Combining imperfect automated annotations of underwater images with human annotations to obtain precise and unbiased population estimates

Jui-Han Chang[1], Deborah R. Hart[1*], Burton V. Shank[1], Scott M. Gallager[2], Peter Honig[2], and Amber D. York[2]

[1]Northeast Fisheries Science Center, National Marine Fisheries Service,
166 Water St., Woods Hole, MA 02543, USA
[2]Biology Department, Woods Hole Oceanographic Institution,
Woods Hole, MA 02543, USA

*Email: deborah.hart@noaa.gov

August 10, 2016

-

Abstract

Optical methods for surveying populations are becoming increasingly popular. These methods often produce hundreds of thousands to millions of images, making it impractical to analyze all the images manually by human annotators. Computer vision software can rapidly annotate these images, but their error rates are often substantial, vary spatially and are autocorrelated. Hence, population estimates based on the raw computer automated counts can be seriously biased. We evaluated four estimators that combine automated annotations of all the images with manual annotations from a random sample to obtain (approximately) unbiased population estimates, namely: ratio, offset, and linear regression estimators as well as the mean of the manual annotations only. Each of these estimators was applied either globally or locally (i.e., either all data were used or only those near the point in question, to take into account spatial variability and autocorrelation in error rates). We also investigated a simple stratification scheme that splits the images into two strata, based on whether the automated annotator detected no targets or at least one target. The 16 methods resulting from a combination of four estimators, global or local estimation, and one stratum or two strata, were evaluated using simulations and field data. Our results indicated that the probability of a false negative is the key factor determining the best method, regardless of the probability of false positives. Stratification was the most effective method in improving the accuracy and precision of the estimates, provided the false negative rate was not too high. If the probability of false negatives are low, stratified estimation with the local ratio estimator or local regression (essentially geographically weighted regression) are best. If the probability of false negatives are high, no stratification with a simple global linear regression or simply the manual sample mean alone is recommended.

Keywords: Underwater imagery; Computer vision; Population estimation; Scallop; Geographically weighted regression

1 Introduction

Underwater optical surveys of fish and invertebrate populations are becoming increasingly common (e.g., Davis et al., 1992; Gallager et al., 2005; Howland et al., 2006; Yoklavich et al., 2007; Rosenkranz et al., 2008; Taylor et al., 2008; Tolimieri et al., 2008; Singh et al., 2013; Gallager et al., 2014). Such surveys have numerous advantages over traditional surveys using fishing gear, including being able to observe populations at all scales under natural conditions, and detection efficiency that potentially approaches 100%.

Optical surveys often generate hundreds of thousands to millions of images. Manually annotating all of the images (i.e., having people identifying the targets of interest in each image) would thus often be impractical. The traditional statistical approach to this problem would be to only manually annotate a sample of the images and obtain inferences on the population (which for our purposes is defined as the targets contained in all of the collected images) based on the sample. Alternatively, computer vision software can produce "automated annotations" that identify the targets in every image. However, automated annotators can make errors, both because they may not detect some targets ("false negatives") and because the annotator mistakenly identifies some objects ("distractors") as targets when they are not ("false positives"). Thus, analyses based on the raw automated counts can be seriously biased. Errors from automated annotations are often autocorrelated and spatially non-stationary due to, for example, a certain region having high densities of distractors or reduced visibility. Manual annotations of a sample of the images can help detect and correct for errors by the automated annotators, in which case the goal is to produce estimators for the population, based on the combination of automated and manual annotations that are more efficient than using the manual annotations alone (i.e., the variances of estimators are less than the variance of the sample mean of the manual images), as well as being at least approximately unbiased.

Although there have been numerous studies devoted to automated detection and classification of marine organisms (e.g., Culverhouse et al., 2006; Marcos et al., 2008; Spampinato et al., 2008; Beijbom et al., 2012), these studies usually conclude with estimating confusion matrices or error rates. The final task of obtaining estimates of the population of targets in all images from automated annotations that contain errors has received less attention. Solow et al. (2001) considered the situation where classification of plankton samples may be in error, which were corrected by inverting the confusion matrix (see also Hu and Davis, 2006; Verikas et al., 2015). The problem they considered is simpler than the one we are considering here because they were only concerned with classification of an object but not its detection, and because errors were assumed to be stationary and not autocorrelated. Beijbom (2014) analyzed what we have termed the offset estimator to bias-correct automated counts using a random sample of manual annotations from a cost reduction point of view.

The purpose of this paper is to explore and compare performance of several methods for estimating population abundance (or biomass) based on automated annotations of all images combined with manual annotations of a random sample of the images. This study is motivated by surveys of sea scallops (*Placopecten magellanicus*) using the HabCam (Habitat Mapping Camera System) towed underwater camera system (Howland et al. 2006; Taylor et al., 2008; NEFSC, 2014; see Figure 1 for an example of HabCam images of sea scallops and sand dollars, a common distractor). Computer vision software for detecting sea scallops is continuing to be developed (Dawkins et al., 2013; Kannappan et al., 2014; Gallager et al., unpublished). The U.S. sea scallop fishery has annual ex-vessel revenue averaging around $500 million in recent years, so obtaining accurate and precise estimates of sea scallop abundance is of immediate practical significance.

2 Methods, Theory, and Calculation

2.1 Global Population Estimators

We tested four different estimators of population size (i.e., the number of true targets in an image set) based on a combination of manual and automated annotations. In the following, it is assumed that each image has been annotated by software, but only a random sample of n images out of a total of N images have been annotated manually, and the manual annotations are without error (it is straightforward to extend the theory to cases where only a sample has been annotated by software). Let X_i and Y_i be the number of targets detected in the ith image by the automated and manual annotators, respectively.

Four global estimators for the total number of targets in the images, Z, are:

Manual sample only: $$Z_m = \bar{Y}N \tag{1}$$

Ratio estimator: $$Z_r = \mu_X N \frac{\bar{Y}}{\bar{X}} \tag{2}$$

Offset estimator: $$Z_o = \sum_{i=1}^{N} X_i - \frac{N}{n}\sum_{j=1}^{n}(X_j - Y_j) \tag{3}$$

Regression estimator: $$Z_g = \sum_{i=1}^{N} \alpha + \beta X_i \tag{4}$$

where $\bar{X}$ and $\bar{Y}$ are the mean number of targets detected by automated and human annotators in the sample of images that have been manually annotated, μ_X is the mean number of targets over all images detected by the automated annotator, and α and β in equation (4) are the intercept and slope obtained by regressing the automated vs. manual annotations. The last three methods can be considered as ways to adjust, or bias correct, the automated counts based on the comparison between the automated and manual counts in the sample. The ratio estimator adjusts the automated counts by a multiplicative constant, the offset estimator adjustment by an additive constant, and the regression estimator combines additive (intercept) and multiplicative (slope) adjustments.

Although the ratio estimator (2) is biased, this bias is negligible for all the simulated datasets because the coefficients of variation of $\bar{X}$ and $\bar{Y}$ are both smaller than 0.1 (Cochran, 1977), which should typically be the case because the sample sizes for both the automated and manual annotations will usually be large. An approximate bias correction can be applied if this is a concern. The Appendix derives analytically the conditions when the variance of the ratio estimator applied to a random sample is lower than manual sampling alone. Beijbom (2014) similarly gave analytic derivations of properties of the offset estimator of a random sample.

2.2 Local Population Estimators

The automated annotator error rate may vary spatially, depending on factors such as water clarity, substrate type, and the densities of targets and distractors. All these factors, and therefore the automated annotator error rates, are typically spatially autocorrelated. If this is the case, it may be more efficient to bias-correct the automated annotations locally, rather than using a single global correction as in equations (1)-(4). In addition, the spatial distribution of the population is often of interest. If the error rates vary spatially, the correction for these errors also needs to vary accordingly to accurately reflect the actual distribution of the population.

For the local estimators, the correction factor is calculated for each data point, and the estimators are similar to the global estimators described above except that only data less than a distance, or "bandwidth", h_j from the point j are used, and the data are weighted as a decreasing function of the distance from the target data point, using an adaptive bisquare distance decay kernel function:

$$w_{(j,k)} = \begin{cases} \left[1 - (\dfrac{d_{(j,k)}}{h_j})^2\right]^2 & d_{(j,k)} \leq h_j \\ 0 & d_{(j,k)} > h_j, \end{cases} \tag{5}$$

where $w_{(j,k)}$ is the weighting factor of point k that is used to calculate the bias correction factor for point j, and $d_{(j,k)}$ is the distance between points j and k. The bandwidth is adapted to the density of the data; it is larger when data are sparser and smaller when data are denser. Even though the bandwidth may vary by location, the number of data points within the bandwidth is the same across locations. The bandwidth (or number of data points to be included at each location) is determined by minimizing the leave-one-out cross-validation squared error:

$$CV = \sum_{j=1}^{n} \left[Y_j - \hat{Y}_{\neq j}(h_j)\right]^2, \tag{6}$$

where $\hat{Y}_{\neq j}(h_j)$ is the fitted value of Y_j with the data points where point j is omitted from the estimation process (Guo et al., 2008).

The local method for the regression estimator is essentially a form of geographically weighted regression (GWR) that is used specifically for situations when the relationship between variables differs across space (i.e., spatial non-stationarity and spatial autocorrelation; Brunsdon et al., 2008). Compared to standard (global) regression models where a single parameter set is estimated for the entire dataset, GWR estimates regression parameters that vary for each data point based on data that is in the local neighborhood of that point.

2.3 Stratification

Population densities from underwater images are often “zero-inflated”, i. e., a high proportion of photos contain no targets. In such a case, the images can be separated into two strata: one where no targets were detected by the automated annotator, and the other where at least one target is detected. Manual annotations are then allocated among the two strata based on the automated annotations and their overall false negative rates, using approximate Neyman optimal allocations. For this purpose, the standard deviation of the true target counts in the zero stratum, s_0, is: $\sqrt{Z_0 P_S(1 - P_S)}$, where Z_0 is the number of targets in the zero stratum (i.e. the number of false negatives), P_S is the probability of detecting a target by the automated annotator, and $1 - P_S$ is the probability of a false negative. In the simulation, Z_0 and P_S are known, but in practice, they would have to be estimated either from previous data or by obtaining a small sample of manual annotations prior to the allocation. The standard deviation of targets in the non-zero stratum, s_1, is approximated by the standard deviation of the automated counts in this stratum. The Neyman optimal allocation is then:

$$n_m = \frac{n N_m s_m}{\sum_{m=0}^{1} N_m s_m}, \tag{7}$$

where n is total number of manual sample size, and N_m is the total number of images in stratum m.

2.4 Simulation Design

We tested the performance of the above methods using simulated data. The simulation design is based on the US sea scallop population characteristics as observed by the HabCam survey. The simulation domain is 70 km (longitude) by 140 km (latitude), with a 50 m grid size, roughly corresponding to the density of annotated images in actual data sets. The spatial distribution of sea scallops is non-stationary due to the influences of physical and biological environment including current, depth, and predator distributions (Brand, 2006).

Therefore, we assumed that the simulated scallop population has large-scale smooth trends in its expected mean (first-order effect) that are added to a stationary autocorrelated random field (second-order effect; Cressie, 1993). We simulated the variations of global mean density using a double logistic function that is constant with latitude but varies with longitude:

$$p(l) = \begin{cases} \dfrac{1}{1+\exp\left(-a(l-b)\right)} & l \leq \dfrac{l_{max}}{2} \\ \dfrac{1}{1+\exp\left(a(l-b-\dfrac{l_{max}}{2})\right)} & l > \dfrac{l_{max}}{2}, \end{cases} \tag{8}$$

where l is longitude, l_{max} is the maximum longitude in the surveyed area, and a and b are the parameters that determine the shape of the logistic curve. The simulated first-order effects are high in the middle and decrease logistically toward the left and right edge of the simulation domain, which is typical of actual scallop distribution patterns (Hart, 2006). The second-order effects were simulated using stationary Gaussian random fields with a spherical isotropic covariance structure (Cressie, 1993):

$$\gamma(d) = \begin{cases} 0 & d = 0 \\ c_0 + c_1\{\frac{3}{2}\frac{d}{r} - \frac{1}{2}(\frac{d}{r})^3\} & 0 < d \leq r, \\ c_0 + c_1 & d \geq r \end{cases} \tag{9}$$

where c_0, c_1, and r are the nugget, partial sill, and range parameter, respectively. The nugget/sill (n/s) ratio $(\dfrac{c_0}{c_0+c_1})$ determines randomness and r determines the aggregation size of the second-order effects. We chose the simulation parameter values based on estimates from the actual HabCam data.

To reflect the highly zero-inflated nature of scallop distributions, those locations where the sum of the first-order and second-order effects values were smaller than its 90th percentile were set to zero. The simulated scallops count for the remaining 10% is simply the sum of the first- and second-order effects (Figure 2). The resultant simulated data is patchy, zero-inflated, and has a large scale trend along one direction, consistent with actual scallop populations. The shape and direction of tracks used to survey the simulated population was designed to mimic the actual HabCam survey design, where more effort was put in the middle high density area (Figure 2; NEFSC, 2014). A total of 9,001 photos were simulated along the track (Figure 2).

False positives were simulated by using distractors. The two most common distractors for sea scallops are sand dollars (*Echinarachnius parma*; Figure 1) and dead scallop shells (Dawkins et al., 2013; Kannappan et al., 2014). The distribution of sand dollars are typically independent or negatively correlated with scallops, whereas dead scallop shells would be expected to be positively related to (live) scallops. The spatial distribution of distractors were simulated similar to scallops, but the distractor's patches were assumed larger (larger range) and less noisy (smaller n/s ratio) than the scallop target distribution, based on actual observations of sand dollars (Figure 2).

Water visibility may affect automated annotation accuracy by reducing the probability of detecting a target or a distractor. We simulated water visibility to be trendless but with spatial autocorrelation. It other words, it is a random field with no first-order effect. It was assumed to have the same noise level but larger patch size as the distractor (larger range; Figure 2).

2.5 Simulation of Automated Count Data

The simulated manually annotated data are assumed to have no errors. For the computer automated counts, each simulated target (S) and distractor (D) has a probability of being detected as a target by the automated annotator:

$$P_S = (1 - F1_S)(1 - F2_S) \text{ and } P_D = 1 - (1 - F1_D)(1 - F2_D), \tag{10}$$

where the $F1_S$ and $F1_D$ are the probabilities of a false negative and false positive with good water visibility, and $F2_S$ and $F2_D$ are the reduced probabilities of detecting targets and distractors due to water visibility. In our simulations, it is assumed that $F2_S = F2_D$. The simulated total number of targets reported by the automated annotator in the ith image is:

$$X_i = \sum_{m=1}^{M} (S_{im} + D_{im}), \tag{11}$$

where M is the total number of objects simulated within image i, S_{im} is the number of correctly identified targets (true positives minus false negatives), and D_{im} is the number of distractors incorrectly identified as targets (false positives).

2.6 Scenarios Tested

To understand whether the estimation methods are robust to changes in the environment, species distributions and the capabilities of the automated annotator, we tested the performance of these methods by varying the following quantities:

(1) Automated annotator's performance: probability of a false negative/positive ($F1_S$ and $F1_D$) from 0 to 1 by 0.05;
(2) Water visibility: good, moderate, or poor (expected value of $F2 = 0, 0.05, 0.1$);
(3) Correlation between scallop and distractor distribution: negative, zero, or positive;
(4) Degree of spatial autocorrelation of distractors: low, medium, and high;
(5) Percent of total sample size that was annotated manually: 1%, 3%, 7%, 11%, and 15%.

A base case was selected where the water visibility is good, the correlation between the spatial distribution of scallops and distractors is negative, the spatial autocorrelation of distractors is medium, and manual annotations were performed on 7% of the photographs. The base case was then varied for each of the attributes (2)-(5) individually, keeping the other three at their base case values. Thus, a total of 14 scenarios were simulated. For each choice of (2)-(5), $F1_S$ and $F1_D$ were varied from 0 to 1 by 0.05 increments, as specified in (1).

For all scenarios, scallops have high densities in middle longitudes of the simulation domain (simulated using equation 8), and water visibility has no first-order effects. Distractors have high first-oder effects on the left (which used only the second part of the equation 8 on $l \leq \dfrac{l_{max}}{2}$ part of the simulation domain), except for the scenarios of zero and positive correlations between scallop and distractor distribution where there are no effects or high effects in the middle, respectively. The partial sill, n/s ratio, and range parameter used to simulate second-order effects are 0.18, 0.6, and 200 for scallops, 0.18, 0.6, and 400 for distractors, and 0.18, 0.6, and 600 for water visibility. For the scenarios where distractors have high and low autocorrelation, the n/s ratio is 0.3 and 0.9, respectively. For the scenarios where water visibility is moderate or poor, the effects of water visibility on the probability of a false negative and false positive is one or two times, respectively, compared to the corresponding scenarios of good water visibility.

For each scenario, the manual annotation subset was resampled 30 times. For each iteration, we tested the combinations of the four estimators applied either globally or locally, and using two strata or one stratum (unstratified) to allocate manual annotations, resulting in 16 different estimation methods.

For stratified estimation, the ratio estimator is undefined in the zero stratum, so the mean of the manual annotations in this stratum was used instead. Since the offset and regression estimators reduce to simply taking the mean of the manual annotations in the zero stratum, all four methods produce the same estimate in this stratum, so any differences among the methods with stratification stem from the non-zero stratum.

2.7 Field Data Analysis

HabCam images from the US sea scallop survey (NEFSC, 2014) were used to illustrate the usefulness of the methods discussed above on real data. For testing purposes, all the images were annotated using computer vision software (Gallager et al., unpublished) and also manually annotated, so that the estimates can be compared to their true values.

The automated annotator used a series of features including texture, color, and shape. A kernel of 100 x 100 pixels was run through each image left to right, top to bottom, extracting each feature set resulting in a feature vector of length 480 by width 3 (texture, color, and shape). Texture features were extracted using a 2-dimensional Gabor wavelet convolved with Gaussian kernels at 360 orientations for each pixel box providing rotational independent texture features (Gallager and Tiwari, 2008). Color was extracted in L*A*B* color space using the color angle approach, where the standard deviation of the gradient between the pixel radius at 10 degree increments was extracted with 128 colors (Gallager and Tiwari, 2008). For each kernel, a Canny edge detection algorithm was used followed by extraction of Fourier shape descriptors. A Principal Component Analysis was run to reduce data dimensionality from > 4000 to 128 principal components. Finally, a linear Support Vector machine was trained on 3800 images containing scallops of various sizes as well as images containing no scallops over varying substrate conditions. The result was a probability of the presence of a scallop; a scallop was considered as detected if this probability was greater than 90%.

One out of every 50 images collected were annotated manually as well as with software (Table 1), and this collection of images served as the data for our analysis. Data from three regions with various probability of a false negative were selected. The probability of a false positive could not be defined for our datasets because number of possible distractors for each image was not identified. For each region, the manual annotations from a 7% random subset of the images were used for estimation along with automated annotations from each image; error rates could therefore be assessed because each image in the datasets were annotated manually, even though only a sample of the manual annotations were used in the analysis. The manual annotation subset was resampled 2000 times, and the various estimation methods were applied to each iteration.

In the field, factors such as vehicle altitude, depth, etc. may also influence the performance of the estimators. We tested an additional method that included auxiliary variables in the two-strata local regression:

$$Y_j = a_0(u_j, v_j) + a_1(u_j, v_j)X_j + \sum_{b=2}^{5} a_b(u_j, v_j)A_{bj} + \epsilon_j \tag{12}$$

where (u_j, v_j) is the coordinates of point j and $a_b(u_j, v_j)$'s are the coefficients of variables A including altitude, depth, squared depth, and latitude at location (u_j, v_j) point j.

2.8 Evaluation of Methods

For both simulation and field data analysis, mean squared error (MSE) and mean absolute error (MAE) were used as the principal measures of precision and bias:

$$\text{MSE} = \frac{1}{K}\sqrt{\sum_{k=1}^{K}(Z_k - \mu)^2} \text{ and } \text{MAE} = \frac{1}{K}\sum_{k=1}^{K}|Z_k - \mu|, \tag{13}$$

where Z is the population estimates based on automated and manual annotations, μ is the true population abundance, and K is the number of iterations. These were reported relative to the global unstratified manual sample mean (M1G):

$$\mathrm{MSE}_{\mathrm{re}} = \frac{\mathrm{MSE} - \mathrm{MSE}_{\mathrm{M1G}}}{\mathrm{MSE}_{\mathrm{M1G}}} \text{ and } \mathrm{MAE}_{\mathrm{re}} = \frac{\mathrm{MAE} - \mathrm{MAE}_{\mathrm{M1G}}}{\mathrm{MAE}_{\mathrm{M1G}}}. \tag{14}$$

MAE and MSE both reflect precision as well as bias but MSE weights more on large errors than small ones.

3 Results

3.1 Simulation Results

Combining automated and manual annotations using our methods increased precision of the estimates over manual counts alone by up to a maximum of 73%, whereas using the uncorrected automated counts could decrease both accuracy and precision up to 717%, compared to using manual counts only (Tables 2 and 3). Increasing the number of manual samples increased the precision of all methods, but only by a modest amount (up to 15%).

In the base case, splitting the annotations into two strata was the most effective way of improving estimation precision, except at very high false negative rates where stratification degraded the estimates (Figures 3 and 4). When both false negative and false positive rates are low, the use of automated data for stratification and/or estimation substantially improves the precision and accuracy of the estimates regardless of the estimator used. Local models were superior to global models only when stratification was employed. For one-stratum allocation and when the probability of a false positive is high, the ratio and regression estimator performed better, whereas the offset estimator was better when the probability of a false negative is high but false positive rate is low. Similar patterns were observed for the other scenarios tested, i.e., the performance of the bias correction methods we tested are robust to changes in the environment and species distributions.

The probability of a false negative is the key factor determining the most effective bias correction methods, regardless of the level of probability of a false positive (Tables 2-5). When the probability of a false negative is low, nearly all the methods can improve the accuracy and precision of the population estimates, but stratification with the local ratio or the local regression estimator was generally superior. If the probability of false negatives is high, no stratification with a simple global linear regression or manual sampling alone tended to have the best performance. If in addition the false positive rate is low, the global offset estimator also performs well.

3.2 Field Data Analysis Results

Results from the field data analysis were consistent with those from the simulations. Estimations of the mean using automated annotations alone were 63% to 498% higher than the simple manual sample mean (Table 1). For the region with low false negative rates (0.31), the two-strata local regression without auxiliary variables and two-strata local ratio estimator were superior; these increased precision over the simple manual sample mean by up to 51% (Table 1). When the false negative rate was higher (0.73-0.75), global regression or simply the manual sample mean were the best, with the global regression model improving precision by at most 11% over the simple manual sample mean. The offset estimator performed better than the ratio estimator in one case, likely because the false positive rate of this dataset is low; however, this is not totally clear since the false positive rates were not available for all of our field data. Auxiliary variables did not improve the performance of local regression for these data.

4 Discussion

The results indicate that combining even a mediocre automated annotator with manual annotations may be able to improve statistical efficiency over manual annotations alone when using the methods presented here. The combination of automated and manual annotations outperformed manual or (unadjusted) automated annotations alone, even when the false positive and false negative rates were as high as 0.5. The results from both simulations and field data analysis are consistent, and indicate that probability of a false negative is the key factor determining the best estimation method. The probability of a false positive does matter to some extent, especially when the probability of a false negative is higher, but even in this case, it is not the main factor determining the best method.

Stratification based on zero and positive automated counts is the most effective technique to improve precision except at very high false negative rates. Stratification directly improves precision when the within-strata variance is less than the between strata variance (Cochran, 1977), which is likely to be the case for even a moderately effective automated annotator. In addition, the allocation of manual samples between the two strata often further increases performance by allocating disproportionately more manual samples to the more variable stratum. Stratified estimates are in particular more precise at high false positive but low false negative rates. The zero stratum has no false positives, and contains a limited number of actual targets when the false negative rates are low. The zero stratum thus tends to have a low variance, so the number of targets in this stratum can be estimated precisely by a relatively small number of manual samples. This allows for higher sampling rates in the non-zero stratum, increasing the precision there.

The simple two-strata stratification presented here is natural for zero-inflated data such as in our examples. In some cases, more complex stratification may give further benefits. For example, there could be three strata, composed of where the automated annotator detects zero, one or more than one targets. We implicitly assumed for simplicity that the cost of a manual annotation is the same in each stratum. In reality, the labor cost of annotating an image tends to go up with the number of targets in the image. If this cost function is known, it can be taken into account in the optimal allocation among strata (Cochran, 1977).

In real world situations, the false negative (and positive) rates may be uncertain. In such cases, we recommend manually annotating a small sample of images to roughly estimate this rate, and select the manual sampling strategy (e.g., stratification scheme) and estimator based on this information. The optimal strategy is fairly robust to modest changes in the automated annotator error rates, so only a crude estimate of the false negative rates is needed to design a sampling strategy.

The offset estimator, by its definition, can account for errors that are independent of the target density, but less efficient in tracking errors that vary with the targets. Conversely, the ratio estimator is more effective without stratification when there are false negatives but few false positives (Figures 3 and 4), because the ratio estimator can take into account errors that are proportional to the target density. The precision of the ratio estimator depends on the correlation between automated and true counts (see Appendix); false positives directly reduce this correlation.

In principle, the regression estimator should be able to account for both these types of errors, but it has the disadvantage of having two parameters that can be confounded with each other, especially at low sample sizes and when the data are zero-inflated. For stratified local regressions, the manual sample size used to estimate the regression parameters at each location is low, and might be one of the reasons why its performance is slightly lower than the stratified local ratio estimator. The difference in performance of stratified local regression estimator and stratified local ratio estimator was larger when the manual sample size is only 1% and became smaller as the manual sample size increased (Tables 2 and 3).

There are nonetheless some advantages of regression methods. For example, multiple regression can be used if there is more than one automated annotator available, using counts from each automated annotator as predictors. Even though in our example field data it was not effective, auxiliary variables such as water depth, latitude, or substrate type may sometimes also be useful as predictors in a multiple regression.

Local estimation methods can improve estimates when the distribution of targets or errors is autocorrelated. In particular, false positives induced by distractors such as sand dollars and dead scallop shells are typically autocorrelated. False negative rates could be in some cases also autocorrelated (caused by e.g., poor visibility), but this would normally be a weaker effect than false positives if it exists at all. Stratification isolates the false positives in one stratum, which may be the reason that it enhances the effectiveness of using local estimation methods. The benefits of local estimation methods are however minor compared to stratification, even in the presence of substantial autocorrelation.

Although computer vision methods are rapidly improving, it is unlikely that automated detection of underwater organisms will be error free in the foreseeable future. Many marine organisms are cryptic, and can adjust their pattern and coloration to match their surroundings, thus making it difficult to totally eliminate false negatives. For scallops in particular, false negatives can be caused by colonization of their shell by epifauna or the shell being covered by marine snow or sediments. In addition, a small percentage ($\sim$5-10%) of sea scallops are “albinos”, with white upper shells, that are difficult to distinguish from dead scallop shells. While we believe that the false positives induced by sand dollars can be reduced considerably compared to present methods, it is also unlikely that false positives can be completely eliminated (for example, it is sometimes difficult to distinguish a dead scallop shell from a live scallop). Thus, combining automated and manual annotations using the methods described here is likely to continue to be an improvement over using either automated or manual annotations alone.

While we have focused on automated annotations of marine organisms, our methods are applicable to a much wider set of problems. For example, our methods could be employed whenever there are at least two observers counting the same things, one of whom is an expert (or is a reference collection) who is considered error free but only observes a sample. Annotations using crowd-sourcing (Simpson et al., 2014) may be subject to higher error rates than those done by experts, which can be corrected using the techniques presented here. Our methods also are applicable to automated or crowd-sourced annotations of a variety of targets beyond those underwater, such as targets from aerial photography, surveillance cameras, medical imaging and testing, and industrial quality control.

5 Acknowledgements

We thank Chris Legault and two anonymous reviewers for their constructive comments. The research was supported in part by funding from the NOAA Automated Image Analysis Strategic Initiative.

References

Beijbom, O. 2014. Random Sampling in an Age of Automation: Minimizing Expenditures through Balanced Collection and Annotation. arXiv preprint arXiv:1410.7074.

Beijbom, O., Edmunds, P.J., Kline, D., Mitchell, B.G., Kriegman, D., 2012. Automated annotation of coral reef survey images. IEEE Conference In Computer Vision and Pattern Recognition (CVPR), p. 1170-1177.

Brand, A.R., 2006. Scallop ecology: distributions and behavior. In: Scallops: Biology, Ecology, and Aquaculture, Elsevier, Amsterdam, p. 651-744.

Brunsdon, C., Fotheringham, A.S., and Charlton, M., 2008. Geographically weighted regression: a method for exploring spatial nonstationarity. Encyclopedia of Geographic Information Science, 558.

Culverhouse, P.F., Williams, R., Benfield, M., Flood, P. R., Sell, A.F., Mazzocchi, M.G., Buttino, I. Sieracki, M., 2006. Automatic image analysis of plankton: Future perspectives. Mar. Ecol. Prog. Ser. 312, 297-309.

Cochran, W.G., 1977. Sampling Techniques: 3rd Ed., Wiley, New York.

Cressie, N.A.C., 1993. Statistics for Spatial Data, revised edition. Wiley, New York.

Davis, C.S., Gallager, S.M., Solow, A.R., 1992. Microaggregations of oceanic plankton observed by towed video microscopy. Science 257(5067), 230-232.

Dawkins, M., Stewart, C., Gallager, S., York, A., 2013. Automatic scallop detection in benthic environments, 2013 IEEE Workshop on Applications of Computer Vision (WACV), p. 160-167.

Gallager, S.M., Singh, H., Tiwari, S., Howland, J., Rago, P., Overholtz, W., Taylor, R., Vine, N., 2005. High resolution underwater imaging and image processing for identifying essential fish habitat. Report of the National Marine Fisheries Service Workshop on Underwater Video analysis. DA Somerton and CT Glendill (eds) NOAA Technical Memorandum NMFS-F/SPO-68. p. 44-54.

Gallager, S.M., Tiwari, S., 2008. Optical method and system for rapid identification of multiple refractive index materials using multiscale texture and color invariants. US patent

Number 7,415,136.

Gallager, S.M., Nordahl, V., Godlewski, J.M., 2014. The habitat mapping camera system (HabCam). Proceedings of the Undersea Imaging Workshop. R Langton and P Rowe, Eds.

Gallager, S.M., Honig, P., York, A.D, Hart D.R., Unpublished. Automated detection and classification of benthic fauna along the US Northeast Continental Shelf. Unpublished Results.

Guo, L., Ma, Z., Zhang, L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Can. J. Fish. Aquat. Sci. 38(9), 2526-2534.

Hart, D.R., 2006. Effects of sea stars and crabs on sea scallop (*Placopecten magellanicus*) recruitment in the Mid-Atlantic Bight. Mar. Ecol. Prog. Ser. 306, 209-221.

Howland, J., Gallager, S.M., Singh Girard, H., Abrams, L., Griner, C., 2006. Development of a towed, ocean bottom survey camera system for deployment by the fishing industry. IEEE Oceans p. 1-10.

Hu, Q., Davis, C.S., 2006. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. Mar. Ecol. Prog. Ser. 306, 51-61.

Kannappan, P., Walker, J.H., Trembanis, A., Tanner, H.G., 2014. Identifying sea scallops from benthic camera images. Limnol. Oceanogr.: Methods 12, 680–693.

Marcos, M.S.A., David, L., Peñaflor, E., Ticzon, V., Soriano, M., 2008. Automated benthic counting of living and non-living components in Ngedarrak Reef, Palau via subsurface underwater video. Environ. Monit. Assess. 145(1-3), 177-184.

Northeast Fisheries Science Center [NEFSC]. 2014. 59th Northeast Regional Stock Assessment Workshop: Assessment Report. Northeast Fisheries Science Center Reference Document 14-09.

Rosenkranz, G.E., Gallager, S.M., Shepard, R.W., Blakesleed, M., 2008. Development of a high-speed, megapixel benthic imaging system for coastal fisheries research in Alaska. Fish. Res. 92, 340–344.

Simpson, R., Page, K.R., De Roure, D., 2014. Zooniverse: observing the world's largest citizen science platform. In: Proceedings of the companion publication of the 23rd international conference on the World Wide Web International World Wide Web Conferences Steering Committee, p. 1049-1054.

Singh, W., Örnólfsdóttir, E.B., Stefansson, G., 2013. A camera-based autonomous underwater vehicle sampling approach to quantify scallop abundance. J. Shellfish Res. 32(3),725-732.

Solow, A., Davis C., Hu, Q., 2001. Estimating the taxonomic composition of a sample when individuals are classified with error. Mar. Ecol. Prog. Ser. 216, 309-311.

Spampinato, C., Chen-Burger, Y.H., Nadarajan, G., Fisher, R.B. 2008. Detecting, tracking and counting fish in low quality unconstrained underwater videos. VISAPP (2), 2008, 514-519.

Taylor, R., Vine, N., York, A., Lerner, S., Hart, D., Howland, J., Prasad, L., Mayer, L., Gallager, S., 2008. Evolution of a benthic imaging system from a towed camera to an automated habitat characterization system. IEEE Oceans p. 1-7.

Tolimieri, N., Clarke, M.E., Singh, H., Goldfinger, C. 2008. In: Reynolds, J.R. and Greene, H.G. (eds.), 2008. Marine Habitat Mapping Technology for Alaska, Alaska Sea Grant College Program, University of Alaska Fairbanks.

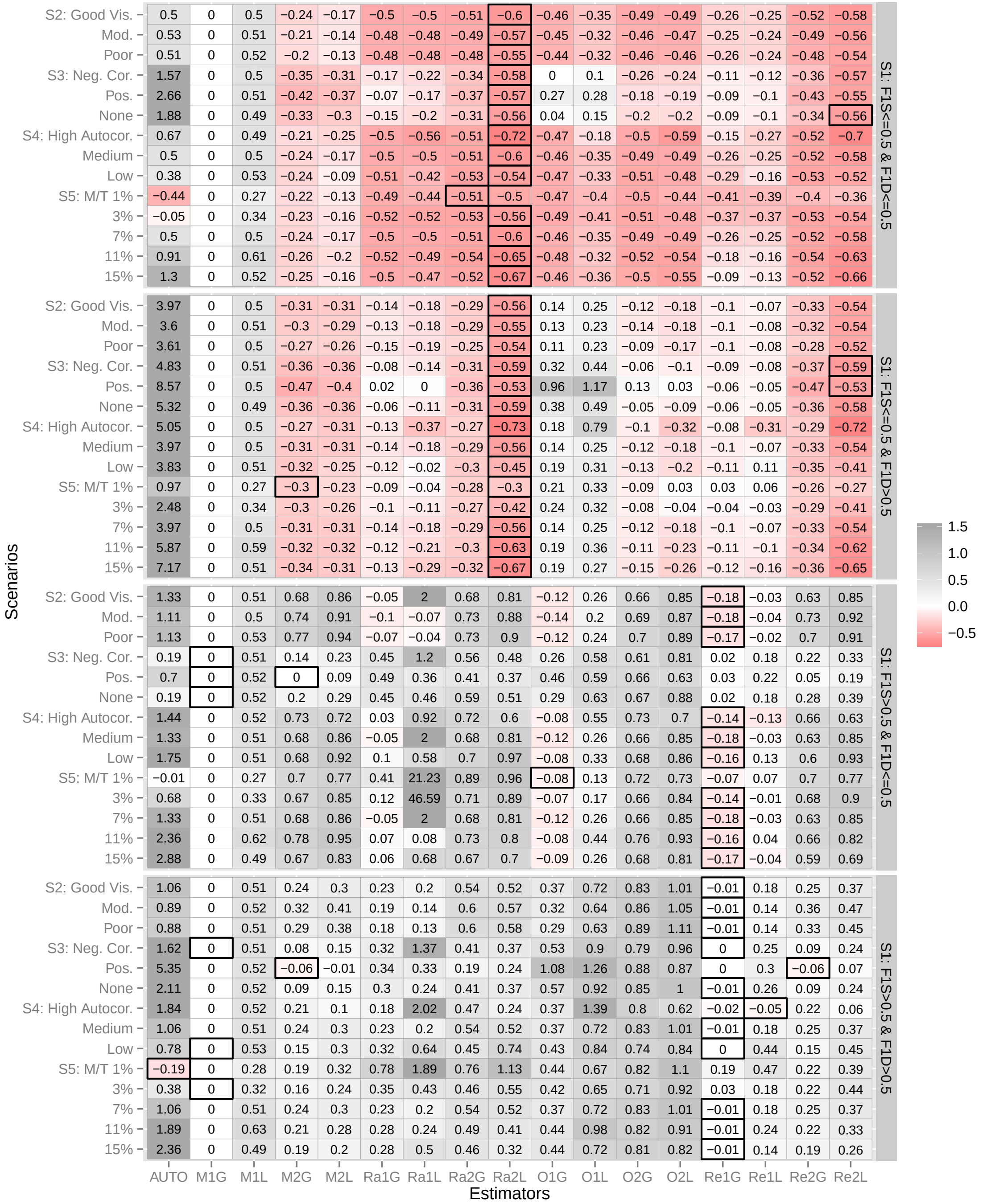
Verikas, A., Gelzinis, A., Bacauskiene, M., Olenina, I., Vaiciukynas, E., 2015. An Integrated Approach to Analysis of Phytoplankton Images. IEEE J. Ocean. Eng. 40(2), 315-326.

Yoklavich, M.M., Love, M.S., Forney, K.A. 2007. A fishery-independent assessment of an overfished rockfish stock, cowcod (*Sebastes levis*), using direct observations from an occupied submersible. Can. J. Fish. Aquat. Sci. 64(12), 1795-1804.

Table 1: Relative mean squared error (MSE_{re}) and relative mean absolute error (MAE_{re}) for each estimator, using unstratified (one-stratum) or two strata estimation, and either local or global estimation for three sets of actual HabCam field data. Error rates are relative to the global unstratified manual mean, which is used as a baseline. "AUTO" represents MSE_{re} or MAE_{re} calculated using only the automated annotations. "L+Var" represents local regression with auxiliary variables. The dark and light grey-shaded entries represent the best and second best method, respectively.

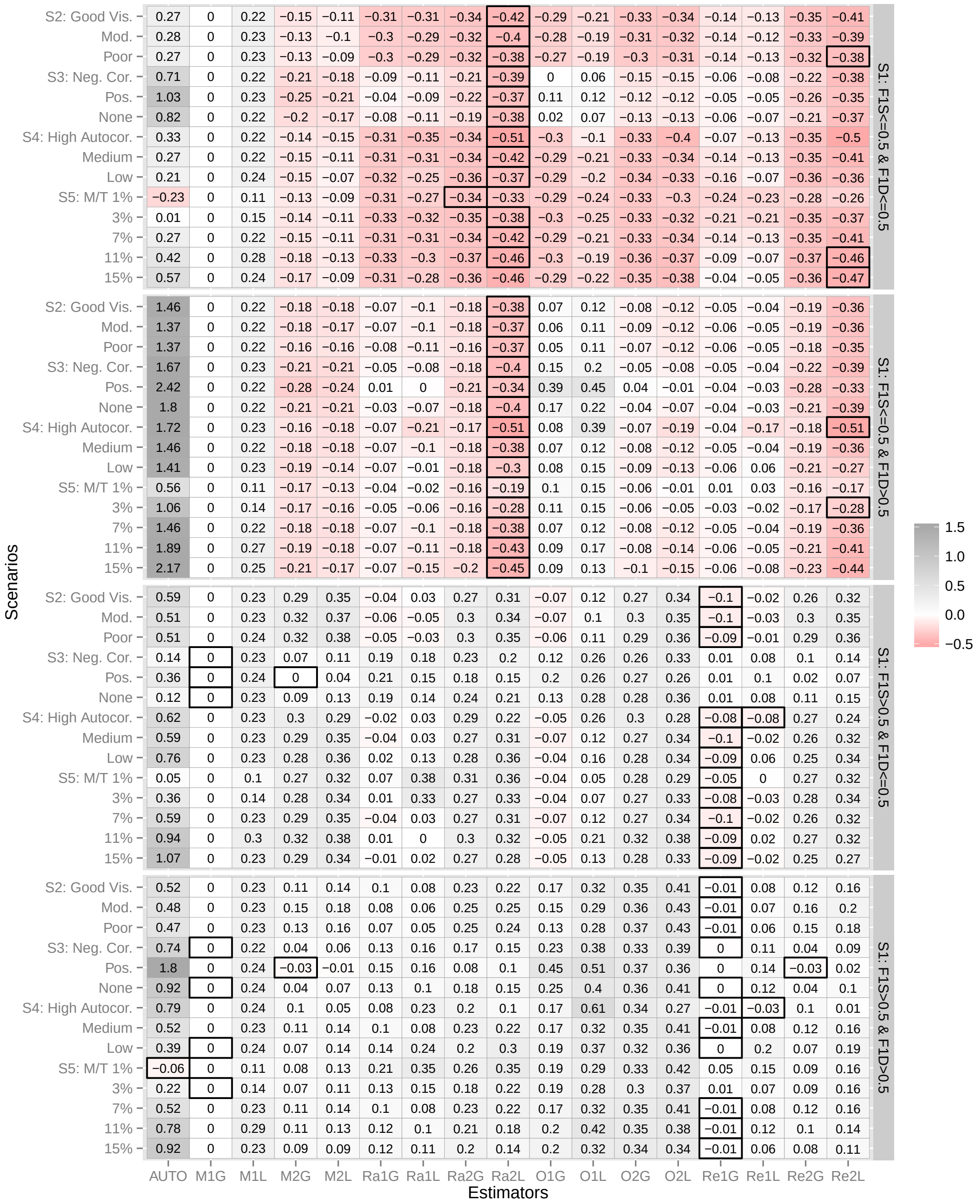
				Manual Mean				Ratio Est.				Offset Est.				Regression Est.				
Sample	False	Stat	Auto	One-stratum		Two-strata		One-stratum		Two-strata		One-stratum		Two-strata		One-stratum		Two-strata		
Size	Negative			Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	L+Var
5057	0.31	MSE_{re}	1.68	0	-0.06	-0.04	-0.21	-0.07	-0.31	-0.09	-0.50	-0.02	0.01	-0.02	-0.19	-0.07	-0.26	-0.10	-0.51	-0.04
		MAE_{re}	0.78	0	-0.03	-0.02	-0.11	-0.04	-0.16	-0.04	-0.29	-0.01	0.01	-0.00	-0.10	-0.04	-0.14	-0.05	-0.31	-0.08
9610	0.73	MSE_{re}	4.98	0	0.04	-0.04	0.02	-0.06	-0.01	-0.03	0.03	-0.10	-0.05	-0.04	0.02	-0.11	-0.07	-0.03	0.03	0.05
		MAE_{re}	1.68	0	0.01	-0.02	0.01	-0.03	-0.01	-0.01	0.01	-0.06	-0.03	-0.02	0.01	-0.06	-0.04	-0.02	0.01	0.01
14856	0.75	MSE_{re}	1.25	0	0.71	0.37	0.16	0.28	2.06	0.89	0.55	0.40	1.90	1.16	1.61	-0.00	0.93	0.37	0.07	0.04
		MAE_{re}	0.63	0	0.36	0.17	0.07	0.13	0.88	0.37	0.26	0.18	0.75	0.47	0.62	-0.00	0.47	0.17	0.03	0.02

Table 2: Relative mean squared error (MSE$_{\mathrm{re}}$, using the global unstratified manual mean as the baseline method) for the five scenarios by types of statistics (M: manual sample mean, Ra: ratio estimator, O: offset estimator, and Re: regression estimator), using global (G) or local (L), and one-statum (1) or two-strata (2) estimation, along with MSE$_{\mathrm{re}}$ calculated using only the automated annotations (AUTO). For each scenario, the cell outlined in bold is the best method.



Scenarios	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S1: F1S<=0.5 & F1D<=0.5																	
S2: Good Vis.	0.5	0	0.5	−0.24	−0.17	−0.5	−0.5	−0.51	**−0.6**	−0.46	−0.35	−0.49	−0.49	−0.26	−0.25	−0.52	−0.58
Mod.	0.53	0	0.51	−0.21	−0.14	−0.48	−0.48	−0.49	**−0.57**	−0.45	−0.32	−0.46	−0.47	−0.25	−0.24	−0.49	−0.56
Poor	0.51	0	0.52	−0.2	−0.13	−0.48	−0.48	−0.48	**−0.55**	−0.44	−0.32	−0.46	−0.46	−0.26	−0.24	−0.48	−0.54
S3: Neg. Cor.	1.57	0	0.5	−0.35	−0.31	−0.17	−0.22	−0.34	**−0.58**	0	0.1	−0.26	−0.24	−0.11	−0.12	−0.36	−0.57
Pos.	2.66	0	0.51	−0.42	−0.37	−0.07	−0.17	−0.37	**−0.57**	0.27	0.28	−0.18	−0.19	−0.09	−0.1	−0.43	−0.55
None	1.88	0	0.49	−0.33	−0.3	−0.15	−0.2	−0.31	**−0.56**	0.04	0.15	−0.2	−0.2	−0.09	−0.1	−0.34	**−0.56**
S4: High Autocor.	0.67	0	0.49	−0.21	−0.25	−0.5	−0.56	−0.51	**−0.72**	−0.47	−0.18	−0.5	−0.59	−0.15	−0.27	−0.52	−0.7
Medium	0.5	0	0.5	−0.24	−0.17	−0.5	−0.5	−0.51	**−0.6**	−0.46	−0.35	−0.49	−0.49	−0.26	−0.25	−0.52	−0.58
Low	0.38	0	0.53	−0.24	−0.09	−0.51	−0.42	−0.53	**−0.54**	−0.47	−0.33	−0.51	−0.48	−0.29	−0.16	−0.53	−0.52
S5: M/T 1%	−0.44	0	0.27	−0.22	−0.13	−0.49	−0.44	**−0.51**	−0.5	−0.47	−0.4	−0.5	−0.44	−0.41	−0.39	−0.4	−0.36
3%	−0.05	0	0.34	−0.23	−0.16	−0.52	−0.52	−0.53	**−0.56**	−0.49	−0.41	−0.51	−0.48	−0.37	−0.37	−0.53	−0.54
7%	0.5	0	0.5	−0.24	−0.17	−0.5	−0.5	−0.51	**−0.6**	−0.46	−0.35	−0.49	−0.49	−0.26	−0.25	−0.52	−0.58
11%	0.91	0	0.61	−0.26	−0.2	−0.52	−0.49	−0.54	**−0.65**	−0.48	−0.32	−0.52	−0.54	−0.18	−0.16	−0.54	−0.63
15%	1.3	0	0.52	−0.25	−0.16	−0.5	−0.47	−0.52	**−0.67**	−0.46	−0.36	−0.5	−0.55	−0.09	−0.13	−0.52	−0.66
S1: F1S<=0.5 & F1D>0.5																	
S2: Good Vis.	3.97	0	0.5	−0.31	−0.31	−0.14	−0.18	−0.29	**−0.56**	0.14	0.25	−0.12	−0.18	−0.1	−0.07	−0.33	−0.54
Mod.	3.6	0	0.51	−0.3	−0.29	−0.13	−0.18	−0.29	**−0.55**	0.13	0.23	−0.14	−0.18	−0.1	−0.08	−0.32	−0.54
Poor	3.61	0	0.5	−0.27	−0.26	−0.15	−0.19	−0.25	**−0.54**	0.11	0.23	−0.09	−0.17	−0.1	−0.08	−0.28	−0.52
S3: Neg. Cor.	4.83	0	0.51	−0.36	−0.36	−0.08	−0.14	−0.31	**−0.59**	0.32	0.44	−0.06	−0.1	−0.09	−0.08	−0.37	**−0.59**
Pos.	8.57	0	0.5	−0.47	−0.4	0.02	0	−0.36	**−0.53**	0.96	1.17	0.13	0.03	−0.06	−0.05	−0.47	**−0.53**
None	5.32	0	0.49	−0.36	−0.36	−0.06	−0.11	−0.31	**−0.59**	0.38	0.49	−0.05	−0.09	−0.06	−0.05	−0.36	−0.58
S4: High Autocor.	5.05	0	0.5	−0.27	−0.31	−0.13	−0.37	−0.27	**−0.73**	0.18	0.79	−0.1	−0.32	−0.08	−0.31	−0.29	−0.72
Medium	3.97	0	0.5	−0.31	−0.31	−0.14	−0.18	−0.29	**−0.56**	0.14	0.25	−0.12	−0.18	−0.1	−0.07	−0.33	−0.54
Low	3.83	0	0.51	−0.32	−0.25	−0.12	−0.02	−0.3	**−0.45**	0.19	0.31	−0.13	−0.2	−0.11	0.11	−0.35	−0.41
S5: M/T 1%	0.97	0	0.27	**−0.3**	−0.23	−0.09	−0.04	−0.28	**−0.3**	0.21	0.33	−0.09	0.03	0.03	0.06	−0.26	−0.27
3%	2.48	0	0.34	−0.3	−0.26	−0.1	−0.11	−0.27	**−0.42**	0.24	0.32	−0.08	−0.04	−0.04	−0.03	−0.29	−0.41
7%	3.97	0	0.5	−0.31	−0.31	−0.14	−0.18	−0.29	**−0.56**	0.14	0.25	−0.12	−0.18	−0.1	−0.07	−0.33	−0.54
11%	5.87	0	0.59	−0.32	−0.32	−0.12	−0.21	−0.3	**−0.63**	0.19	0.36	−0.11	−0.23	−0.11	−0.1	−0.34	−0.62
15%	7.17	0	0.51	−0.34	−0.31	−0.13	−0.29	−0.32	**−0.67**	0.19	0.27	−0.15	−0.26	−0.12	−0.16	−0.36	−0.65
S1: F1S>0.5 & F1D<=0.5																	
S2: Good Vis.	1.33	0	0.51	0.68	0.86	−0.05	2	0.68	0.81	−0.12	0.26	0.66	0.85	**−0.18**	−0.03	0.63	0.85
Mod.	1.11	0	0.5	0.74	0.91	−0.1	−0.07	0.73	0.88	−0.14	0.2	0.69	0.87	**−0.18**	−0.04	0.73	0.92
Poor	1.13	0	0.53	0.77	0.94	−0.07	−0.04	0.73	0.9	−0.12	0.24	0.7	0.89	**−0.17**	−0.02	0.7	0.91
S3: Neg. Cor.	0.19	**0**	0.51	0.14	0.23	0.45	1.2	0.56	0.48	0.26	0.58	0.61	0.81	0.02	0.18	0.22	0.33
Pos.	0.7	**0**	0.52	**0**	0.09	0.49	0.36	0.41	0.37	0.46	0.59	0.66	0.63	0.03	0.22	0.05	0.19
None	0.19	**0**	0.52	0.2	0.29	0.45	0.46	0.59	0.51	0.29	0.63	0.67	0.88	0.02	0.18	0.28	0.39
S4: High Autocor.	1.44	0	0.52	0.73	0.72	0.03	0.92	0.72	0.6	−0.08	0.55	0.73	0.7	**−0.14**	−0.13	0.66	0.63
Medium	1.33	0	0.51	0.68	0.86	−0.05	2	0.68	0.81	−0.12	0.26	0.66	0.85	**−0.18**	−0.03	0.63	0.85
Low	1.75	0	0.51	0.68	0.92	0.1	0.58	0.7	0.97	−0.08	0.33	0.68	0.86	**−0.16**	0.13	0.6	0.93
S5: M/T 1%	−0.01	0	0.27	0.7	0.77	0.41	21.23	0.89	0.96	**−0.08**	0.13	0.72	0.73	−0.07	0.07	0.7	0.77
3%	0.68	0	0.33	0.67	0.85	0.12	46.59	0.71	0.89	−0.07	0.17	0.66	0.84	**−0.14**	−0.01	0.68	0.9
7%	1.33	0	0.51	0.68	0.86	−0.05	2	0.68	0.81	−0.12	0.26	0.66	0.85	**−0.18**	−0.03	0.63	0.85
11%	2.36	0	0.62	0.78	0.95	0.07	0.08	0.73	0.8	−0.08	0.44	0.76	0.93	**−0.16**	0.04	0.66	0.82
15%	2.88	0	0.49	0.67	0.83	0.06	0.68	0.67	0.7	−0.09	0.26	0.68	0.81	**−0.17**	−0.04	0.59	0.69
S1: F1S>0.5 & F1D>0.5																	
S2: Good Vis.	1.06	0	0.51	0.24	0.3	0.23	0.2	0.54	0.52	0.37	0.72	0.83	1.01	**−0.01**	0.18	0.25	0.37
Mod.	0.89	0	0.52	0.32	0.41	0.19	0.14	0.6	0.57	0.32	0.64	0.86	1.05	**−0.01**	0.14	0.36	0.47
Poor	0.88	0	0.51	0.29	0.38	0.18	0.13	0.6	0.58	0.29	0.63	0.89	1.11	**−0.01**	0.14	0.33	0.45
S3: Neg. Cor.	1.62	**0**	0.51	0.08	0.15	0.32	1.37	0.41	0.37	0.53	0.9	0.79	0.96	**0**	0.25	0.09	0.24
Pos.	5.35	0	0.52	**−0.06**	−0.01	0.34	0.33	0.19	0.24	1.08	1.26	0.88	0.87	0	0.3	**−0.06**	0.07
None	2.11	0	0.52	0.09	0.15	0.3	0.24	0.41	0.37	0.57	0.92	0.85	1	**−0.01**	0.26	0.09	0.24
S4: High Autocor.	1.84	0	0.52	0.21	0.1	0.18	2.02	0.47	0.24	0.37	1.39	0.8	0.62	−0.02	**−0.05**	0.22	0.06
Medium	1.06	0	0.51	0.24	0.3	0.23	0.2	0.54	0.52	0.37	0.72	0.83	1.01	**−0.01**	0.18	0.25	0.37
Low	0.78	**0**	0.53	0.15	0.3	0.32	0.64	0.45	0.74	0.43	0.84	0.74	0.84	**0**	0.44	0.15	0.45
S5: M/T 1%	**−0.19**	0	0.28	0.19	0.32	0.78	1.89	0.76	1.13	0.44	0.67	0.82	1.1	0.19	0.47	0.22	0.39
3%	0.38	**0**	0.32	0.16	0.24	0.35	0.43	0.46	0.55	0.42	0.65	0.71	0.92	0.03	0.18	0.22	0.44
7%	1.06	0	0.51	0.24	0.3	0.23	0.2	0.54	0.52	0.37	0.72	0.83	1.01	**−0.01**	0.18	0.25	0.37
11%	1.89	0	0.63	0.21	0.28	0.28	0.24	0.49	0.41	0.44	0.98	0.82	0.91	**−0.01**	0.24	0.22	0.33
15%	2.36	0	0.49	0.19	0.2	0.28	0.5	0.46	0.32	0.44	0.72	0.81	0.82	**−0.01**	0.14	0.19	0.26

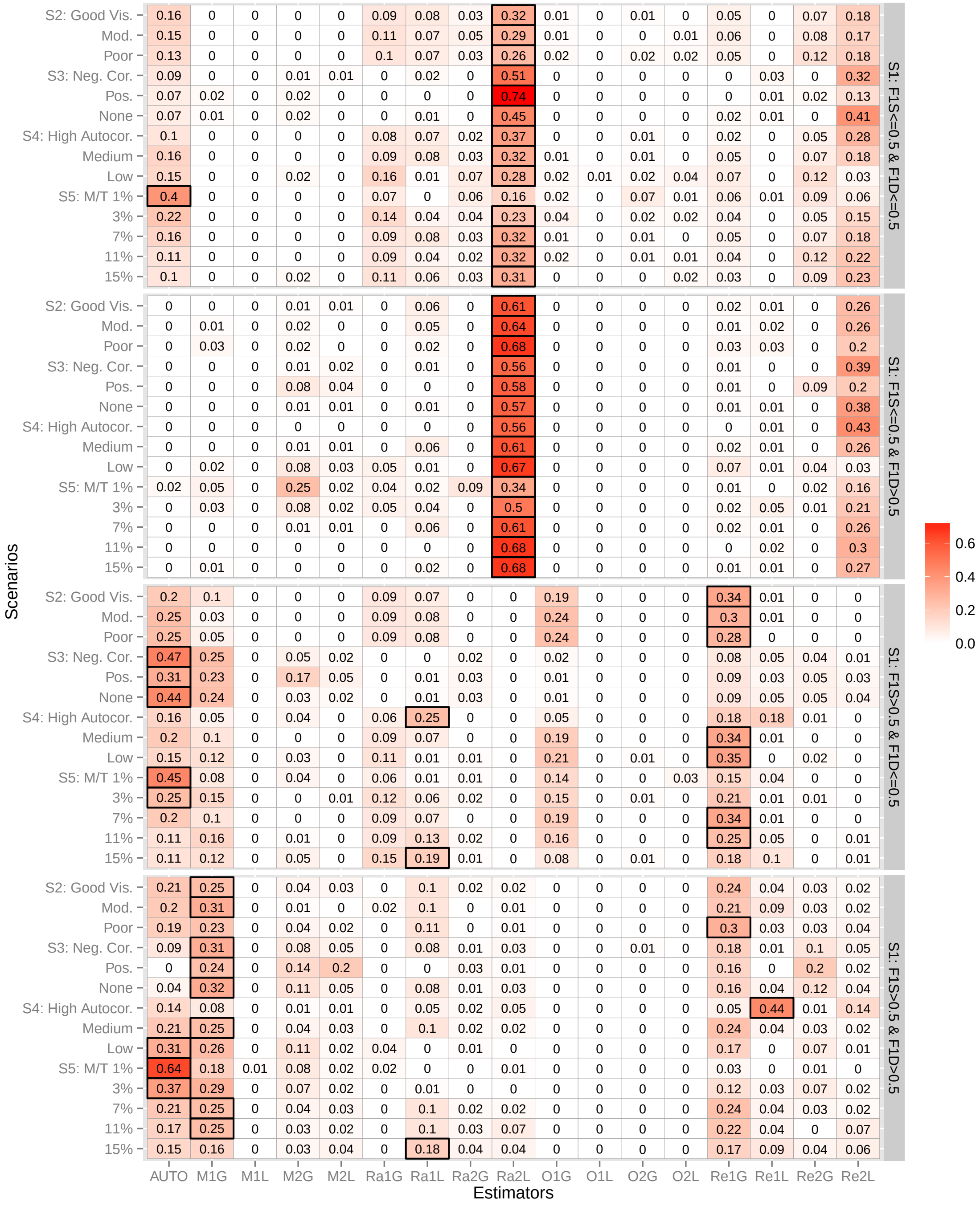
Table 3: Relative mean absolute error ($\mathrm{MAE}_{\mathrm{re}}$, using the global manual mean as the baseline method) for the five scenarios by type of estimators. See Table 2 for explanations of the notations.



	Scenarios	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S1: F1S<=0.5 & F1D<=0.5	S2: Good Vis.	0.27	0	0.22	−0.15	−0.11	−0.31	−0.31	−0.34	−0.42	−0.29	−0.21	−0.33	−0.34	−0.14	−0.13	−0.35	−0.41
	Mod.	0.28	0	0.23	−0.13	−0.1	−0.3	−0.29	−0.32	−0.4	−0.28	−0.19	−0.31	−0.32	−0.14	−0.12	−0.33	−0.39
	Poor	0.27	0	0.23	−0.13	−0.09	−0.3	−0.29	−0.32	−0.38	−0.27	−0.19	−0.3	−0.31	−0.14	−0.13	−0.32	−0.38
	S3: Neg. Cor.	0.71	0	0.22	−0.21	−0.18	−0.09	−0.11	−0.21	−0.39	0	0.06	−0.15	−0.15	−0.06	−0.08	−0.22	−0.38
	Pos.	1.03	0	0.23	−0.25	−0.21	−0.04	−0.09	−0.22	−0.37	0.11	0.12	−0.12	−0.12	−0.05	−0.05	−0.26	−0.35
	None	0.82	0	0.22	−0.2	−0.17	−0.08	−0.11	−0.19	−0.38	0.02	0.07	−0.13	−0.13	−0.06	−0.07	−0.21	−0.37
	S4: High Autocor.	0.33	0	0.22	−0.14	−0.15	−0.31	−0.35	−0.34	−0.51	−0.3	−0.1	−0.33	−0.4	−0.07	−0.13	−0.35	−0.5
	Medium	0.27	0	0.22	−0.15	−0.11	−0.31	−0.31	−0.34	−0.42	−0.29	−0.21	−0.33	−0.34	−0.14	−0.13	−0.35	−0.41
	Low	0.21	0	0.24	−0.15	−0.07	−0.32	−0.25	−0.36	−0.37	−0.29	−0.2	−0.34	−0.33	−0.16	−0.07	−0.36	−0.36
	S5: M/T 1%	−0.23	0	0.11	−0.13	−0.09	−0.31	−0.27	−0.34	−0.33	−0.29	−0.24	−0.33	−0.3	−0.24	−0.23	−0.28	−0.26
	3%	0.01	0	0.15	−0.14	−0.11	−0.33	−0.32	−0.35	−0.38	−0.3	−0.25	−0.33	−0.32	−0.21	−0.21	−0.35	−0.37
	7%	0.27	0	0.22	−0.15	−0.11	−0.31	−0.31	−0.34	−0.42	−0.29	−0.21	−0.33	−0.34	−0.14	−0.13	−0.35	−0.41
	11%	0.42	0	0.28	−0.18	−0.13	−0.33	−0.3	−0.37	−0.46	−0.3	−0.19	−0.36	−0.37	−0.09	−0.07	−0.37	−0.46
	15%	0.57	0	0.24	−0.17	−0.09	−0.31	−0.28	−0.36	−0.46	−0.29	−0.22	−0.35	−0.38	−0.04	−0.05	−0.36	−0.47
S1: F1S<=0.5 & F1D>0.5	S2: Good Vis.	1.46	0	0.22	−0.18	−0.18	−0.07	−0.1	−0.18	−0.38	0.07	0.12	−0.08	−0.12	−0.05	−0.04	−0.19	−0.36
	Mod.	1.37	0	0.22	−0.18	−0.17	−0.07	−0.1	−0.18	−0.37	0.06	0.11	−0.09	−0.12	−0.06	−0.05	−0.19	−0.36
	Poor	1.37	0	0.22	−0.16	−0.16	−0.08	−0.11	−0.16	−0.37	0.05	0.11	−0.07	−0.12	−0.06	−0.05	−0.18	−0.35
	S3: Neg. Cor.	1.67	0	0.23	−0.21	−0.21	−0.05	−0.08	−0.18	−0.4	0.15	0.2	−0.05	−0.08	−0.05	−0.04	−0.22	−0.39
	Pos.	2.42	0	0.22	−0.28	−0.24	0.01	0	−0.21	−0.34	0.39	0.45	0.04	−0.01	−0.04	−0.03	−0.28	−0.33
	None	1.8	0	0.22	−0.21	−0.21	−0.03	−0.07	−0.18	−0.4	0.17	0.22	−0.04	−0.07	−0.04	−0.03	−0.21	−0.39
	S4: High Autocor.	1.72	0	0.23	−0.16	−0.18	−0.07	−0.21	−0.17	−0.51	0.08	0.39	−0.07	−0.19	−0.04	−0.17	−0.18	−0.51
	Medium	1.46	0	0.22	−0.18	−0.18	−0.07	−0.1	−0.18	−0.38	0.07	0.12	−0.08	−0.12	−0.05	−0.04	−0.19	−0.36
	Low	1.41	0	0.23	−0.19	−0.14	−0.07	−0.01	−0.18	−0.3	0.08	0.15	−0.09	−0.13	−0.06	0.06	−0.21	−0.27
	S5: M/T 1%	0.56	0	0.11	−0.17	−0.13	−0.04	−0.02	−0.16	−0.19	0.1	0.15	−0.06	−0.01	0.01	0.03	−0.16	−0.17
	3%	1.06	0	0.14	−0.17	−0.16	−0.05	−0.06	−0.16	−0.28	0.11	0.15	−0.06	−0.05	−0.03	−0.02	−0.17	−0.28
	7%	1.46	0	0.22	−0.18	−0.18	−0.07	−0.1	−0.18	−0.38	0.07	0.12	−0.08	−0.12	−0.05	−0.04	−0.19	−0.36
	11%	1.89	0	0.27	−0.19	−0.18	−0.07	−0.11	−0.18	−0.43	0.09	0.17	−0.08	−0.14	−0.06	−0.05	−0.21	−0.41
	15%	2.17	0	0.25	−0.21	−0.17	−0.07	−0.15	−0.2	−0.45	0.09	0.13	−0.1	−0.15	−0.06	−0.08	−0.23	−0.44
S1: F1S>0.5 & F1D<=0.5	S2: Good Vis.	0.59	0	0.23	0.29	0.35	−0.04	0.03	0.27	0.31	−0.07	0.12	0.27	0.34	−0.1	−0.02	0.26	0.32
	Mod.	0.51	0	0.23	0.32	0.37	−0.06	−0.05	0.3	0.34	−0.07	0.1	0.3	0.35	−0.1	−0.03	0.3	0.35
	Poor	0.51	0	0.24	0.32	0.38	−0.05	−0.03	0.3	0.35	−0.06	0.11	0.29	0.36	−0.09	−0.01	0.29	0.36
	S3: Neg. Cor.	0.14	0	0.23	0.07	0.11	0.19	0.18	0.23	0.2	0.12	0.26	0.26	0.33	0.01	0.08	0.1	0.14
	Pos.	0.36	0	0.24	0	0.04	0.21	0.15	0.18	0.15	0.2	0.26	0.27	0.26	0.01	0.1	0.02	0.07
	None	0.12	0	0.23	0.09	0.13	0.19	0.14	0.24	0.21	0.13	0.28	0.28	0.36	0.01	0.08	0.11	0.15
	S4: High Autocor.	0.62	0	0.23	0.3	0.29	−0.02	0.03	0.29	0.22	−0.05	0.26	0.3	0.28	−0.08	−0.08	0.27	0.24
	Medium	0.59	0	0.23	0.29	0.35	−0.04	0.03	0.27	0.31	−0.07	0.12	0.27	0.34	−0.1	−0.02	0.26	0.32
	Low	0.76	0	0.23	0.28	0.36	0.02	0.13	0.28	0.36	−0.04	0.16	0.28	0.34	−0.09	0.06	0.25	0.34
	S5: M/T 1%	0.05	0	0.1	0.27	0.32	0.07	0.38	0.31	0.36	−0.04	0.05	0.28	0.29	−0.05	0	0.27	0.32
	3%	0.36	0	0.14	0.28	0.34	0.01	0.33	0.27	0.33	−0.04	0.07	0.27	0.33	−0.08	−0.03	0.28	0.34
	7%	0.59	0	0.23	0.29	0.35	−0.04	0.03	0.27	0.31	−0.07	0.12	0.27	0.34	−0.1	−0.02	0.26	0.32
	11%	0.94	0	0.3	0.32	0.38	0.01	0	0.3	0.32	−0.05	0.21	0.32	0.38	−0.09	0.02	0.27	0.32
	15%	1.07	0	0.23	0.29	0.34	−0.01	0.02	0.27	0.28	−0.05	0.13	0.28	0.33	−0.09	−0.02	0.25	0.27
S1: F1S>0.5 & F1D>0.5	S2: Good Vis.	0.52	0	0.23	0.11	0.14	0.1	0.08	0.23	0.22	0.17	0.32	0.35	0.41	−0.01	0.08	0.12	0.16
	Mod.	0.48	0	0.23	0.15	0.18	0.08	0.06	0.25	0.25	0.15	0.29	0.36	0.43	−0.01	0.07	0.16	0.2
	Poor	0.47	0	0.23	0.13	0.16	0.07	0.05	0.25	0.24	0.13	0.28	0.37	0.43	−0.01	0.06	0.15	0.18
	S3: Neg. Cor.	0.74	0	0.22	0.04	0.06	0.13	0.16	0.17	0.15	0.23	0.38	0.33	0.39	0	0.11	0.04	0.09
	Pos.	1.8	0	0.24	−0.03	−0.01	0.15	0.16	0.08	0.1	0.45	0.51	0.37	0.36	0	0.14	−0.03	0.02
	None	0.92	0	0.24	0.04	0.07	0.13	0.1	0.18	0.15	0.25	0.4	0.36	0.41	0	0.12	0.04	0.1
	S4: High Autocor.	0.79	0	0.24	0.1	0.05	0.08	0.23	0.2	0.1	0.17	0.61	0.34	0.27	−0.01	−0.03	0.1	0.01
	Medium	0.52	0	0.23	0.11	0.14	0.1	0.08	0.23	0.22	0.17	0.32	0.35	0.41	−0.01	0.08	0.12	0.16
	Low	0.39	0	0.24	0.07	0.14	0.14	0.24	0.2	0.3	0.19	0.37	0.32	0.36	0	0.2	0.07	0.19
	S5: M/T 1%	−0.06	0	0.11	0.08	0.13	0.21	0.35	0.26	0.35	0.19	0.29	0.33	0.42	0.05	0.15	0.09	0.16
	3%	0.22	0	0.14	0.07	0.11	0.13	0.15	0.18	0.22	0.19	0.28	0.3	0.37	0.01	0.07	0.09	0.16
	7%	0.52	0	0.23	0.11	0.14	0.1	0.08	0.23	0.22	0.17	0.32	0.35	0.41	−0.01	0.08	0.12	0.16
	11%	0.78	0	0.29	0.11	0.13	0.12	0.1	0.21	0.18	0.2	0.42	0.35	0.38	−0.01	0.12	0.1	0.14
	15%	0.92	0	0.23	0.09	0.09	0.12	0.11	0.2	0.14	0.2	0.32	0.34	0.34	−0.01	0.06	0.08	0.11

Estimators

Table 4: Proportion of runs with the least mean square error (MSE) for the five scenarios by type of estimators. See Table 2 for explanations of the notations.



S1: F1S<=0.5 & F1D<=0.5

Scenarios	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0.16	0	0	0	0	0.09	0.08	0.03	0.32	0.01	0	0.01	0	0.05	0	0.07	0.18
Mod.	0.15	0	0	0	0	0.11	0.07	0.05	0.29	0.01	0	0	0.01	0.06	0	0.08	0.17
Poor	0.13	0	0	0	0	0.1	0.07	0.03	0.26	0.02	0	0.02	0.02	0.05	0	0.12	0.18
S3: Neg. Cor.	0.09	0	0	0.01	0.01	0	0.02	0	0.51	0	0	0	0	0	0.03	0	0.32
Pos.	0.07	0.02	0	0.02	0	0	0	0	0.74	0	0	0	0	0	0.01	0.02	0.13
None	0.07	0.01	0	0.02	0	0	0.01	0	0.45	0	0	0	0	0.02	0.01	0	0.41
S4: High Autocor.	0.1	0	0	0	0	0.08	0.07	0.02	0.37	0	0	0.01	0	0.02	0	0.05	0.28
Medium	0.16	0	0	0	0	0.09	0.08	0.03	0.32	0.01	0	0.01	0	0.05	0	0.07	0.18
Low	0.15	0	0	0.02	0	0.16	0.01	0.07	0.28	0.02	0.01	0.02	0.04	0.07	0	0.12	0.03
S5: M/T 1%	0.4	0	0	0	0	0.07	0	0.06	0.16	0.02	0	0.07	0.01	0.06	0.01	0.09	0.06
3%	0.22	0	0	0	0	0.14	0.04	0.04	0.23	0.04	0	0.02	0.02	0.04	0	0.05	0.15
7%	0.16	0	0	0	0	0.09	0.08	0.03	0.32	0.01	0	0.01	0	0.05	0	0.07	0.18
11%	0.11	0	0	0	0	0.09	0.04	0.02	0.32	0.02	0	0.01	0.01	0.04	0	0.12	0.22
15%	0.1	0	0	0.02	0	0.11	0.06	0.03	0.31	0	0	0	0.02	0.03	0	0.09	0.23

S1: F1S<=0.5 & F1D>0.5

Scenarios	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0	0	0	0.01	0.01	0	0.06	0	0.61	0	0	0	0	0.02	0.01	0	0.26
Mod.	0	0.01	0	0.02	0	0	0.05	0	0.64	0	0	0	0	0.01	0.02	0	0.26
Poor	0	0.03	0	0.02	0	0	0.02	0	0.68	0	0	0	0	0.03	0.03	0	0.2
S3: Neg. Cor.	0	0	0	0.01	0.02	0	0.01	0	0.56	0	0	0	0	0.01	0	0	0.39
Pos.	0	0	0	0.08	0.04	0	0	0	0.58	0	0	0	0	0.01	0	0.09	0.2
None	0	0	0	0.01	0.01	0	0.01	0	0.57	0	0	0	0	0.01	0.01	0	0.38
S4: High Autocor.	0	0	0	0	0	0	0	0	0.56	0	0	0	0	0	0.01	0	0.43
Medium	0	0	0	0.01	0.01	0	0.06	0	0.61	0	0	0	0	0.02	0.01	0	0.26
Low	0	0.02	0	0.08	0.03	0.05	0.01	0	0.67	0	0	0	0	0.07	0.01	0.04	0.03
S5: M/T 1%	0.02	0.05	0	0.25	0.02	0.04	0.02	0.09	0.34	0	0	0	0	0.01	0	0.02	0.16
3%	0	0.03	0	0.08	0.02	0.05	0.04	0	0.5	0	0	0	0	0.02	0.05	0.01	0.21
7%	0	0	0	0.01	0.01	0	0.06	0	0.61	0	0	0	0	0.02	0.01	0	0.26
11%	0	0	0	0	0	0	0	0	0.68	0	0	0	0	0	0.02	0	0.3
15%	0	0.01	0	0	0	0	0.02	0	0.68	0	0	0	0	0.01	0.01	0	0.27

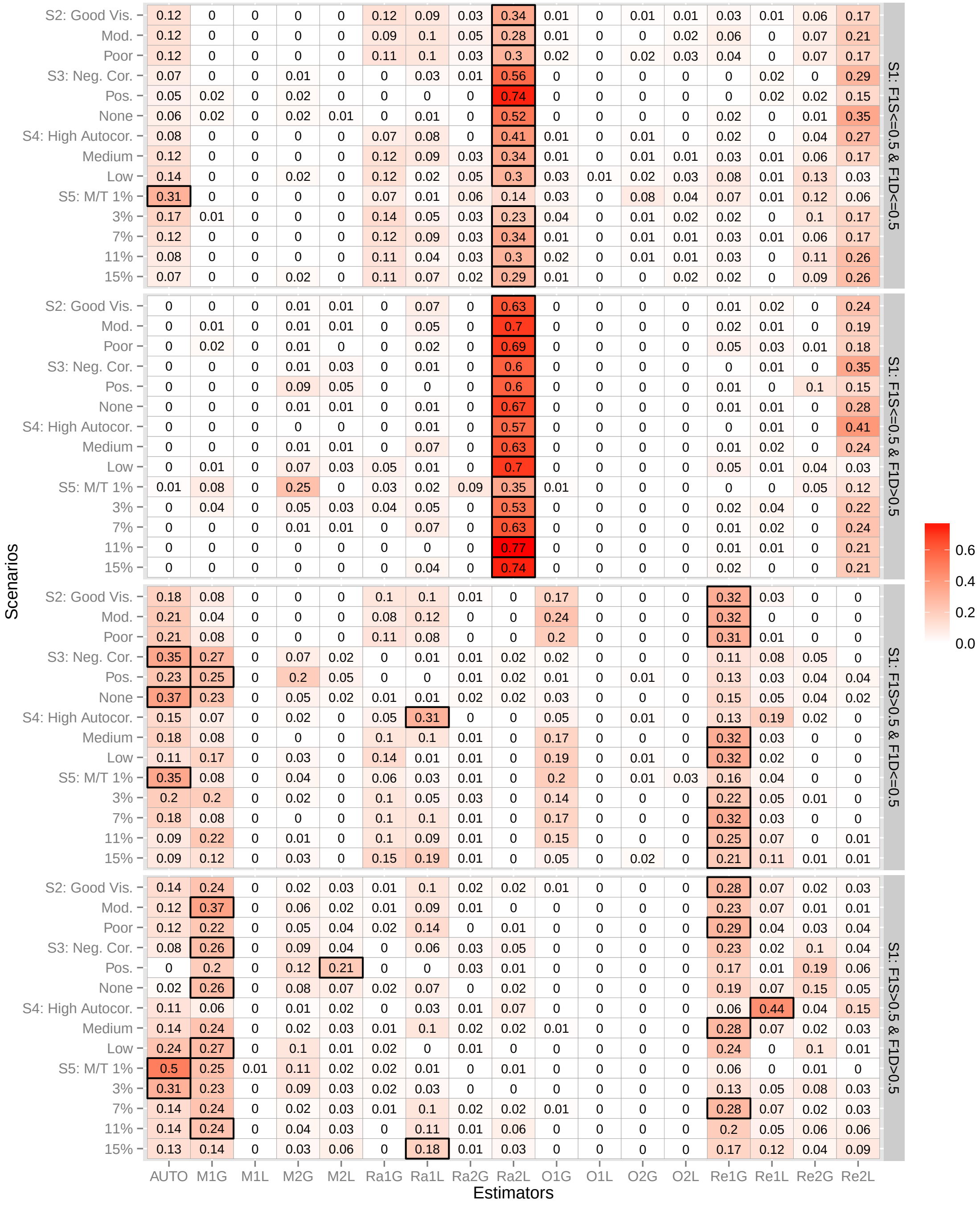
S1: F1S>0.5 & F1D<=0.5

Scenarios	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0.2	0.1	0	0	0	0.09	0.07	0	0	0.19	0	0	0	0.34	0.01	0	0
Mod.	0.25	0.03	0	0	0	0.09	0.08	0	0	0.24	0	0	0	0.3	0.01	0	0
Poor	0.25	0.05	0	0	0	0.09	0.08	0	0	0.24	0	0	0	0.28	0	0	0
S3: Neg. Cor.	0.47	0.25	0	0.05	0.02	0	0	0.02	0	0.02	0	0	0	0.08	0.05	0.04	0.01
Pos.	0.31	0.23	0	0.17	0.05	0	0.01	0.03	0	0.01	0	0	0	0.09	0.03	0.05	0.03
None	0.44	0.24	0	0.03	0.02	0	0.01	0.03	0	0.01	0	0	0	0.09	0.05	0.05	0.04
S4: High Autocor.	0.16	0.05	0	0.04	0	0.06	0.25	0	0	0.05	0	0	0	0.18	0.18	0.01	0
Medium	0.2	0.1	0	0	0	0.09	0.07	0	0	0.19	0	0	0	0.34	0.01	0	0
Low	0.15	0.12	0	0.03	0	0.11	0.01	0.01	0	0.21	0	0.01	0	0.35	0	0.02	0
S5: M/T 1%	0.45	0.08	0	0.04	0	0.06	0.01	0.01	0	0.14	0	0	0.03	0.15	0.04	0	0
3%	0.25	0.15	0	0	0.01	0.12	0.06	0.02	0	0.15	0	0.01	0	0.21	0.01	0.01	0
7%	0.2	0.1	0	0	0	0.09	0.07	0	0	0.19	0	0	0	0.34	0.01	0	0
11%	0.11	0.16	0	0.01	0	0.09	0.13	0.02	0	0.16	0	0	0	0.25	0.05	0	0.01
15%	0.11	0.12	0	0.05	0	0.15	0.19	0.01	0	0.08	0	0.01	0	0.18	0.1	0	0.01

S1: F1S>0.5 & F1D>0.5

Scenarios	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0.21	0.25	0	0.04	0.03	0	0.1	0.02	0.02	0	0	0	0	0.24	0.04	0.03	0.02
Mod.	0.2	0.31	0	0.01	0	0.02	0.1	0	0.01	0	0	0	0	0.21	0.09	0.03	0.02
Poor	0.19	0.23	0	0.04	0.02	0	0.11	0	0.01	0	0	0	0	0.3	0.03	0.03	0.04
S3: Neg. Cor.	0.09	0.31	0	0.08	0.05	0	0.08	0.01	0.03	0	0	0.01	0	0.18	0.01	0.1	0.05
Pos.	0	0.24	0	0.14	0.2	0	0	0.03	0.01	0	0	0	0	0.16	0	0.2	0.02
None	0.04	0.32	0	0.11	0.05	0	0.08	0.01	0.03	0	0	0	0	0.16	0.04	0.12	0.04
S4: High Autocor.	0.14	0.08	0	0.01	0.01	0	0.05	0.02	0.05	0	0	0	0	0.05	0.44	0.01	0.14
Medium	0.21	0.25	0	0.04	0.03	0	0.1	0.02	0.02	0	0	0	0	0.24	0.04	0.03	0.02
Low	0.31	0.26	0	0.11	0.02	0.04	0	0.01	0	0	0	0	0	0.17	0	0.07	0.01
S5: M/T 1%	0.64	0.18	0.01	0.08	0.02	0.02	0	0	0.01	0	0	0	0	0.03	0	0.01	0
3%	0.37	0.29	0	0.07	0.02	0	0.01	0	0	0	0	0	0	0.12	0.03	0.07	0.02
7%	0.21	0.25	0	0.04	0.03	0	0.1	0.02	0.02	0	0	0	0	0.24	0.04	0.03	0.02
11%	0.17	0.25	0	0.03	0.02	0	0.1	0.03	0.07	0	0	0	0	0.22	0.04	0	0.07
15%	0.15	0.16	0	0.03	0.04	0	0.18	0.04	0.04	0	0	0	0	0.17	0.09	0.04	0.06

Table 5: Proportion of runs with the least mean absolute error (MSE) for the five scenarios by type of estimators. See Table 2 for explanations of the notations.



S1: F1S<=0.5 & F1D<=0.5

	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0.12	0	0	0	0	0.12	0.09	0.03	0.34	0.01	0	0.01	0.01	0.03	0.01	0.06	0.17
Mod.	0.12	0	0	0	0	0.09	0.1	0.05	0.28	0.01	0	0	0.02	0.06	0	0.07	0.21
Poor	0.12	0	0	0	0	0.11	0.1	0.03	0.3	0.02	0	0.02	0.03	0.04	0	0.07	0.17
S3: Neg. Cor.	0.07	0	0	0.01	0	0	0.03	0.01	0.56	0	0	0	0	0	0.02	0	0.29
Pos.	0.05	0.02	0	0.02	0	0	0	0	0.74	0	0	0	0	0	0.02	0.02	0.15
None	0.06	0.02	0	0.02	0.01	0	0.01	0	0.52	0	0	0	0	0.02	0	0.01	0.35
S4: High Autocor.	0.08	0	0	0	0	0.07	0.08	0	0.41	0.01	0	0.01	0	0.02	0	0.04	0.27
Medium	0.12	0	0	0	0	0.12	0.09	0.03	0.34	0.01	0	0.01	0.01	0.03	0.01	0.06	0.17
Low	0.14	0	0	0.02	0	0.12	0.02	0.05	0.3	0.03	0.01	0.02	0.03	0.08	0.01	0.13	0.03
S5: M/T 1%	0.31	0	0	0	0	0.07	0.01	0.06	0.14	0.03	0	0.08	0.04	0.07	0.01	0.12	0.06
3%	0.17	0.01	0	0	0	0.14	0.05	0.03	0.23	0.04	0	0.01	0.02	0.02	0	0.1	0.17
7%	0.12	0	0	0	0	0.12	0.09	0.03	0.34	0.01	0	0.01	0.01	0.03	0.01	0.06	0.17
11%	0.08	0	0	0	0	0.11	0.04	0.03	0.3	0.02	0	0.01	0.01	0.03	0	0.11	0.26
15%	0.07	0	0	0.02	0	0.11	0.07	0.02	0.29	0.01	0	0	0.02	0.02	0	0.09	0.26

S1: F1S<=0.5 & F1D>0.5

	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0	0	0	0.01	0.01	0	0.07	0	0.63	0	0	0	0	0.01	0.02	0	0.24
Mod.	0	0.01	0	0.01	0.01	0	0.05	0	0.7	0	0	0	0	0.02	0.01	0	0.19
Poor	0	0.02	0	0.01	0	0	0.02	0	0.69	0	0	0	0	0.05	0.03	0.01	0.18
S3: Neg. Cor.	0	0	0	0.01	0.03	0	0.01	0	0.6	0	0	0	0	0	0.01	0	0.35
Pos.	0	0	0	0.09	0.05	0	0	0	0.6	0	0	0	0	0.01	0	0.1	0.15
None	0	0	0	0.01	0.01	0	0.01	0	0.67	0	0	0	0	0.01	0.01	0	0.28
S4: High Autocor.	0	0	0	0	0	0	0.01	0	0.57	0	0	0	0	0	0.01	0	0.41
Medium	0	0	0	0.01	0.01	0	0.07	0	0.63	0	0	0	0	0.01	0.02	0	0.24
Low	0	0.01	0	0.07	0.03	0.05	0.01	0	0.7	0	0	0	0	0.05	0.01	0.04	0.03
S5: M/T 1%	0.01	0.08	0	0.25	0	0.03	0.02	0.09	0.35	0.01	0	0	0	0	0	0.05	0.12
3%	0	0.04	0	0.05	0.03	0.04	0.05	0	0.53	0	0	0	0	0.02	0.04	0	0.22
7%	0	0	0	0.01	0.01	0	0.07	0	0.63	0	0	0	0	0.01	0.02	0	0.24
11%	0	0	0	0	0	0	0	0	0.77	0	0	0	0	0.01	0.01	0	0.21
15%	0	0	0	0	0	0	0.04	0	0.74	0	0	0	0	0.02	0	0	0.21

S1: F1S>0.5 & F1D<=0.5

	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0.18	0.08	0	0	0	0.1	0.1	0.01	0	0.17	0	0	0	0.32	0.03	0	0
Mod.	0.21	0.04	0	0	0	0.08	0.12	0	0	0.24	0	0	0	0.32	0	0	0
Poor	0.21	0.08	0	0	0	0.11	0.08	0	0	0.2	0	0	0	0.31	0.01	0	0
S3: Neg. Cor.	0.35	0.27	0	0.07	0.02	0	0.01	0.01	0.02	0.02	0	0	0	0.11	0.08	0.05	0
Pos.	0.23	0.25	0	0.2	0.05	0	0	0.01	0.02	0.01	0	0.01	0	0.13	0.03	0.04	0.04
None	0.37	0.23	0	0.05	0.02	0.01	0.01	0.02	0.02	0.03	0	0	0	0.15	0.05	0.04	0.02
S4: High Autocor.	0.15	0.07	0	0.02	0	0.05	0.31	0	0	0.05	0	0.01	0	0.13	0.19	0.02	0
Medium	0.18	0.08	0	0	0	0.1	0.1	0.01	0	0.17	0	0	0	0.32	0.03	0	0
Low	0.11	0.17	0	0.03	0	0.14	0.01	0.01	0	0.19	0	0.01	0	0.32	0.02	0	0
S5: M/T 1%	0.35	0.08	0	0.04	0	0.06	0.03	0.01	0	0.2	0	0.01	0.03	0.16	0.04	0	0
3%	0.2	0.2	0	0.02	0	0.1	0.05	0.03	0	0.14	0	0	0	0.22	0.05	0.01	0
7%	0.18	0.08	0	0	0	0.1	0.1	0.01	0	0.17	0	0	0	0.32	0.03	0	0
11%	0.09	0.22	0	0.01	0	0.1	0.09	0.01	0	0.15	0	0	0	0.25	0.07	0	0.01
15%	0.09	0.12	0	0.03	0	0.15	0.19	0.01	0	0.05	0	0.02	0	0.21	0.11	0.01	0.01

S1: F1S>0.5 & F1D>0.5

	AUTO	M1G	M1L	M2G	M2L	Ra1G	Ra1L	Ra2G	Ra2L	O1G	O1L	O2G	O2L	Re1G	Re1L	Re2G	Re2L
S2: Good Vis.	0.14	0.24	0	0.02	0.03	0.01	0.1	0.02	0.02	0.01	0	0	0	0.28	0.07	0.02	0.03
Mod.	0.12	0.37	0	0.06	0.02	0.01	0.09	0.01	0	0	0	0	0	0.23	0.07	0.01	0.01
Poor	0.12	0.22	0	0.05	0.04	0.02	0.14	0	0.01	0	0	0	0	0.29	0.04	0.03	0.04
S3: Neg. Cor.	0.08	0.26	0	0.09	0.04	0	0.06	0.03	0.05	0	0	0	0	0.23	0.02	0.1	0.04
Pos.	0	0.2	0	0.12	0.21	0	0	0.03	0.01	0	0	0	0	0.17	0.01	0.19	0.06
None	0.02	0.26	0	0.08	0.07	0.02	0.07	0	0.02	0	0	0	0	0.19	0.07	0.15	0.05
S4: High Autocor.	0.11	0.06	0	0.01	0.02	0	0.03	0.01	0.07	0	0	0	0	0.06	0.44	0.04	0.15
Medium	0.14	0.24	0	0.02	0.03	0.01	0.1	0.02	0.02	0.01	0	0	0	0.28	0.07	0.02	0.03
Low	0.24	0.27	0	0.1	0.01	0.02	0	0.01	0	0	0	0	0	0.24	0	0.1	0.01
S5: M/T 1%	0.5	0.25	0.01	0.11	0.02	0.02	0.01	0	0.01	0	0	0	0	0.06	0	0.01	0
3%	0.31	0.23	0	0.09	0.03	0.02	0.03	0	0	0	0	0	0	0.13	0.05	0.08	0.03
7%	0.14	0.24	0	0.02	0.03	0.01	0.1	0.02	0.02	0.01	0	0	0	0.28	0.07	0.02	0.03
11%	0.14	0.24	0	0.04	0.03	0	0.11	0.01	0.06	0	0	0	0	0.2	0.05	0.06	0.06
15%	0.13	0.14	0	0.03	0.06	0	0.18	0.01	0.03	0	0	0	0	0.17	0.12	0.04	0.09

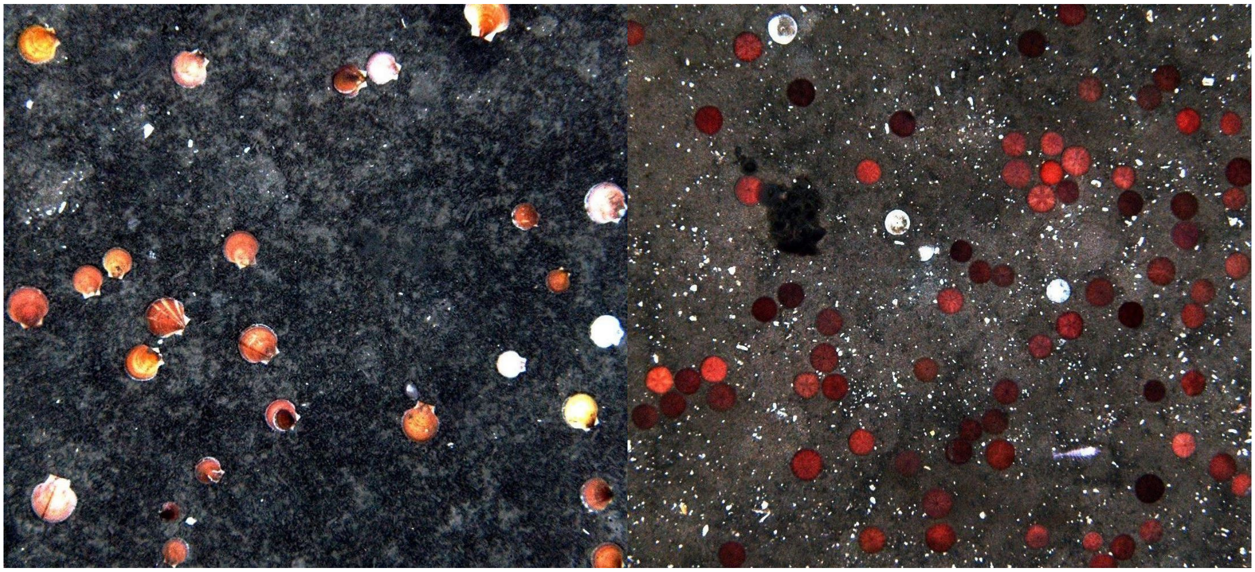


Figure 1: HabCam Images with scallops (left) and its common distractor sand dollars (right).

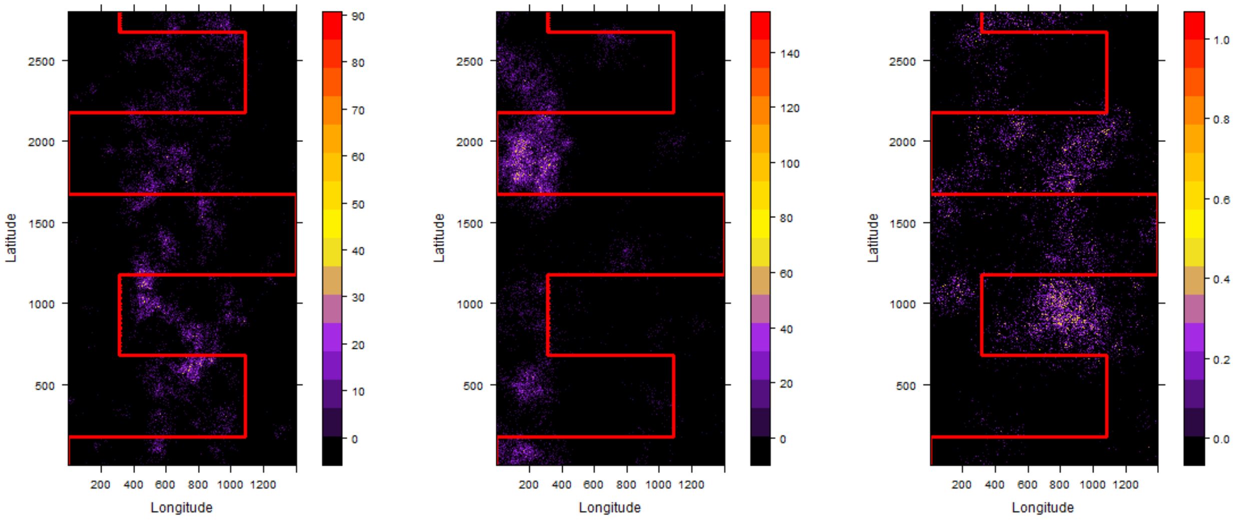


Figure 2: Example simulated distributions of scallops (left), distractors (center; moderate autocorrelation and negatively correlated with scallop distribution), and water visibility (right; poor) with an over-layed sampling track (red line). The colors represent counts per m^2 for scallops and distractors and the reduced probabilities of detecting scallops and distractors due to poor water visibility.

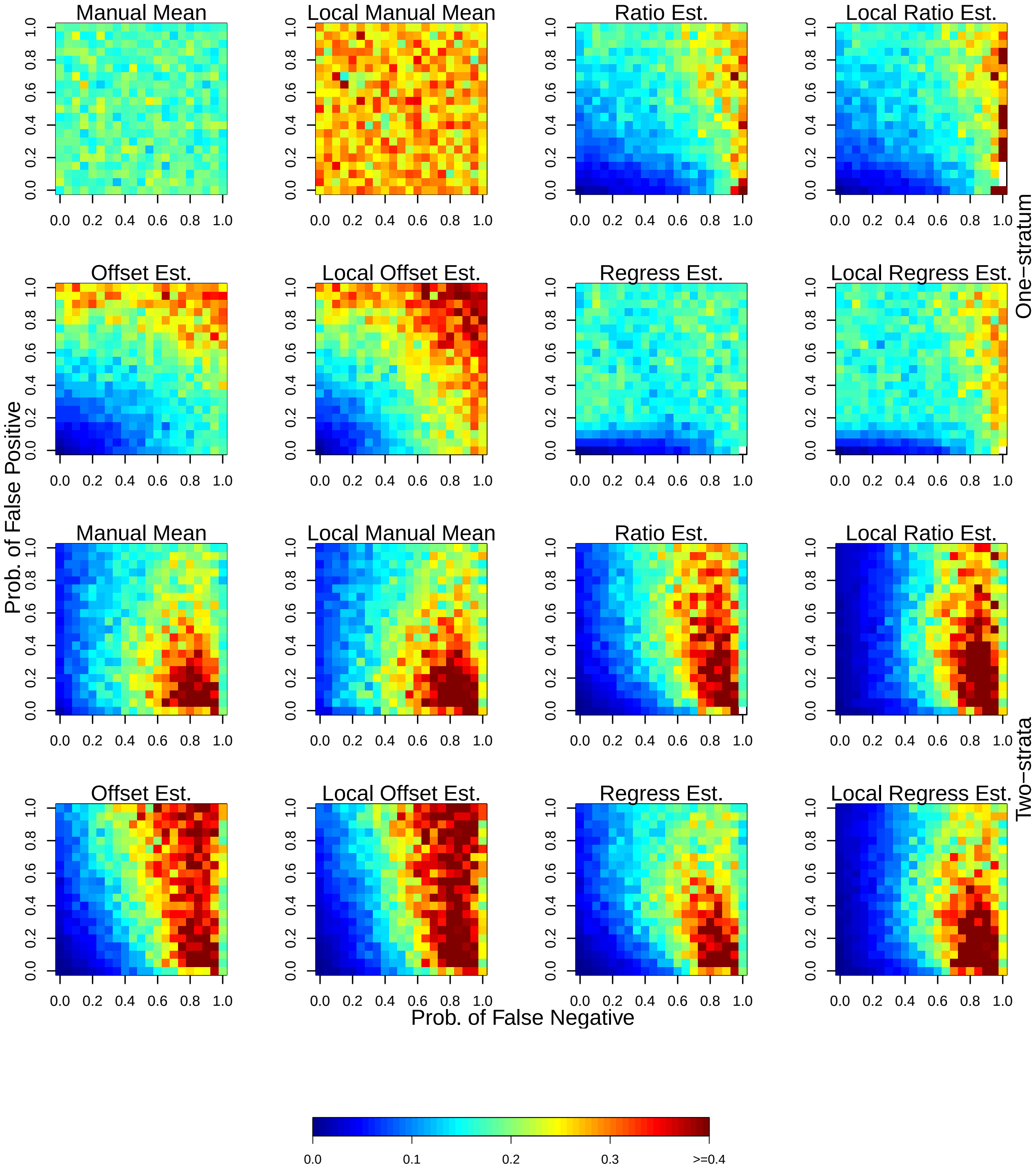


Figure 3: Mean squared error (MSE, indicated by color) at various false negative and false positive rates in the base case scenario, by estimator type, global or local estimation, and unstratified (one-statum) or two-strata estimation.

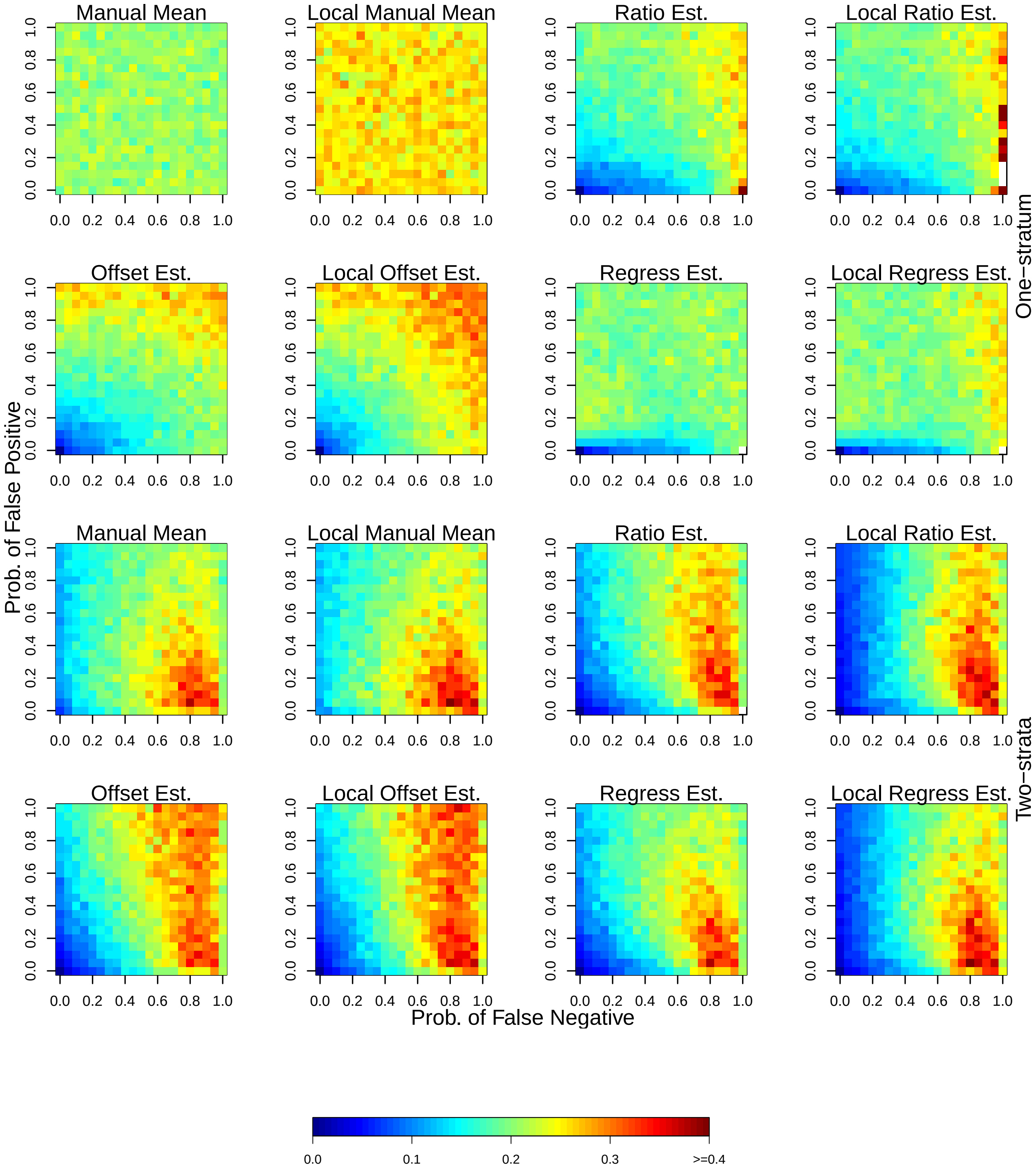


Figure 4: Mean absolute error (MAE, indicated by color) at various false negative and false positive rates in the base case scenario, by estimator type, global or local estimation, and unstratified (one-statum) or two-strata estimation.

Appendix - Analytic derivation of properties of the ratio estimator

Let Y_i be the number of targets in the ith randomly chosen image; it will be assumed that manual processing is perfect, so that Y_i is also the number of targets that were detected manually. Let X_i be the number of targets detected by the automated software in the ith image. We will consider the following ratio estimator for the mean number of targets:

$$T = \mu_X \frac{Y_1 + Y_2 + ... + Y_n}{X_1 + X_2 + ... + X_n} = \mu_X \frac{\bar{Y}}{\bar{X}} \tag{15}$$

where μ_X is the mean of the automated counts over all photographs, and $\bar{X}$ and $\bar{Y}$ are the sample means for the automated and manual counts for a randomly chosen sample of n images. Let $\mu_X = E(X_i)$ and $\mu_Y = E(Y_i)$, σ_X and σ_Y be the standard deviations of X_i and Y_i, respectively, and let ρ be the correlation between X_i and Y_i. Assuming for simplicity that the finite population correction factor is negligible (i.e., that the total number of images is large relative to n; this does not affect the main results below), using the approximate variance for a ratio (Cochran, 1977),

$$\begin{aligned} \mathrm{Var}(T) &= \mu_X^2 \mathrm{Var}\frac{\bar{Y}}{\bar{X}} \simeq \mu_X^2 \frac{1}{\mu_X^2}[\sigma_Y^2 + \frac{\sigma_X^2 \mu_Y^2}{\mu_X^2} - 2\rho\sigma_X\sigma_Y \frac{\mu_Y}{\mu_X}]/n & (16) \\ &= [\sigma_Y^2 + \sigma_X \frac{\mu_Y}{\mu_X}(\sigma_X \frac{\mu_Y}{\mu_X} - 2\rho\sigma_Y)]/n. & (17) \end{aligned}$$

Hence, $\mathrm{Var}(T)$ decreases linearly with ρ. If $\mu_X = \mu_Y$ and $\sigma_X = \sigma_Y$, this reduces to $\mathrm{Var}(T) \simeq 2\sigma_Y^2(1-\rho)/n$.

By comparison, a simple random sample of n manual images has variance $\mathrm{Var}(\bar{Y}) = \sigma_Y^2/n$, which is the first term of equation (17). Thus, the ratio estimator T has lower variance than simply using the manual images (i.e., $\mathrm{Var}(T) < \mathrm{Var}(\bar{Y})$) if and only if $\sigma_X \frac{\mu_Y}{\mu_X} - 2\rho\sigma_Y < 0$, i.e.,

$$\rho > \frac{\sigma_X \mu_Y}{2\sigma_Y \mu_X}. \tag{18}$$

In particular, if the X_is and Y_is have the same means and variances, then the ratio estimator is an improvement over simple random sampling of the manual images if and only if $\rho > 1/2$.