8,367	>1,530/8

B. Object Correlation

Object correlation is defined as the process of associating detected objects with known SATCAT object to determine if an object being tracked is known or newly discovered. As with probability of track, object correlation can be expressed in terms of a Bernoulli random variable, allowing the use of the same binomial methods previously described. In this case, we replace p_t, with the probability of correlation, p_c, and set it equal to 0.97 to illustrate the dependency of sample size on different probabilities. Following the same procedures as used for probability of track, we generate Table VI. The higher requirement threshold (0.97 for correlation versus 0.5 for probability of track) leads to much smaller sample size needs (81 versus 268) for the same 5 percent effect size, 5 percent α error, and 95 percent power.

We used the logistic regression / Monte Carlo method described in the previous section to estimate the sample size needed to determine whether altitude or inclination are statistically significant factors. A total of 540 data points, or 60 evenly divided over 9 factor/level combinations, are needed to achieve a post-hoc power of 95 percent at the 5 percent α error using a simulated probability of correlation of 0.97, 0.92, and 0.87 for the high, mid, and low levels of altitude or inclination, respectively.

TABLE VI.
POWER AND SAMPLE SIZE FOR OBJECT CORRELATION

Operational Requirement	Effect Size	Power	Alpha	Sample Size
Object Correlation (p_c=0.97)	10% ($p_c \leq 0.87$)	95%	5%	81
		95%	1%	118
	5% ($p_c \leq 0.92$)	95%	5%	228
		95%	1%	335

[7] The column labeled "real tracks/min test days" indicates the number of days needed to obtain 170 tracks.

V. SMALL OBJECT SENSITIVITY

The International Space Station is shielded to withstand the impact of debris as large as 1 centimeter in diameter [20]. Therefore, the protection of manned space flight will depend on Space Fence's ability to track space debris in the 1-10 centimeter size regime, with 10 centimeter being the lower limit for routine tracking by the current space surveillance system. Objects at these small sizes present a particular challenge to detection and tracking because their irregular shape and motion might cause the radar to miss them on any particular pass. However, small spheres present a repeatable, constant signal level to the radar and thus are ideal objects to test sensitivity without the effects of shape or motion

Unfortunately, there are no spherical dedicated objects on-orbit under 10 centimeters and only a few at 10 centimeters [21]. One approach is to track existing small spherical debris that are not readily tracked by the current space surveillance network. A family of possible objects includes sodium-potassium (NaK) ejecta debris from inactive Soviet Radar Ocean Reconnaissance Satellite (RORSAT) satellites. A Haystack debris campaign detected many objects in that size regime that are believed to be spherical, based on Mott Polarization analyses [22].

Because the mission of that debris campaign was to establish debris flux rather than to catalog debris objects, we suspect that there are numerous small spherical objects on orbit, even though that campaign did not provide their orbital elements. There are two potential approaches to this problem: (1) another small debris campaign could be initiated to identify and catalog 1- to 10- centimeter class NaK debris targets or (2) Space Fence could participate in its own debris campaign by initially detecting candidate objects and handing them off to specialized sensors, such as the Haystack radars, for verification.

Table VII shows the number of 1- to 3-centimeter NaK debris objects detected in the Haystack debris campaign, organized by inclination, in the 250 to 600 kilometers altitude band.

A Monte Carlo logistic regression test of the effect of inclination on tracking these potentially spherical objects indicates that such a test would have 88 percent post-hoc power of detecting a 10 percent effect at a 5 percent α error with 175 targets. Should Space Fence attempt to detect and track these objects, there would be a sufficient number of targets above 45 degrees inclination, but M&S would be necessary for targets at inclinations between 9 and 45 degrees, which were not surveyed in the debris campaign. Alternatively, the test could just comprise debris targets in the two inclination bands of 45 to 80 and 80 to 171 degrees. Those targets could be tracked over 13 days.

TABLE VII.
NAK DEBRIS DETECTED BY HAYSTACK

Inclination (degrees)	NaK 1-3 centimeter quantity	Tracks/ min test days
45-80	14	>175/13
80-171	26	>175/7
Total	40	350/13

VI. TESTING DATA LATENCY

Data latency is defined as the time from the end of sensor collects to the moment the data is received by the user, in this case the JSpOC. For Space Fence, we assume data latency is to be no more than 2 minutes, 99 percent of the time.

Historically, for other fence-style radars, radar message latency times cluster around a time well away from zero and close to the requirement, suggesting modeling latency time, L, as a normal random variable with mean μ and standard deviation σ rather than the exponential or lognormal that may be more common in other applications. The upper 99 percentile for L satisfies $P\{[(L - \mu) / \sigma] \leq 2.33\} = .99$, and a statistical test of the requirement is equivalent to testing the null hypothesis that $\mu + 2.33\ \sigma \leq 2$.

Since μ and σ are not known, we use an Upper Tolerance Interval (UTI) random variable T_u, defined as $T_u = m + k_u * s$, where k_u is a function of the sample size, and m and s are estimates of the unknown μ and σ that incorporate their uncertainty. There are various methods in the literature to determine k_u for specified percentiles, sample size, and α [23].

Figure 1 shows the value of k_u for 99 percent/alpha UTIs with alpha at 0.01, 0.05, and 0.1 as a function of sample size for data from a standard normal distribution based on Monte Carlo simulations of various sample sizes. This data was generated using the R function normtol.int [24], and the Howe method [25]. For alpha = 0.05, k_u is converging to the percentile value of 2.33 of known μ and σ from above, because m and s are converging to μ and σ as sample size increases.

Sample sizes above 1,000 are beyond the "knee" of the curve, and the difference between the UTI and the upper percentile point will be relatively minor. This large sample behavior is of particular benefit to Space Fence and most radars because message latency is not typically sensitive to target characteristics. As such, sample sizes of thousands of messages per day should be routine, from which reliable UTI estimates could be made.

Power analyses for tolerance intervals can be challenging because of the complexity of alternative hypotheses and distribution assumptions. The problem is simplified under the assumption that L is normally distributed with an unknown μ and a known σ, say σ_o, obtainable either from history or from estimates from precursor tests.

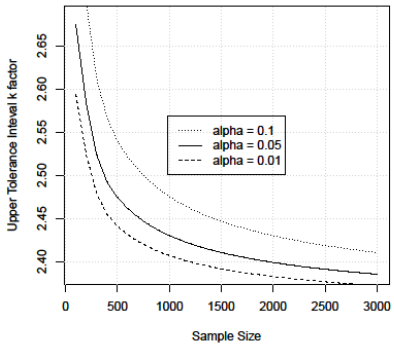

Fig. 1. Upper 99%/95% Normal Tolerance Interval versus Sample Size.

Under this approach, the power of the test can be stated in terms of an alternative hypothesis on the mean, H_1: $\mu < \mu_o (1+ d)$, where $d > 0$ is the percent effect size in terms of the known σ_o. Using the R pwr.norm.test function, for a 10% effect size and a 5% α, power levels of 70, 80, and 90% can be achieved with sample sizes of 471, 618, and 856, respectively.

A 10 percent effect size is selected because it corresponds to a 12 second delay in the 2 minute latency threshold. This delay might be significant for certain conjunction alerts and consequent collision avoidance maneuvers. For the International Space Station, with a 0.5 to 1 m/s collision avoidance maneuver velocity [26], a 12 second delay translates to being 6 to 12 meters closer to a potential conjunction.

VII. SUMMARY

We presented a statistical methodology, using both well-understood satellites and a larger subset of SATAT objects as target sources to ascertain the probabilities of meeting the Space Fence requirements written in terms of measurement errors, binomial responses, and percentiles. Our approach quantified the chances of determining whether Space Fence meets its intended metric accuracy, tracking, correlation, and data latency performance and if the individual performances (except for data latency) depends on an object's altitude, inclination, and/or size. Based on these results, we identified the type and minimum number of on-orbit test targets, determined that 125 orbital tracks may require M&S within specific orbital regimes, and estimated the test duration of approximately 25 days to evaluate the effectiveness of the system. Additionally, we suggest an approach to test Space Fence against small objects not readily tracked by the current space surveillance network. These results provide testers and users a mathematical basis for evaluation and acceptance decisions based on timely prediction of its operational effectiveness against the complete set of observations.

ACKNOWLEDGMENT

The authors would like to thank Dr. Kelly McGinnity, Dr. Kassandra Fronczyk, and Mr. Michael Tuley for useful discussions.

REFERENCES

[1] Air Force Space Command (2014, June) "Space Fence Contract Award," [online]. Available: http://www.afspc.af.mil/news/story.asp?id=123413302)

[2] NASA Orbital Debris Program Office (2012, March) "Orbital Debris Frequently Asked Questions", [online]. Available: http://orbitaldebris.jsc.nasa.gov/faqs.html.

[3] J. Mochan and R.A. Stophel, "Dynamic Calibration of Space Object Tracking Systems," Space Congress Proc, 5th Session, 1968."

[4] C. Noll, and M. Pearlman, "International Laser Ranging Services 2009-2010 Report", National Aeronautics and Space Administration TP 2013-217507, 2012.

[5] "Using Satellites for Radar Performance Monitoring and Calibration", Joint Range Instrument Accuracy Improvement Group Range Commanders Council, White Sands Missile Range, May 03, 1995

[6] L. Martin, N. Fisher, W. Jones, J. Furumo, J. Ah Heong Jr., et al., "Ho'oponopono: A Radar Calibration CubeSat" Proceedings of the AIAA/USU Conference on Small Satellites, Mission Lessons, SSC11-VI-7,

August 2011[online]. Availible: http://digitalcommons.usu.edu/smallsat/2011/all2011/42/
[7] Analyses is based on the June 2013 publicly available SATCAT found on space-track.org
[8] L. A Cortes and D. Bergstrom, "Using design of experiments for the assessment of surface radar detection performance and compliance with critical technical parameters." 80th MORS Symposium, June 2012.
[9] Space Fence Technical Requirements Document, Space Command and Control and Surveillance Division, Air Force Life Cycle Management Center.
[10] J. S. Milton and J.C Arnold, "Introduction to Probability and Statistics: Principles and Application for Engineering and the Computing Sciences" Irwin/McGraw-Hill 3rd ed. (1986).
[11] C. Noll, and M. Pearlman, International Laser Ranging Services 2009-2010 Report, National Aeronautics and Space Administration TP 2013-217507, June 2012."
[12] M.F. Storz, B.R. Bowman, J.I. Branson, S.J. Casali, and W.K. Tobiska, Adv. Space Res. 36, 2497 (2005).
[13] http://www.itl.nist.gov/div898/handbook/ppc/section2/ ppc231.htm
[14] D. R Cox, Principles of statistical inference. Cambridge New York: Cambridge University Press. ISBN 978-0-521-68567-2" (2006).
[15] W. G. Cochran, "The planning of observational studies of human populations (with Discussion)," Journal of the Royal Statistical Society. Series A 128, 134–155, 1965
[16] ANOVA FOR UNBALANCED DATA: AN OVERVIEW, Shaw and Mitchell-Olds, Ecological Society of America1993. pp. 1638-1645"
[17] JMP® 11.0.0, 64-Bit Edition; Copyright 2013 SAS Institute Inc."
[18] R Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statitical Commuting, Vienna, Austria, R version 3.1.0 (2014, April). Platform: x86_64-w64-mingw32/x64 (64-bit).
[19] F. Ortiz, "Dealing with Categorical Data Types in a Designed Experiment Part II: Sizing a Designed Experiment When Using a Binary Response," Scientific Test and Analysis Techniques, Test and Evaluation Center of Excellence, 2014.
[20] "Limiting Future Collision Risk to Spacecraft: An Assessment of NASA's Meteoroid and Orbital Debris Programs" The national Academies, (2011)."
[21] Center for Atmospheric & Space Sciences (2015, June) "POPACS Spheres in Space: Sensors for the Impact of Solar Storms on Our Atmosphere", Utah State University [online] Available: http://cass.usu.edu/files/uploads/POPACS-WebPage.pdf
[22] C. L. Stokely, J. L. Foster, Jr., E.G. Stansbery, J. R. Benbrook, and Q. Juarez "Haystack and HAX Radar Measurements of the Orbital Debris Environment; 2003," NASA JSC-62815 (2006)."
[23] NIST/SEMATECH (2012, April, "e-Handbook of Statistical Methods", [Online]. Available: www.itl.nist.gov/div898/handbook/prc/section2/ prc263.htm
[24] D. S. Young, An {R} Package for Estimating Tolerance Intervals, J. STAT SOFTW (2010) [Online]. Available: http://cran.r-project.org/web/packages/tolerance/.
[25] W. G. Howe, "Two-Sided Tolerance Limits for Normal Populations, Some Improvements," Journal of the American Statistical Association Volume 64, pages 610-620, June 1969.
[26] L. Hutchinson, "How NASA steers the Internaional Space Station around space junk," Arstechnica, (2013,July) [Online]. Available: http://arstechnica.com/science/2013/07/how-nasa-steers-the-international-space-station-around-space-junk/, July 2013.

Daniel L. Pechkis received the B.S. degree majoring in physics, minoring in mathematics and computer science, from Southern Connecticut State University in New Haven, Connecticut in 2002, the M.S. in physics from the College of William and Mary in Williamsburg, VA, in 2003, and the Ph.D. in Physics from the College of William and Mary in 2011.

He was a research assistant with the Material Characterization Laboratory at Southern Connecticut State University, from 2000 to 2002, performing surface and cross-sectional experimental studies of thin film ferroelectric materials and nanoporous silica using atomic force, scanning tunneling, and transmission electron microscopy. From 2002 to 2003 to 2005, he worked at the Nuclear Magnetic Resonance Laboratory (NMR) at the College of William and Mary, performing solid-state NMR spectroscopy experiments on high dielectric perovskite polycrystalline materials and cyclic dipeptides. Between 2005 and 2011, he worked for the Computational Material Science group at the College of William and Mary, performing first-principles quantum mechanical calculations of NMR chemical shielding tensors in piezoelectric/ferroelectric perovskite crystals and alloys using quantum chemistry methods and embedded cluster techniques. This work resulted in two publications in the Journal of Chemical Physics. He is currently a Research Staff Member with the Institute for Defense Analyses, Alexandria, VA, supporting the Pentagon's Director of Operational Test and Evaluation in his oversight of space surveillance systems.

Dr. Pechkis' awards and honors include the Virginia Space Grant Consortium Aerospace Graduate Research Fellowship (2006-2009) and the Microscopy Society of America Undergraduate research fellowship (2001).

Nelson S. Pacheco received the B.S. degree in Mathematics from St. Mary's University in San Antonio, Texas in 1966, the M.S. in Mathematics from the University of Colorado in 1971, and the Ph.D. in Mathematical Statistics from Colorado State University in 1979.

He was a research assistant with the Structural Research Division of Southwest Research Institute in San Antonio, Texas, from 1964 to 1966, where he worked on stress analysis for the Apollo moon rocket transporter. From 1967 to 1973 he served in U.S. Air Force as a Minuteman Targeting Officer, an Orbital Analyst, and a radar signature analyst at the FPS-85 phased array radar and the FPS-79 dish radar. While at the FPS-79 he performed radar analysis on numerous space events, including the Skylab anomaly. Between 1974 and 1985 he taught mathematics at the Air Force Academy, and served as Acting Department Head in 1986 and 1987. He also collaborated with the Center for Energy and Environmental Research at the University of Puerto Rico in 1981 and worked as a principal scientist at the SHAPE Technical Centre, The Hague, Netherlands, in 1982 and 1983. Between 1988 and 1989 he was Analysis Group Leader for MITRE at the SDIO National Test Bed in Colorado Springs, CO, and between 1990 and 2008 he was space surveillance and C3 systems test analyst and group leader at the Operational Evaluation Division of the Institute for Defense Analyses in Alexandria, Virginia. Since 2009 he has worked as an independent aerospace consultant.

Dr. Pacheco's awards and honors include the James L. Madison memorial award for outstanding PhD Statistics student at Colorado State University, and the Harold Brown Research Award nominee at the Air Force Academy for statistical prediction of alternative energy potential across all Department of Defense installations.

He has several publications in the open statistics and engineering professional literature, and multiple internal military publications.

Tye W. Botitng received the B.S. degree in chemistry from Texas A&M University in College Station, Texas, in 1987 and the Ph.D. degree in physical/nuclear chemistry from Texas A&M University in 1999.

From 1988-1989, he was a service engineer for the environmental chemical analytical equipment company, OI Corporation in College Station, Texas. In 1990, he went to graduate school, elucidating fission dynamics and timescales directly by comparing results of novel neutron and gamma-ray emission "clocks." In 1999, he began post-doctoral work with the Department of Nuclear Engineering at Texas A&M University, running a tandem Van de Graaf accelerator and participating in microdosimetry experiments. In 2003, he went on to become Joint Faculty Researcher in conservation science with the National Center for Preservation Training and Technology (part of the United States Department of the Interior) while also teaching chemistry at Northwestern State University, both in Natchitoches, Louisiana. Since 2007, he has been a Research Staff Member at the non-profit Institute for Defense Analyses in Alexandria, Virginia, where he continues to support government oversight of operational test and evaluation for Department of Defense space surveillance acquisitions.

Dr. Botting's awards and honors include being named two-time recipient of the Oliver Torry Fuller Award for Technical Excellence and Innovation, 2007 and 2008; Robert A. Welch Fellow from 1991-1999; George C. Bauer Memorial Scholar, 1987, and Lechner Undergraduate Fellow from 1985-1987. Dr. Botting has several nuclear chemistry and conservation science articles in peer-reviewed publications, although his current work is produced primarily for his Department of Defense sponsor, the Director of Operational Test and Evaluation.