EBLUP Small Area Estimation for Red Snapper Bycatch from the Gulf of Mexico Shrimp Fleet

Beatrix Jones

March 12, 2004

1 Introduction

There are many difficulties in estimating bycatch taken by the Gulf of Mexico shrimp fleet. Foremost among these is a paucity of data. Nichols (2004) outlines the many bycatch data collection programs that have occurred since 1972. This paper is primarily concerned with the effort to augment data from the various bycatch observer programs with data from the research surveys carried out by the vessel Oregon II. The data from observer programs will be referred to as the "bycatch" data, and the data from the Oregon II will be referred to as the "abundance" data. These two data sets were previously integrated by using both types of data to fit a GLM:

$$\log(CPUE{+}1)_{ijklmn} = mean{+}dataset_i{+}year_j{+}season_k{+}area_l{+}depth_m{+}\epsilon_{ijklmn}.$$

(This approach was introduced in Nichols *et al.* 1987, used in Nichols *et al.* 1990, Nichols and Pellegrin 1992, and modified to incorporate a delta distribution in Ortiz *et al* 2000.) In a 1997 peer review (see MRAG Americas Inc., 1997) Mark Kaiser suggested small area estimation (specifically the EBLUP procedure) as an alternative. The "small areas" we wish to produce estimates for are cells representing combinations of geographic area, depth, season, and year. This approach releases us from the restrictions of the additive GLM structure; instead, the mean from each cell is estimated primarily from the bycatch and abundance data collected within that cell. The across cell data is combined only to determine the relationship between the bycatch and abundance measurements. The model relating these measurements is also more flexible; a two parameter linear model is used rather than just a mean difference.

2 Model Overview

EBLUP estimation is an empirical version of the best linear unbiased prediction method. The estimates produced are a compromise between raw estimates and predictions from a statistical model. For bycatch estimation, this model is a simple linear model relating bycatch to abundance measurements. Both bycatch and abundance are transformed to the scale log(CPUE+1). It is then assumed the abundance for a particular cell x_i and the true bycatch rate y_i are related by a regression model:

$$y_i = \beta x_i + v_i,\ v_i \sim N(0, \sigma_v^2)$$

y_i is then measured with error $\epsilon_i \sim N(0, \psi_i)$. This is a random effects model with a random effect v_i for each cell. These (independent) random effects are intended to capture differences from cell to cell that are independent of abundance, perhaps as a result of different fishing practices or weather. The best linear unbiased property of EBLUP only requires that the random effects and ϵ_i's have a symmetric distributions, with a common variance for the random effects. However, standard error estimates require the third and larger moments of the distribution to be zero for both these quantities.

Because the measurement errors ϵ_i seem to be large (as indicated by some extreme observations), shrinking the estimate toward the value predicted via the regression model using the measured abundance for that cell may improve the estimates over a simple cell average $\hat{\theta}_i$. We then use the estimate

$$\tilde{\theta} = \gamma\hat{\theta} + (1-\gamma)x_i^T\tilde{\beta}$$

where

$$\tilde{\beta} = \left[\sum_i x_i x_i^T/(\sigma_v^2 + \psi_i)\right]^{-1} \left[\sum_i x_i\hat{\theta}_i/(\sigma_v^2 + \psi_i)\right]$$

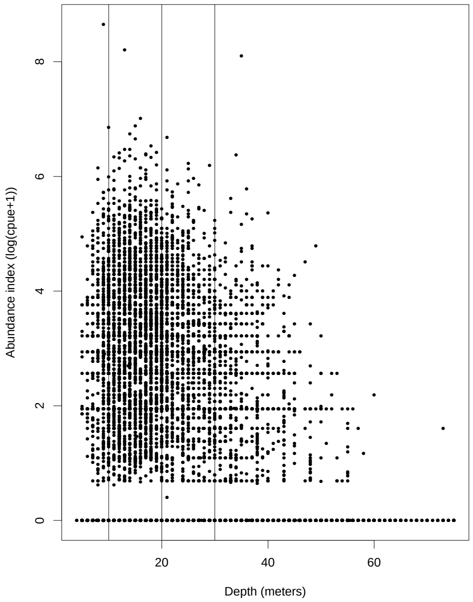
$$\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$$

Throughout we follow the approach outlined in Section 5.1 of Ghosh and Rao (1994).

3 Implementation Details

The observer data sets used were: all data available from 1972-1982, and the Characterization data and control net Evaluation data from both the Regional Research Program (1992-1997) and the summer 1998 BRD evaluations. The data are the same as those used in the GLM 'ALL' updates in the last red snapper assessment. The abundance data used were taken from survey measurements for the same time period (1972-1998). Depth was divided into 4 zones, as seen in Figure 1. These zones were chosen so that each zone had sufficient observations, but also to capture the trends of abundance with depth. Other factors were maintained as they were in the previous GLM model. Binning by these factors resulted in 260 cells with bycatch estimates. The abundance measurement for each cell was taken to be the average of $\log(r + 1)$, where the r's are the raw measurements for that cell in the oregon1 dataset. We refer to this cell average as the cell's "abundance index."

Figure 1: Abundance index vs depth, with the depth categories used in the analysis marked.



To determine the ψ_i, we assumed each observation had a common error variance σ_e^2. The final "measurement" $(\hat{\theta}_i)$ for a cell was taken to be the mean of all J_i observations in that cell; thus $\psi_i = \sigma_e^2/J_i$. We will call the individual observations the y_{ij}'s. σ_e^2 was estimated from all cells with observations, regardless of whether they had abundance estimates available. This included 4129 observations in 260 cells, so the estimate was $1/3869 \sum_{i,j}(y_{ij} - \hat{\theta}_i)^2$.

To estimate σ_v^2 we use the moment estimator given in equation (5.6) of Ghosh and Rao (1994):

$$\hat{\sigma_v^2} = 1/(t-p)\left[\sum_i (y_i - x_i^T\beta^*)^2 - \sum_i \psi_i \left\{1 - x_i^T(\sum_i x_i x_i^T)^{-1}x_i\right\}\right]$$

where β^* is the ordinary least squares estimate of β, t is the number of cells with abundance estimates (in our case 156), and p is the length of β, in our case 2 (the slope and the intercept). Under normality of both the random effects and measurement errors, this estimate of σ_v^2 has approximate variance:

$$\bar{V}(\hat{\sigma_v^2}) = 2/t^2 \sum_i (\sigma_v^2 + \psi_i)^2.$$

This gives mean squared error for $\tilde{\theta}_i$ of $g_{1i} + g_{2i} + 2g_{3i}$, where

$$\begin{aligned} g_{1i} &= \gamma_i\psi_i \\ g_{2i} &= (1-\gamma_i)^2 x_i^T \left[x_i x_i^T x_i^T/(\sigma_v^2 + \psi)\right] x_i \\ g_{3i} &= \psi_i^2 \bar{V}(\sigma_v^2)/(\sigma_v^2 + \psi_i)^3. \end{aligned}$$

4 Results

Of the 260 cells with observer data, only 156 also have abundance data. Figure 2a shows the abundance index vs the raw estimate $\hat{\theta}$. Figure 2b shows abundance vs. the EBLUP estimates $\tilde{\theta}$, with error bars representing the square root of the mean squared error. In Figure 3, $\tilde{\theta}$ is plotted against $\hat{\theta}$. Figures 2-3 show that many estimates are essentially unchanged, but a few extreme points have been reigned in by the EBLUP process. Some of the zero values have also been raised, indicating that while zero bycatch was observed, positive abundance measurements indicate a potential for some bycatch. Our estimate of the measurement error variance σ_e^2 is 0.56 , and is 0.78 for σ_v^2.

Figure 2: Abundance index vs bycatch (log(CPUE+1)): (A) direct bycatch estimates, (B) EBLUP bycatch estimates.

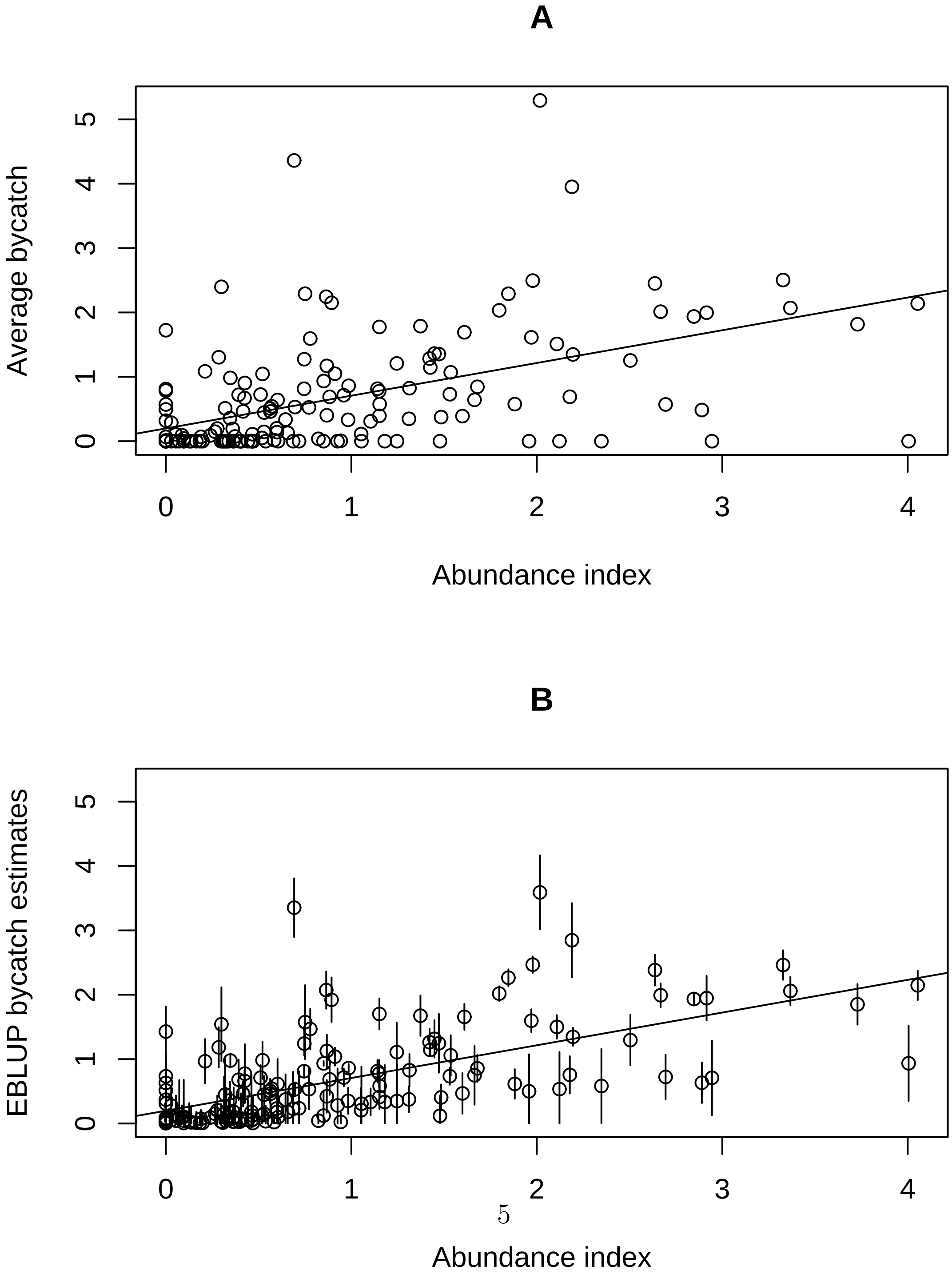


Figure 3: EBLUP vs Direct estimates. The line y=x is also shown.

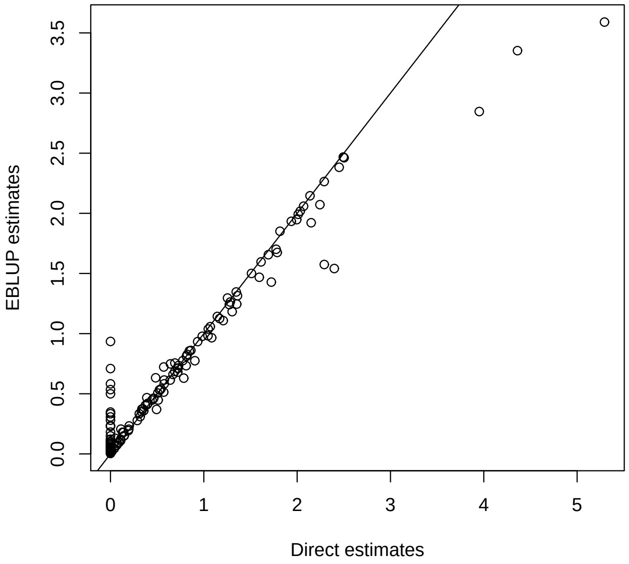
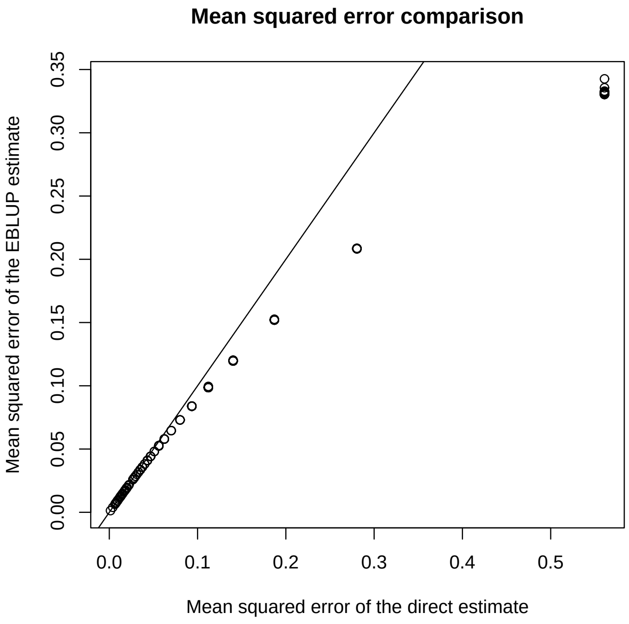


Figure 4 shows the mean squared error of $\tilde{\theta}$ vs that of $\hat{\theta}$. (The mean squared error of $\hat{\theta}$ is just its variance, σ_e^2/J_i). The mean squared error is dramatically reduced for points whose raw estimate has high variance. The sum of the 156 mean squared errors is 17.59 for $\tilde{\theta}$, and 24.77 for $\hat{\theta}$.

Figure 5 shows abundance index vs point estimates for the average untransformed bycatches, i. e. the means of the log normal distributions with parameters θ_i and σ_e^2, $\exp\theta_i + 0.5\sigma_e^2$. It is natural to estimate this quantity by plugging in estimates of $\exp\theta_i$ and σ_e^2. Our estimate of σ_e^2 is based on a large number of observations (all the deviations of the individual bycatch estimates from their bin means) and thus should have low variance. If we estimate $\exp(\theta)$ with $\exp(\tilde{\theta})$, the variance of $\tilde{\theta}$ will induce a bias in the estimate of $\exp(\theta)$. We partially correct for this in the following way: first, note that the $\tilde{\theta}$'s are each a mixture of a regression prediction and a sample mean. Both these quantities have approximately normal distributions. If we assume $\tilde{\theta}$ has a normal distribution, then $\exp(\tilde{\theta})$ has a log normal distribution, with variance parameter at least $\gamma_i^2\psi_i$; so the expected value of $\exp(\tilde{\theta})$ is at least $\exp\theta + 0.5\gamma_i^2\psi + i$. Thus, we use $\exp(\tilde{\theta}) + 0.5(\sigma_e^2 - \gamma_i^2\psi_i)$ as our estimate of

Figure 4: Mean squared error of EBLUP estimates vs Mean Squared error of direct estimates. The line y=x is also shown.



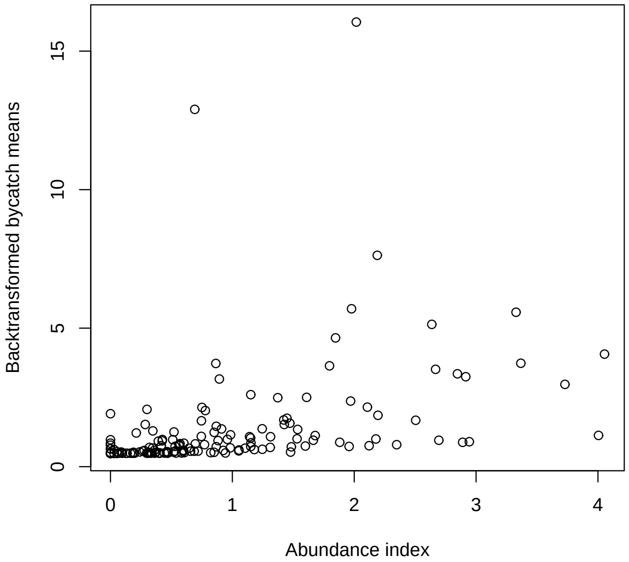
$\exp(\theta) + 0.5(\sigma_e^2)$. This still may be an over estimate since $\gamma_i^2 \psi_i$ is a lower bound for the variance of $\tilde{\theta}$.

Even after the EBLUP correction, there are still at least two points with quite large bycatch estimates. Comparison of the two panels in Figure 2 shows that the estimates for these cells have been modified considerably by the EBLUP procedure, but are still substantially above the bulk of the data.

5 Advantages and Shortcomings of EBLUP

The primary advantage of EBLUP is its simplicity. However, within this simple framework there is limited flexibility to make our model more realistic. One important factor we ought to take into account is the variability in the quality of the abundance measurements. The measurements are based on variable numbers of trawls. A response to this would be to fit different random effects variances for different trawl numbers; however, this would substantially complicate EBLUP framework. A related issue is bycatch esti-

Figure 5: Abundance index vs. estimated average catch.



mates for cells without abundance estimates. An abundance index for these cells could be estimated from neighboring cells; however, this abundance estimate almost certainly has a higher variance than abundances that have been measured directly.

Another issue is the form of our distribution for the random effects (v_i's) and measurement errors ($y_{ij} - \theta_i$). In reality these have truncated distributions, as our measurements cannot go below zero. This phenomenon primarily effects those cells with low abundance. As a result, the fitted values of these quantities deviate somewhat from our assumption of symmetry. Both distributions (seen in Figure 6) are somewhat skew, with the right tail longer than left tail. In addition, despite the log transformation the random effects do not have a constant variance with respect to the predicted values from the linear relationship with the abundance index (Figure 7).

The EBLUP model used does not exploit any effect on bycatch of season, area, depth, or year not reflected in the abundance estimate. (One consequence of this is that the EPLUP procedure cannot provide any estimates for cells where neither bycatch nor abundance estimates have been collected.)

Figure 6: Histograms of the estimated random effects and measurement errors, (on the scale which the model was fit on, both bycatch and abundance shifted by 1 and logged).

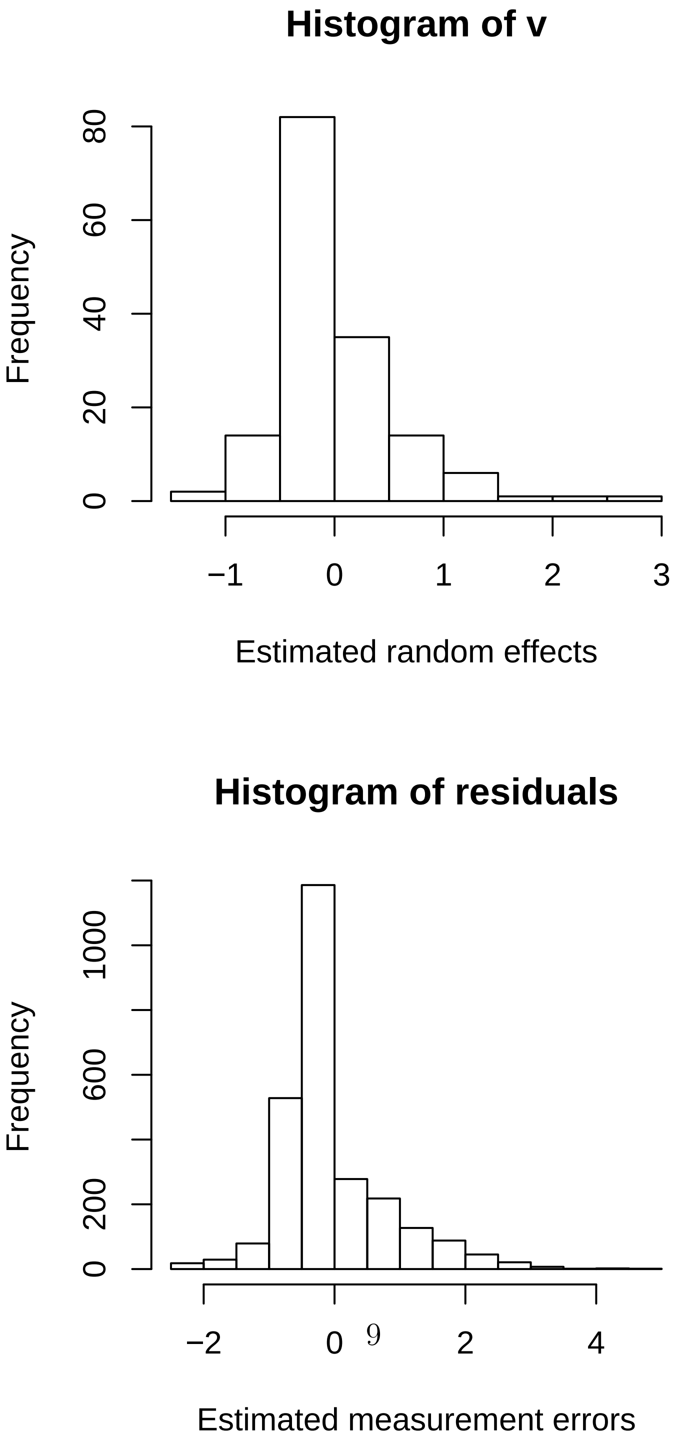
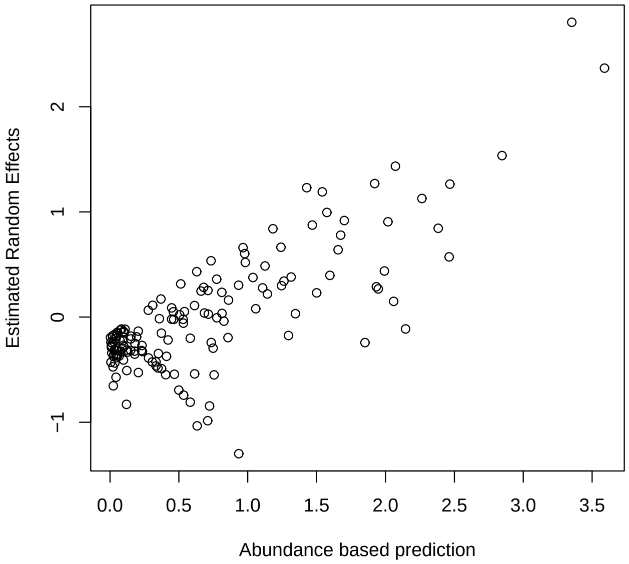


Figure 7: Random effects vs values predicted from the linear abundance relationship (on the scale which the model was fit on, both bycatch and abundance shifted by 1 and logged).

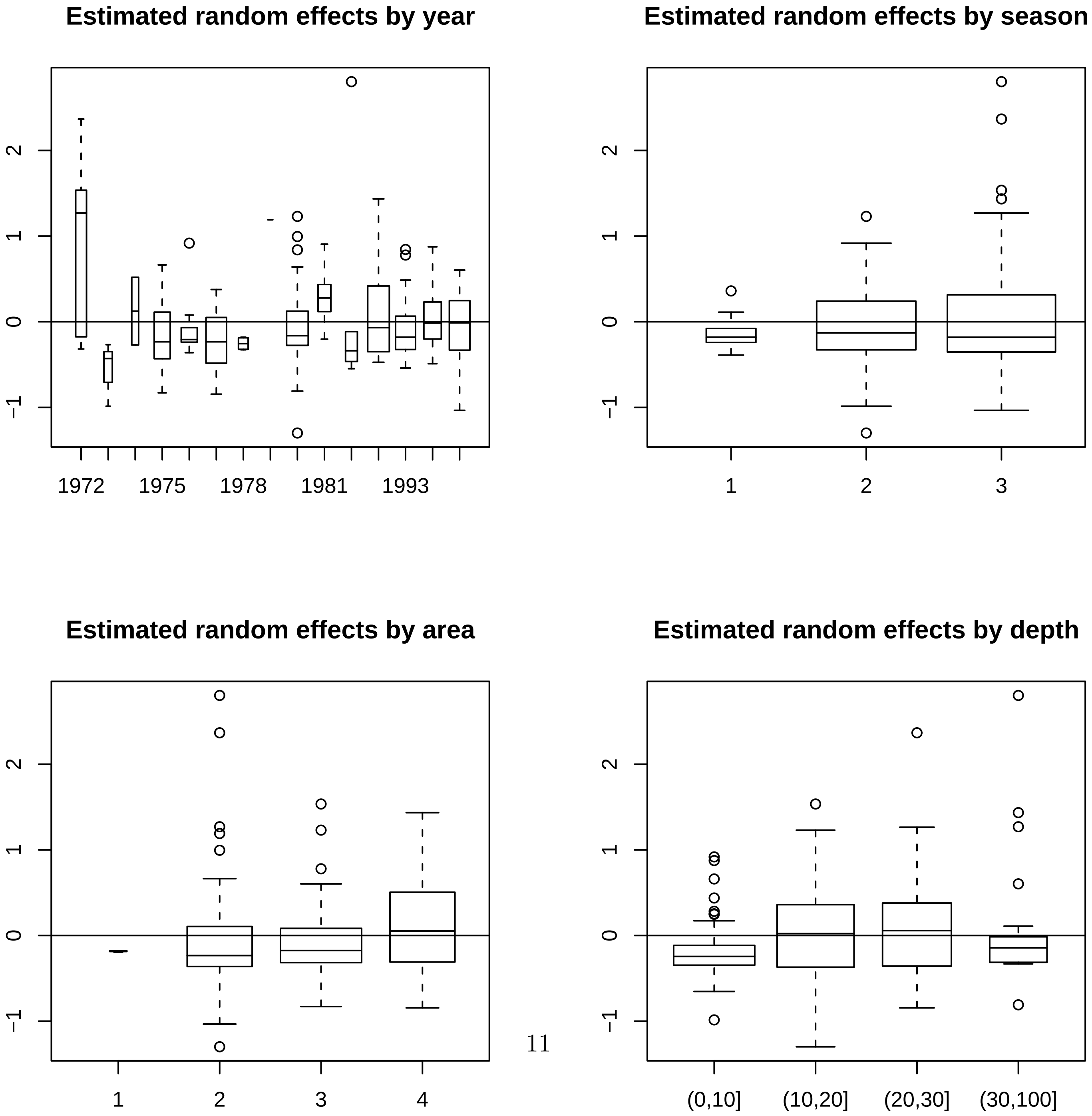


However, examining the random effects grouped by these factors, there are no pronounced trends suggesting this would be helpful (see Figure 8). The lack of trends may also be a result of the rather coarse grouping of data: spatial or temporal correlations may be evident at smaller scales.

6 Conclusions

The EBLUP estimates use the relationship between the abundance measurements and the bycatch measurements to improve the raw estimates for the cells that have abundance estimates. Assuming the bycatch-abundance relationship is constant, the method reduces mean squared error for cells that have abundance measurements but few (or outlying) bycatch measurements. One possibility for treating the entire data set is to simply use the raw estimates for cells without covariate information. However, the cells without abundance estimates include at least a few observations akin to the ones

Figure 8: Random effects vs binning factors. Box plot width is proportional to the square root of the number of observations in that category. Circles are observations lying more than 1.5 times the interquartile range outside the box (outliers).



that were significantly impacted by the EBLUP procedure. In addition, the abundance does show trends in time and space, so we are not completely without abundance information for these cells, although this information is substantially less than what is available when there are direct observations.

The presence of covariate information for each of the "small areas" is a fairly fundamental premise of the EBLUP framework. One could imagine a two stage procedure for cells without abundance observations where abundance is estimated by some model, and the fitted values used as predictors. This would, at the very least, require a different σ_v^2 to be fitted for these observations. This would compromise the simplicity of the EBLUP procedure; it would also be theoretically unsatisfying as it would not exploit our knowledge about the underlying source of the increased variance: the reduced precision of the abundance "measurement." This concern about the differing variances for abundance measurements also extends to the results of the current analysis, where the abundance measurements are based on widely varying numbers of measurements for different cells.

The various problems with few or no observations of abundance, bycatch, or both in some cells points to the fact that the data is simply not rich enough to estimate each cell mean essentially in isolation from the other cells, despite the attractive freedom from modeling assumptions such an approach would provide. Rather than trying to extend EBLUP to a situation outside its intended purpose, I would recommend using models that are designed to use data from cells that are "nearby" in time and/or space, and modify them to suit the idiosyncrasies of this data (especially the presence of many zero measurements). Nichols (2004) introduces a few such approaches that use ideas from Bayesian hierarchical models to add additional flexibility to the GLM model mentioned above. Jones (2004) explores exploiting the spatial structure of the data; while developing attractive models, the paper found that it was not possible to fit them using the computational methods currently available.

7 References

Ghosh, M. and J. N. K. Rao (1994). "Small area estimation: an appraisal." *Statistical Science* **9**: 55-76.

Jones, B. (2004). "Spatial modeling of red snapper shrimp fleet bycatch in the Gulf of Mexico." Report to the Gulf of Mexico Fishery Management Council.

Nichols, S. (2004). "Some Bayesian approaches to the measurement of shrimp fleet bycatch." Report to the Gulf of Mexico Fishery Management Council.

MRAG Americas Inc. (1997). "Consolidated report on the peer review of red snapper (*Lutjanus campechanus*) research and management in the Gulf of Mexico." Report the National Marine Fisheries Service.

Nichols, S. and G. J. Pellegrin Jr. (1992). "Revision and update of estimates of shrimp fleet bycatch, 1972-1991." Report to the Gulf of Mexico Fishery Management Council.

Nichols, S., A. Shah, G. J. Pellegrin Jr. and K. Mullin (1987). "Estimates of annual shrimp fleet bycatch for thirteen finfish species in the offshore waters of the Gulf of Mexico." Report to the Gulf of Mexico Fishery Management Council.

Nichols, S., A. Shah, G. J. Pellegrin Jr. and K. Mullin (1987). "Updated estimates of shrimp fleet bycatch in the offshore waters of the U. S. Gulf of Mexico." Report to the Gulf of Mexico Fishery Management Council.