CBP/TRS-316-17

Recommendation Report for the Establishment of Uniform Evaluation Standards for Application of Remote Sensing to Identify and Inventory Agricultural Conservation Practices for the Chesapeake Bay Program Partnership's Watershed Model

Chesapeake Bay Program
Science. Restoration. Partnership.

DRAFT FINAL

January 24, 2017

Prepared for

Chesapeake Bay Program
410 Severn Avenue
Annapolis, MD 21403

Prepared by

Chesapeake Bay Agricultural Remote Sensing Technical Team:

Dr. Lee Norfleet, CEAP Team Leader – USDA-NRCS
Dan Mullarkey, Resource Assessment Division Director – USDA-NRCS
Dan Good, Resource Inventory Division Director – USDA-NRCS
Patrick Flanagan, Resource Inventory Division National Statistician – USDA-NRCS
Kelly Shenk, Agricultural Advisor – US EPA Region 3
Ted Tesler, Chesapeake Bay Technical Coordinator - Pennsylvania Department of Environmental Protection

With:
Mark Dubin
Agricultural Technical Coordinator
Chesapeake Bay Program Office
University of Maryland Extension

Steven Dressing, Tetra Tech
Jon Harcum, Tetra Tech

Contract Support Provided by

TETRA TECH

EPA Contract No. EP-C-12-055
Task Order No. 003

Cover Photo Credit
Tim Sexton, Virginia Department of Conservation and Recreation

Introduction

An assessment of remote sensing accuracy was performed to determine the suitability of remote sensing to identify agricultural conservation practices for credit in the Chesapeake Bay Program (CBP) partnership's watershed model. The assessment combined the findings from a literature search and a detailed evaluation of the strengths and weaknesses of selected metrics.

Literature Review

A limited literature search was performed to locate publications that recommended levels of remote sensing accuracy and the extent of ground-truthing needed in the identification of agricultural conservation practices, otherwise referred to as best management practices (BMPs). Multiple documents/reports provide methods for performing and quantifying analyses of remote sensing accuracy (e.g., Congalton and Green 2008). Other reports evaluated the efficacy of different methods for identifying selected features. For example, Waz (2016) developed/evaluated an automated framework to identify lost and restorable wetlands. Waz performed ground-truthing at 200 sites. Tetra Tech (2014) developed an approach for identifying potential sites for drainage water management practices and included sample sizes necessary to quantify remote sensing error. However, none of these sources provided acceptability guidelines.

One publication did, as part of its objective, evaluate the effect of different levels of ground-truthing to develop a reliable conservation tillage model (Sullivan et al. 2008). They (Sullivan et al. 2008) reviewed similar studies with 15 sites (van Deventeer et al. 1997) and 84 sites (Gowda et al. 2001). The researchers noted that follow-up studies found that the models developed by van Deventeer et al. (1997) performed poorly when applied to a broader array of fields and soils. Sullivan et al. (2008) evaluated the usefulness of Landsat TM data to calculate four indices: Normalized Difference Vegetation Index (NDVI), Crop Residue Cover Index (CRCI), Normalized Difference Tillage Index (NDTI), and the Simple Tillage Index (STI). Ground truth data consisted of a windshield survey, assigning each site a tillage regime (conventional or conservation tillage) at 138 locations throughout the watershed and surrounding areas.

Subsets of the ground-truth sites (n = 20 or n = 44) were used to develop regression models. Statistically significant models were developed for NDTI and STI using the n=20 and n=44 subsets of data. Statistically significant models were also developed for NDVI and CRC using the n=44 subset. When the NDTI models were applied to the Little River experimental watershed image, overall accuracy ranged from 71% to 78% for models developed using 20 and 44 survey points, respectively. With the model fit using 44 sites, errors of commission (i.e., erroneous inclusion) represented 15% of all surveyed locations, compared to 20% for the model fit with 20 ground control points.

Metrics for Remote Sensing Accuracy and Completeness

For the purposes of this assessment, remote sensing was assumed to include both remote sensing and field verification of a percentage of sites or BMPs included in the remote sensing activity. The accuracy of the remote sensing approach is determined by comparing remote sensing conclusions with conclusions based on field or ground examination. Possible outcomes for remote sensing are summarized in Figure 1 and Table 1. Note that "d" may be unknown.

BMP Observed Remotely and Confirmed on Ground (a)	BMP Observed Remotely but not Confirmed on Ground (b)	BMP Not Observed Remotely but Found on Ground (c)	BMP Not Observed Remotely and Not Found on Ground (d)

Figure 1. Possible Outcomes for Remote Sensing.

Metrics considered to assess method accuracy are the False Alarm Ration (FAR), Hit Rate (HR), Critical Success Index (CSI), Post Agreement Rate (PAG), and Proportion Correct (PC) (Schaefer 1990). Formulas for each metric are provided in Table 1.

The ***False Alarm Ratio (FAR)*** is the fraction of remotely-detected BMPs that were not confirmed via farm visits. FAR is also known as the commission error or error of inclusion. The ***Post Agreement Rate (PAG)*** is equal to 1-FAR and is also known as the user's accuracy.

The ***Hit Rate (HR)*** is the fraction of remotely-detected BMPs that were confirmed or found through farm visits. HR ranges from 0 to 1, with a value of 1 indicating all BMPs were found. HR is also known as producer's accuracy. The omission error or error of exclusion is equal to 1-HR.

The ***Critical Success Index (CSI)*** is a measure of the accuracy of the remote sensing method as the percentage of observations that are confirmed on the ground, i.e., BMP presence was correctly determined. The range for the CSI is 0 to 1, with a value of 1 indicating perfect remote sensing. The CSI is a frequently used measure because, unlike the FAR, it takes into account both false positives and missed events, and is therefore a more balanced score.

The ***Proportion Correct (PC)*** is a measure of the accuracy of the survey method calculated as the percentage of survey responses that are confirmed via farm visit, i.e., BMP presence or absence was correctly determined. The range for the PC is 0 to 1, with a value of 1 indicating a perfect survey. The PC is a frequently used measure because, unlike the FAR, it takes into account both false positives and missed events, and is therefore a more balanced score.

The last metric identified in Table 1 is the ***Frequency Bias (FB)*** which has a range from 0 to infinity. An FB of 1 indicates an unbiased result for which the event is forecast exactly as often as it is observed. An FB greater than 1 indicates the BMPs are over-identified while FBs less than 1 indicate under-identification.

The metrics chosen for this assessment are FAR, HR, and FB.

Table 1. Data Elements Used in Measures of Remote Sensing Accuracy and Completeness

		Field Observed		
		Yes	No	Row Total
Remote Sensing	Yes	a	b	a+b
	No	c	d	c+d
	Column Total	a+c	b+d	a+b+c+d=n

Metric	**Formula**
Critical Success Index (CSI)	a/(a+b+c)
False Alarm Rate (FAR)	b/(a+b)
Hit Rate (HR)	a/(a+c)
Post Agreement Rate (PAG)	a/(a+b)
Proportion Correct (PC)	(a+d)/(a+b+c+d)
Frequency Bias (FB)	(a+b)/(a+c)

Confidence Interval for Proportions

The confidence interval for any of the proportions shown in Table 1 i.e., CSI, FAR, HR, PAG, and PC) can be represented by a binomial distribution and confidence intervals estimated with the following equation[1]:

$$p \pm Z_{1-\propto/2}\sqrt{\left(\frac{p(1-p)}{n}\right)\cdot\left(\frac{N-n}{N-1}\right)} \quad (1)$$

where

p = proportion from a particular metric such as CSI, FAR, HR, PAG, or PC
n = number of samples (denominator of corresponding metric)
N = total number of population units in sample population
$Z_{1-\alpha/2}$ = value corresponding to cumulative area of 1-α/2 using the normal distribution (e.g., 1.645 for 90% confidence level, 1.96 for 95% confidence level)

Correcting for Bias

Dividing the total number of remote sensing detected BMPs results by FB could be used to scale the remote sensing results to correct for bias. Algebraically, dividing by FB can be shown to be equal to multiplying by the ratio of PAG/HR.

[1] https://onlinecourses.science.psu.edu/stat414/node/264. The second term under the square root operator accounts for finite populations (N).

$$\frac{PAG}{HR} = \frac{a/a+b}{a/a+c} = \frac{a+c}{a+b} = \frac{1}{FB} \quad (2)$$

The confidence interval for the ratio of PAG/HR can be computed with the following formula[2]:

$$\frac{p_1}{p_2} e^{\pm Z_{1-\alpha/2}\sqrt{\frac{(1-p_1)}{p_1 n_1}+\frac{(1-p_2)}{p_2 n_2}}} \quad (3)$$

where

p_1 and p_2 are PAG and HR, respectively
n_1 = a+b
n_2 = a+c

The National Institute of Standards and Technology (NIST) uses Bayes estimates of proportions rather than PAG and HR directly to accommodate situations where both proportions are equal to one (NIST 2010). In these cases, alternative estimates of p_1 and p_2 can be made.

$$p_1 = \frac{a+0.5}{a+b+1} \quad \text{and} \quad p_2 = \frac{a+0.5}{a+c+1} \quad (4)$$

Results

A series of calculations was performed to determine half-width confidence intervals (α=0.10) for both FAR (Figure 2) and HR (Figure 3). Calculations were performed assuming a range of ρ values from 0 to 1 and sample sizes from 5 to 40. The population size is assumed to be very large in both cases. The black diagonal line in both figures shows the mean sample value of ρ.

The lines for *n*=5 through *n*=40 in Figure 3 show the upper confidence limit (UCL) for FAR. This represents the worst case or upper limit of the 90 percent confidence interval (UCL90). For example, where the mean value for FAR is 0.20 (ρ =0.20 on the x-axis), the UCL90 for *n*=5 is 0.50. For *n*=10, 20, and 30, the UCL90s are 0.41, 0.35, and 0.32, respectively. In other words, where field verification yields a FAR of 0.20 (i.e., 20 percent false alarms), the UCL value (α=0.10) is 0.50 (50 percent false alarms) for a sample size of 5 and 0.32 (32 percent false alarms) for a sample size of 30.

The lines for *n*=5 through *n*=40 in Figure 4 show the lower confidence limit (LCL) value for HR. This represents the worst case or lower limit of the 90 percent confidence interval (LCL90). For example, where the mean value for HR is 0.80 (ρ =0.80), the LCL90 for *n*=5 is 0.50. For *n*=10, 20, and 30, the LCL90s are 0.60, 0.65, and 0.67, respectively. In other words, where field verification yields a HR of 0.80 (i.e., 80 percent of BMPs detected), the LCL90 is 0.50 (50 percent of BMPs found) for a sample size of 5 and 0.67 (67 percent of BMPs found) for a sample size of 30.

[2] http://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/binoraci.htm

It is important to note that the UCL90 for FAR and the LCL90 for HR are determined both by the mean value (ρ) and the sample size (*n*). By setting thresholds on the FAR UCL90 and HR LCL90, the field practitioner has multiple ways to achieve the threshold. For example, it would be possible to satisfy a FAR UCL90 of 0.3 with ρ=0.10 and *n*=10 (UCL90=0.256) or with ρ=0.175 and *n*=40 (upper limit=0.274).

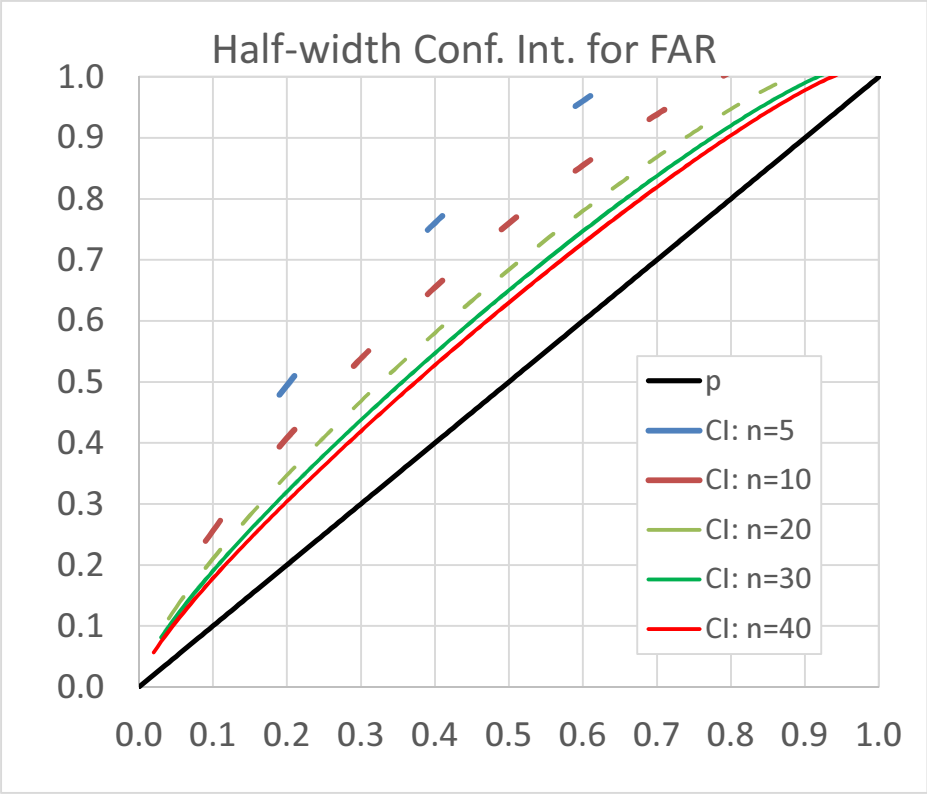

Figure 2. Half-Width Confidence Intervals for FAR with Large N.

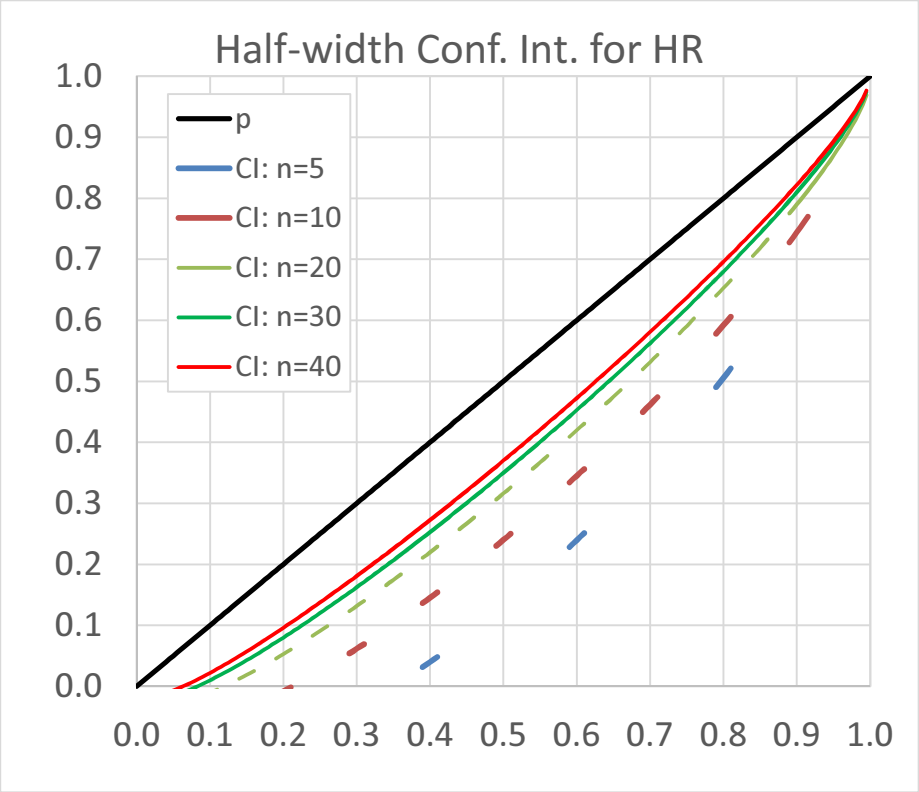

Figure 3. Half-Width Confidence Intervals for HR with Large N.

Example Bias Correction

As stated earlier, it is possible to examine FB to determine the extent of under- or over-reporting. Data in Table 2 illustrate an example of this process using strip cropping data from the USDA-NRCS Remote Sensing Pilot Study. In this case, the FAR UCL90 and HR LCL90 are 0.22 and 0.68, respectively. The FB is 0.88, indicating under-reporting. This value can be scaled by multiplying by the ratio of PAG/HR (or dividing by FB). Using Equation 3 the LCL90 and UCL90 on the PAG/HR ratio are 1.02 and 1.26, respectively. Multiplying these values by the number of remotely sensed BMPs (1,371) yields an estimate that the number of strip cropping practices falls between 1,405 and 1,727 at the 90% confidence level.

Table 2. Application of Bias Correction Using USDA-NRCS Remote Sensing Pilot Study Data for Strip Cropping

	(a)	**(b)**	**(c)**	**(n)**	**N**
Practice Name	BMP Observed Remotely and Confirmed on Ground	BMP Observed Remotely but Not Confirmed on Ground	BMP Not Observed Remotely but Found on Ground	Total Field Verification Sample Size	Population Size
Strip cropping	110	22	40	172	1371

	FAR	**HR**	**CSI**	**PC**	**PAG**
	False Alarm Rate	Hit Rate	Critical Success Index	Proportion Correct	Post Agreement Rate (1-FAR)
Estimate	0.17	0.73	0.64	0.64	0.83
UCL90 (Equation 1)	0.22				
LCL90 (Equation 1)		0.68	0.58	0.58	0.78

	FB			
	Frequency Bias	PAG / HR	Overall Remotely Sensed BMPs	Bias Corrected Estimate of BMPs
Estimate	0.88	1.14	1,371	1,558
LCL90		1.02		1,405
UCL90		1.26		1,727

Discussion

Conclusions reached by Sullivan et al. (2008) indicate that a remote sensing method with an overall accuracy of 71% to 78% is acceptable. Results also indicated that a ground verification sample size of 20-44 is suitable with overall accuracy improving by 7% in this case when *n* was increased from 20 to 44. Similarly, the error of commission, or FAR, value ranged from 0.15 for $n = 44$ to 0.20 for $n = 20$. It is acknowledged that Sullivan et al.'s objectives for using remote sensing are somewhat different from the present purpose, and these differences should be considered if a minimum sample size for ground verification is selected.

Half-width confidence intervals for FAR and HR can be used in combination with mean values calculated from ground verification of remote sensing determinations to assess whether the method was approximately as successful as that employed by Sullivan et al. (2008). Specific threshold values for FAR, HR, or both could be used to determine whether the method is acceptable for a specific BMP.

If a threshold is based on an upper confidence level value for FAR or a lower confidence level value for HR, sample size need not be specified. However, based on the work of Sullivan et al. (2008), a minimum sample size of 20 or more could be set.

This analysis used a two-sided 90 percent confidence level. Alternative levels can be used, but $\alpha=0.10$ is fairly common for environmental work and a broad spectrum of other disciplines.

For cases where the remote sensing method is deemed suitable, the BMP count estimate (and resulting BMP extent estimate) can be scaled to correct for bias using the ratio of PAG/HR. A conservative result could be obtained by applying the lower confidence level value ($\alpha=0.10$) to the measured mean value. It is important to note that the analyses presented here assume that each BMP occurrence is equal. In other words, each occurrence of a fence is of equal length, or each occurrence of conservation tillage is of equal acreage.

Recommendations

Multiple options exist for applying the findings here to decisions regarding suitability of remote sensing efforts for estimating BMP implementation data to report and credit in the Bay model. Recommendations are presented assuming that the primary purpose for using remote sensing is to identify and inventory BMPs. All recommendations presented here assume that field verification meets CBPO requirements for identifying and characterizing BMPs and that the results of the ground verification study are representative of the full remote sensing study. In addition, the assessment performed here assumed that all occurrences of a specific BMP represent equivalent BMP extent (e.g., acres, feet).

Options for assessing the suitability of remote sensing methods within the context described here include:

- Comparing single performance measures such as FAR and HR with threshold values
- Comparing multiple performance measures such as FAR and HR with threshold values

Options for crediting the results of data generated via suitable remote sensing include:

- Crediting all identified BMPs
- Scaling estimates of BMP quantity to correct for bias using the ratio of PAG/HR

Based on the available literature (Sullivan et al. 2008) a FAR value of 0.2 or 0.3 and a HR value of 0.7 or 0.8 may be appropriate thresholds for determining whether a remote sensing effort generates suitable BMP data. Single application of the FAR threshold would be appropriate when over-counting of BMPs is the primary concern. It is recommended that the upper confidence level (α=0.10) be used as the threshold value. Single application of the lower confidence level (α=0.10) of HR be used when under-counting of BMPs is the primary concern. Where over-counting and under-counting of BMPs are of equal concern it is recommended that thresholds for both FAR and HR be established.

A minimum sample size is also recommended. Analyses presented here and the study by Sullivan et al. (2008) indicate that a minimum sample size of 20 (i.e., 20 instances of the BMP ground verified) may be suitable. Although smaller minimum sample sizes might be appropriate for relatively uncommon conservation practices (i.e., accounting for finite population correction on confidence intervals), insufficient information was found in the literature to support an alternative strategy for addressing minimum sample sizes for ground truthing.

A two-step process may be appropriate where the first step requires that the following conditions be met:

- Sample size is at least 20
- FAR (upper confidence limit value is recommended) is at or below the threshold value
- HR (lower confidence limit value is recommended) is at or above the threshold value

If these conditions are met, the estimate of BMP quantity would then be corrected for bias using the ratio of PAG/HR:

- Lower confidence limit value is recommended for a conservative estimate.

References

Congalton RG, and K. Green. 2008. *Sample Design Considerations. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. 2nd edition. CRC Press, Boca Raton, Florida.

Gowda, P.H., B.J. Dalzell, D.J. Mulla, and F. Kollman. 2001. Mapping tillage practice with Landsat Thematic Mapper based logistic regression models. *Journal of Soil and Water Conservation* 56(2):91-96.

NIST (National Institute of Standards and Technology). 2010. *Binomial Ratio Confidence Limits*. Last updated October 5, 2010. Accessed January 19, 2017. http://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/binoraci.htm

Schaefer, J.T. 1990. The critical success index as an indicator of warning skill. *Weather and Forecasting*. 5:570-575.

http://www.nssl.noaa.gov/users/brooks/public_html/feda/papers/schaefer1990.pdf (Accessed November 1, 2016).

Sullivan, D.G., T.C. Strickland, and M.H. Masters. 2008. Satellite mapping of conservation tillage adoption in the Little River experimental watershed, Georgia. *Journal of Soil and Water Conservation* 63(3):112-119.

Tetra Tech. 2014. *Tile Drainage Management Opportunities*. Prepared for U.S. Environmental Protection Agency, Region 5, by Tetra Tech Inc., Cleveland, OH.

van Deventeer, A.P., A.D. Ward, P.H. Gowda, and J.G. Lyon. 1997. Using Thematic Mapper data to identify contrasting soil plains and tillage practices. *Photogrammetric Engineering and Remote Sensing* 63:87-93.

Waz, A. 2016. An Automated Framework to Identify Lost and Restorable Wetlands in the Prairie Pothole Region. Master's thesis, University of Western Ontario.