***BAM Internal Report Series:* Dealing with the uncertainty of habitat classification in species-habitat modelling through covariate measurement error models**

By: Dr. Erin Bayne

Dated: Thursday, December-12-13

Note: This report is a draft product from a preliminary exploration. It has not yet yielded any manuscripts.

Rationale: Ecologists are increasingly using geographic information systems (hereafter GIS) and predictive statistical models to map the occurrence of species. Significant effort has been made to improve these models by collecting more species occurrence data at ever increasing spatial extents. Similarly, much research has been done to correct for errors in the estimation of species occurrence through approaches like multiple visit occupancy models. While such efforts undoubtedly will improve our ability to predict where species do and do not occur, there often remains considerable uncertainty in model predictions. Sources of model uncertainty are often unknown but are typically attributed to observation error and/or a failure to include all relevant covariates. This has resulted in ecologists seeking out more and more covariates AKA predictor variables to fully explain the variation in species occurrence that is observed.

A source of uncertainty that most ecologists ignore is the reality that predictor variables are also measured with error. This is particularly problematic when using GIS layers. All GIS layers I am familiar with come with a standard caveat that the layer has a certain level of classification error. Yet when modelling the distribution of species, ecologists typically ignore these caveats and use statistical models that assume the GIS layer they are using is completely accurate. Simply comparing two different GIS layers purporting to measure the same phenomena reveals this is rarely true.

For scientists creating remote sensing products, this uncertainty is dealt with by continually improving their models to achieve greater accuracy in classification. For the ecologist using these GIS layers however, the standard approach is to select “the best GIS layer” to describe the predictor variable of interest. How “best” is decided upon and why a particular GIS layer is used as a predictive layer is rarely discussed. More often than not, I suspect the GIS layer chosen as a tool for prediction is the one most readily available and when two or more are available the one with the highest spatial resolution is chosen. The problem with this approach is that another ecologist may make different decisions about what GIS layers to use, create a model using similar occurrence data, and draw very different conclusions about the ecological factors that influence a species. This reduces the perceived utility of species habitat models by managers and reduces the quality of conservation science.

Rather than trying to fix such uncertainty by continually improving the GIS layer in the hope of reaching the “absolute truth”, an alternative approach is to incorporate the uncertainty between GIS layers and use it in when estimating species – habitat relationships. An approach commonly used in the medical literature to deal with data where there is error in the predictor variable is known as the covariate measurement model. Covariate measurement models accept that the predictor variables are measured with error. Using covariates that have error is known to lead to biased estimates of how the response variable is influenced by a predictor variable. Various types of covariate measurement models can be used to correct for this bias if additional information is available such as the true covariate value in a validation subsample using some type of gold standard, instrumental variables, knowledge of the measurement error variance, or in the case examined here replicate measurements of the same phenomena.

A brief overview of CME and its application in wildlife habitat modelling:

Here I evaluate how different GIS layers and/or use of GIS layers can be used to obtain replicate measurements of the same habitat phenomena to see how correcting for errors in GIS layers influences estimation of species occurrence patterns. I highlight the use of CME when a single covariate is estimated with error from two or more different GIS layers. The underlying rationale to the CME approach is that the true response to an environmental covariate by a species is an unobservable or latent "true" variable that can be estimated.

When an environmental covariate is measured with error it is known as a fallible measure. When an analyst has multiple measures of the same covariate, the "degree of fallibility" can be estimated and incorporated into the model that predicts the probability of observing a species or number of individuals of a species. The covariate measurement error model evaluated here is an extension of generalized linear models using maximum likelihood as implemented through an adaptive quadrature method available in the GLLAMM command and CME wrapper in the program Stata.

The CME model consists of three sub-models. The outcome model describes how the species reacts to the estimated "true" value of the environmental covariate. The measurement model specifies how the true covariate is related to the fallible measures used to estimate it. Typically, this is expressed as a measure of reliability. Finally, the true covariate model is a regression of the true covariate on additional covariates that are observed without error. The true covariate model allows estimation of direct and indirect effects of the fallible measures on the overall outcome when additional covariates that are measured without error are included in the model. Detailed proofs and the mathematical derivation of the CME model is available in (Rabe-Hesketh, Skrondal, & Pickles, 2003).

Bird occurrence data:

I selected 10,050 point count locations from the Boreal Avian Modelling database ver3. All of these points were within the boreal forest ecoregion of the province of Alberta, Canada. Point counts involved people standing at a point and listening for birds of all species for a period ranging from 3 to 10 minutes. The distance over which birds were surveyed varied between the data included in this study but did so more or less randomly with respect to spatial location. In this analysis, I only use presence-not detected data to reduce the effects caused by differential survey duration and distance. I chose to model the Ovenbird because my personal experience provides me with a very good understanding of the factors that are correlated with its occurrence.

Reclassification differences in GIS layers:

One of the most commonly used remote sensing products for predicting land-cover in North America and hence species – habitat relationships come from the LANDSAT satellite. The LANDSAT satellite program was launched in 1972 and is the longest running satellite imagery acquisition program. Using information from the Landsat-7 satellite, the Earth Observation for Sustainable Development of Forests (hereafter EOSD) program in Canada created a wall-to-wall map of Canadian forest cover *circa* 2000 using data from different Landsat scenes collected during cloud free conditions across different months and years. The EOSD layer has 44 classes that were derived by unsupervised clustering via the K-means method. The end product is a raster image that classifies each 25 x 25 metre area in Canada's forests.

Several agencies have created value-added products from the EOSD layer. The Alberta Biodiversity Monitoring Institute (hereafter ABMI) Wall-to-wall Land Cover Map is a polygon-based representation of Alberta's land cover *circa* 2010 that was derived from the EOSD. The map describes the spatial distribution of 11 land cover classes across the province of Alberta. The map consists of approximately 1 million non-overlapping polygons of various sizes with a minimum size of 0.5 hectares (ha) for aquatic features and 2 ha for all others. Each polygon represents a contiguous area relatively homogeneous in terms of land cover, where the specific land cover class of each polygon is different from that of adjacent polygons. By design this layer emphasizes human footprints, such that: 1) The width of roads is systematically exaggerated to a minimum of 60 m (two Landsat pixels); 2) Forest areas harvested or burned between 2000 and 2010 were assigned to the 'Shrub' class. Thus, this class also represents young forests and; 3) the accuracy of the natural shrub class is low (30%). Many shrub polygons are in reality forest. Thus, the exact definition of what is forested differs between the two GIS layers and creates a good example of covariate measurement error caused by different reclassification rules to define the same environmental phenomena. From these two layers, I estimated the proportion of a 150 metre radius buffer around the bird survey points that was wooded. WOODED was the summed area of trees or shrubs, as defined by each layer, divided by the total area of the 150 metre buffer.

The correlation between the EOSD and ABMI estimate of WOODED was low (r = 0.32). Reclassification rules create considerable differences in how WOODED predicts observation of Ovenbirds. Because of the systematic exaggeration of roads in the ABMI layer, a relatively shallow increase (0.258 to 0.374) in the probability of observing of an Ovenbird as WOODED increased from 0 to 1 was seen. In contrast, the EOSD layer changed from 0.061 to 0.475 over the same range of WOODED. The outcome model from CME, while consistent in having a positive slope, indicated the estimated true effect of WOODED (0.004 to 0.574) was much larger than either standard logistic regression model using EOSD or ABMI. As well, when the value of WOODED in the CME model was low (less than 0.5) the probability of observing an Ovenbird had a very shallow slope that began to rise rapidly thereafter. From personal experience is more consistent with this species life history as they are a forest specialist (Figure 1). The exaggeration of roads in ABMI is the primary reason for the magnitude of this difference. The low reliability (0.28) of the fallible measure of WOODED and high variance (0.004) results in a 4-fold increase in the size of the standard error for the true estimate of WOODED based on CME.

Table 1 – Coefficient values for regression models predicting probability of observing an Ovenbird as a function of different definitions of the predictor variable WOODED based on the EOSD and ABMI GIS layers. The CME model describes the outcome model from a covariate measurement error model that combines the uncertainty in the definition of EOSD and ABMI layers to estimate the true response of Ovenbirds to WOODED. Data are reported as raw coefficients ± standard error.

PREDICTOR	EOSD MODEL	ABMI MODEL	CME MODEL
WOODED	2.634 ± 0.111	0.541 ± 0.092	5.900 ± 0.394
CONSTANT	-2.734 ± 0.097	-1.057 ± 0.085	-5.601 ± 0.340

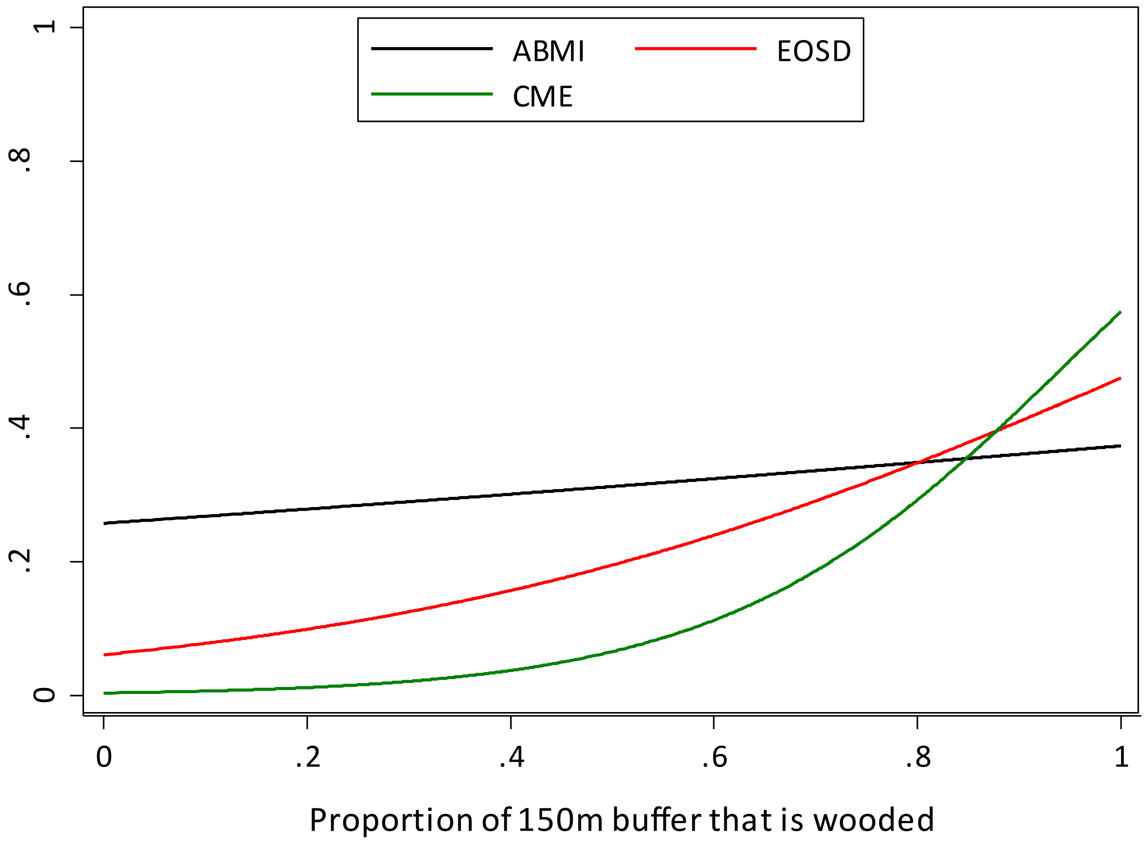

Figure 1 – Probability of observing an Ovenbird as a function of the proportion of a 150 metre buffer being WOODED. Probability curves are shown for individual logistic regressions using the ABMI and EOSD layers as the predictor of WOODED. The CME curve shows the estimated probability when the true estimate of WOODED is applied.

Next, I evaluated how Ovenbirds react to forest composition while controlling for WOODED. In this analysis, I dropped those point counts where either the EOSD or ABMI layer indicated there was no forest cover. Forest composition cannot be calculated for areas without woody plants. I estimated the proportion of the area in each layer that was comprised of coniferous trees relative to the summed total of all woody vegetation (hereafter CONIFER). Mixedwood pixels or polygons were assumed to be 50% conifer. I then created two standard logistic regression models and one CME model:

A) logit oven WOODED-EOSD CONIFER-EOSD
B) logit oven WOODED-EOSD CONIFER-ABMI
C) cme oven WOODED-EOSD (conifer: CONIFER-EOSD CONIFER-ABMI)

In this CME model, I am testing how CONIFER influences probability of observing an Ovenbird when there is error in estimating CONIFER but assuming that WOODED is measured without error based on the EOSD layer. In the CME wrapper, only one covariate measured with error can be estimated so I chose to include WOODED-EOSD because it described more of the variation in the probability of observing an Ovenbird than WOODED-ABMI (see **Next Steps** section of how to extend this approach to generate models that have multiple covariates measured with error).

Unlike WOODED, where the exaggeration of roads in the ABMI layer creates a fundamental change in the way an analyst perceives the amount of WOODED area relative to EOSD, CONIFER was more similar although not identical between layers (r = 0.705). The differences between the two layers are driven by the fact that: 1) in an unsupervised classification like EOSD some pixels are misclassified which creates a “salt and pepper” appearance. The ABMI layer used a variety of filters, smoothing of class boundaries, and removing small isolated regions to create a layer of contiguous homogeneous polygons; 2) the minimum mapping unit in the ABMI layer was 2 hectares which incorporates ~16 EOSD pixels; 3) polygon boundaries were smoothed in the ABMI whereas in the EOSD rasters exist as square cells, so that when the proportion CONIFER or WOODED estimates are calculated in the GIS the exact areas clipped, buffered, and intersected do not match perfectly in areal coverage; and 4) The areas exaggerated by the road polygons in ABMI result in fewer trees than the EOSD and it is possible that this does not occur randomly with respect to forest composition (i.e. roads are preferentially put in upland areas more likely to be dominated by deciduous forest types).

The consequence of these slight variations in GIS processing when estimating the probability of observing an Ovenbird are shown in Table 2 and Figure 3. When the 150 metre buffer was completely WOODED (1), I found that the probability of observing an Ovenbird declined as CONIFER increased. In the EOSD layer, increasing CONIFER from 0 to 1 resulted in the probability of observing an Ovenbird decreasing from 0.698 to 0.158, the ABMI layer from 0.619 to 0.210, and the CME model from 0.741 to 0.095. As with WOODED, the CME estimate for CONIFER resulted in a steeper slope. Reliability of the CONIFER variable was much higher than WOODED (0.674) which is also reflected in the fact that the SE for the true coefficient value for CONIFER was only ~1.5 times higher than the standard logistic regression models using EOSD or ABMI predictors of CONIFER alone.

Table 2 - Coefficient values for regression models predicting probability of observing an Ovenbird as a function of different definitions of the predictor variable CONIFER based on the EOSD and ABMI GIS layers. The CME model describes the outcome model from a covariate measurement error model that combines the uncertainty in the definition of EOSD and ABMI layers to estimate the true response of Ovenbirds to CONIFER while controlling for the proportion of a 150 metre buffer that was WOODED as based on the EOSD GIS layer. Data are reported as raw coefficients ± standard error.

PREDICTOR	EOSD MODEL	ABMI MODEL	CME MODEL
CONIFER	-2.510 ± 0.068	-1.811 ± 0.063	-3.312 ± 0.104
WOODED	2.144 ± 0.121	2.636 ± 0.119	2.271 ± 0.126
CONSTANT	-1.306 ± 0.108	-2.152 ± 0.104	-1.218 ± 0.114

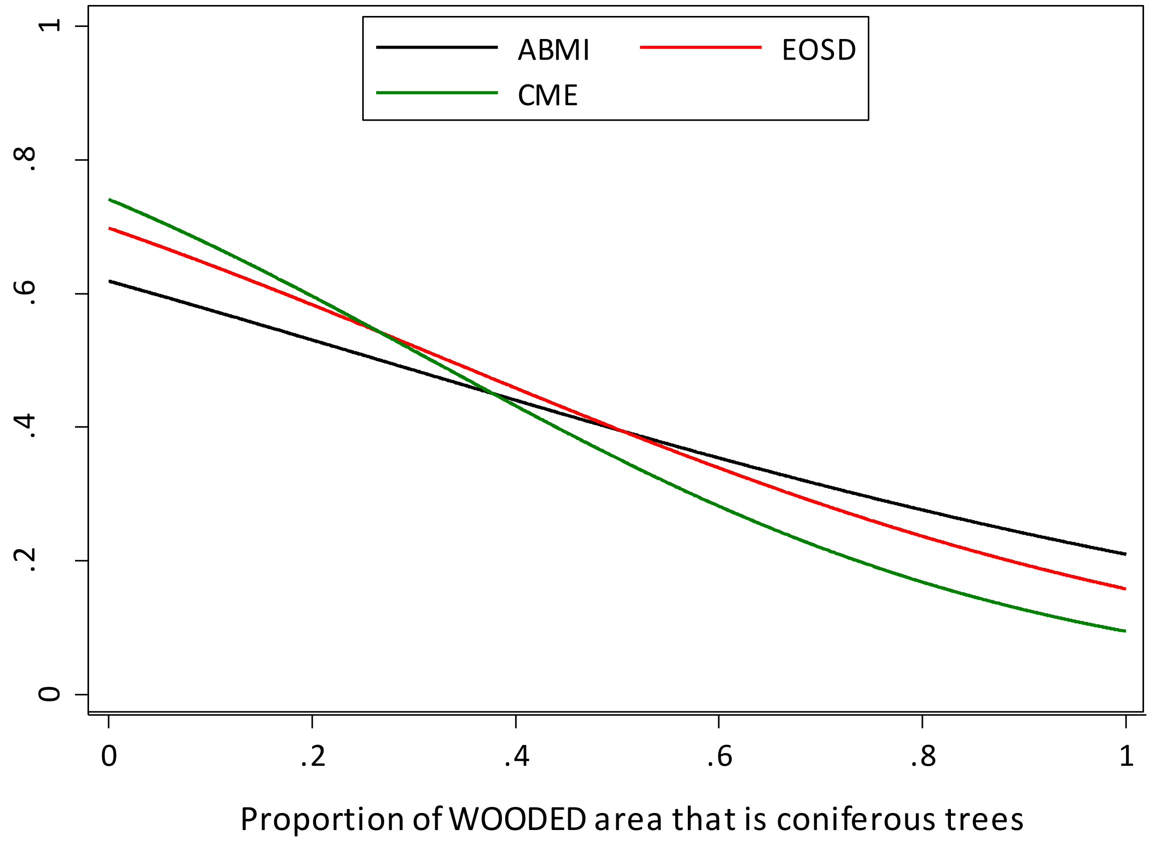

Figure 2 – Probability of observing an Ovenbird as a function of the proportion of WOODED area that is comprised of coniferous trees within a 150 metre buffer. Probability curves are shown for individual logistic regressions using the ABMI and EOSD layers as the predictor of CONIFER and WOODED-EOSD. The CME curve shows the estimated probability when the true estimate of CONIFER is applied and WOODED-EOSD is set to a value of 1.

Image resolution from different satellites:

Every day different satellites take thousands of photographs of earth. As such, numerous options exist for the types of imagery that can be used to categorize land-cover. The Land Cover Classification (hereafter LCC) of Canada *circa* 2005 is another product available for all of Canada that theoretically maps things similarly to what is done by EOSD. This map was created using the MODIS satellite which has a spatial resolution of 500 metre pixels that are downscaled into 250 metre pixels. While the classification algorithm and grain size of the LCC are quite different from the EOSD layer, they have a similar classification scheme that should allow users to “collapse” information to accurately derive WOODED and CONIFER. The fundamental difference is that the large grain size results in much more variation within a pixel in LCC than occurs within the EOSD because the LCC is classified primarily based on the pixel’s dominant land-cover type. Because of the large grain size of the LCC, I used a buffer of 1000 metres when estimating WOODED and CONIFER for both the EOSD and LCC layers in this analysis. If I had used a 150 metre buffer then I simply would have had two possible values of WOODED (0 or 1) and three possible values of CONIFER (0,0.5,1). Future efforts should explore how downscaling the LCC to 25 x 25 metre resolution influences this analysis.

Using the same modelling approach described above, I saw a very similar pattern to what was observed in **Reclassification differences in GIS layers.** The LCC model had the shallowest slope with probability of observing an Ovenbird increased from 0.106 to 0.456 as WOODED increased from 0 to 1, EOSD increasing from 0.042 to 0.548, and CME from 0.002 to 0.731. Reliability of the two WOODED estimates was relatively low (0.366) with an error variance of 0.028.

Controlling for WOODED-EOSD in a 1000m buffer, I found that the probability of observing an Ovenbird decreased from 0.799 to 0.302 as CONIFER increased from 0 to 1 with LCC, 0.795 to 0.147 with EOSD, and 0.872 to 0.133 with CME. Reliability of the two CONIFER estimates was relatively high (0.721) with an error variance of 0.024. The LCC layer had the shallowest slope suggesting that some of the variation within the relatively large grain size of an LCC pixel was masking important variation in forest composition that Ovenbirds were responding to and that was better described in the EOSD.

Table 3 - Coefficient values for regression models predicting probability of observing an Ovenbird as a function of different definitions of the predictor variable WOODED based on the EOSD and ABMI GIS layers. The CME model describes the outcome model from a covariate measurement error model that combines the uncertainty in the definition of EOSD and ABMI layers to estimate the true response of Ovenbirds to WOODED. Data are reported as raw coefficients ± standard error.

PREDICTOR	EOSD MODEL	LCC MODEL	CME MODEL
WOODED	3.315 ± 0.146	1.959 ± 0.103	7.178 ± 0.366
CONSTANT	-3.122 ± 0.116	-2.135 ± 0.087	-6.176 ± 0.290

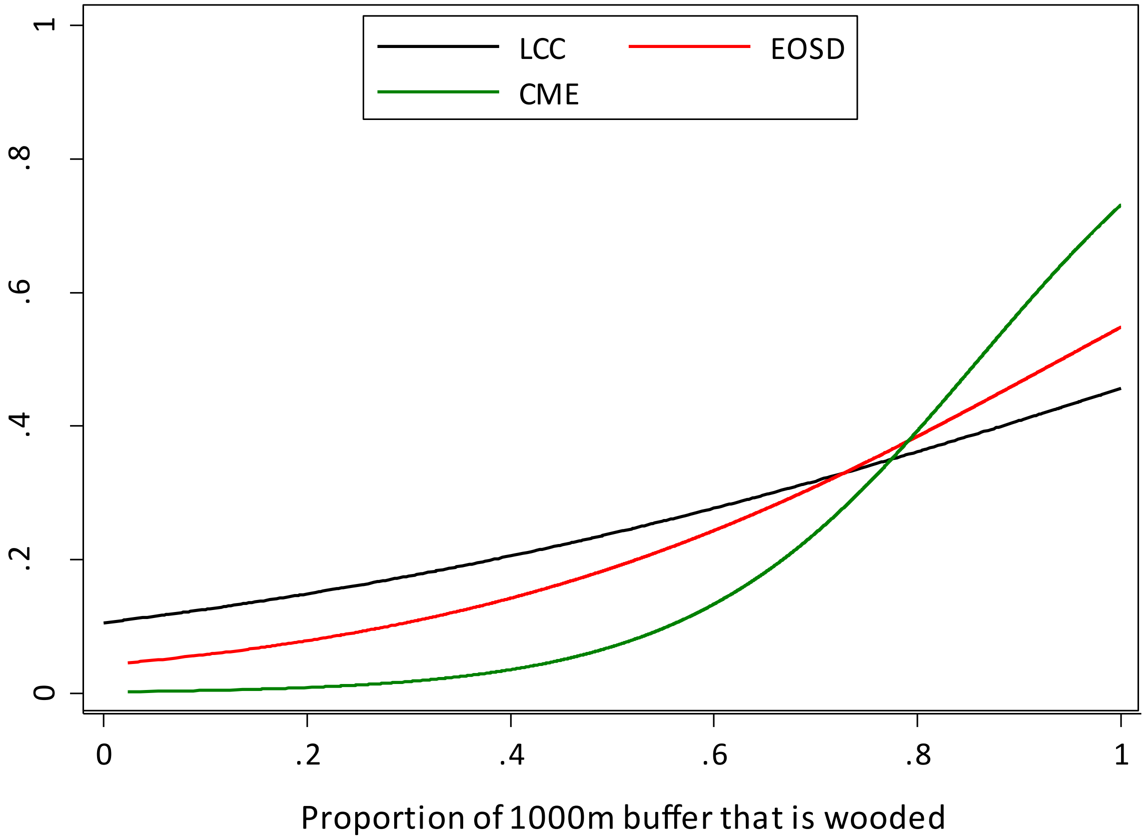

Figure 3 – Probability of observing an Ovenbird as a function of the proportion of a 1000 metre buffer being WOODED. Probability curves are shown for individual logistic regressions using the EOSD and LCC layers as the predictor of WOODED. The CME curve shows the estimated probability when the true estimate of WOODED is applied.

Table 4 - Coefficient values for regression models predicting probability of observing an Ovenbird as a function of different definitions of the predictor variable CONIFER based on the LCC and EODS GIS layers. The CME model describes the outcome model from a covariate measurement error model that combines the uncertainty in the definition of LCC and EOSD layers to estimate the true response of Ovenbirds to CONIFER while controlling for the proportion of a 1000 metre buffer that was WOODED as based on the EOSD GIS layer. Data are reported as raw coefficients ± standard error.

PREDICTOR	EOSD MODEL	LCC MODEL	CME MODEL
CONIFER	-3.111 ± 0.082	-2.218 ± 0.083	-3.796 ± 0.0.119
WOODED	1.792 ± 0.156	3.337 ± 0.149	2.411 ± 0.157
CONSTANT	-0.438± 0.136	-1.954 ± 0.124	-0.493± 0.142

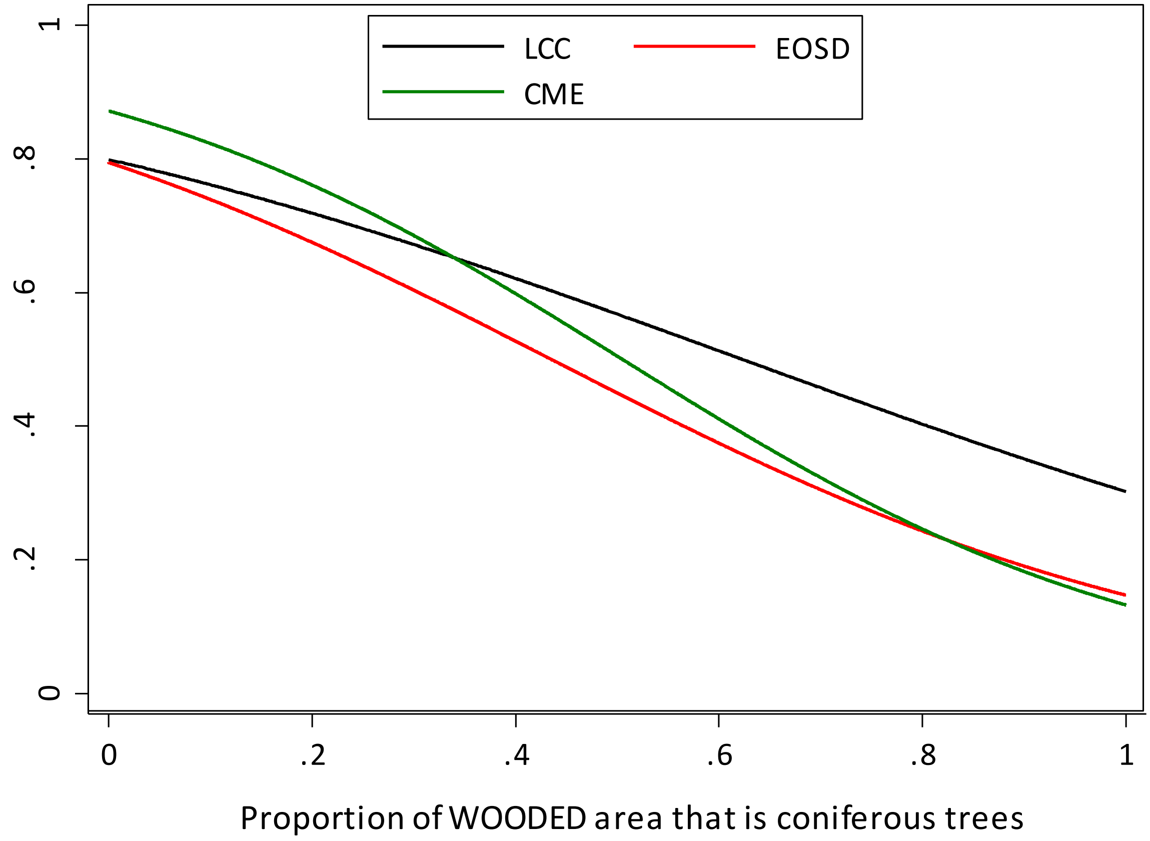

Figure 4 – Probability of observing an Ovenbird as a function of the proportion of the WOODED area that is comprised of coniferous trees within a 1000 metre buffer. Probability curves are shown for individual logistic regressions using the LCC and EOSD layers as the predictor of CONIFER and WOODED-EOSD. The CME curve shows the estimated probability when the true estimate of CONIFER is applied and WOODED-EOSD is set to a value of 1.

Remote sensing versus aerial photography - The additional issue of missing data:

One of the great strengths of satellite based remote sensing is the vast spatial extent that often is available. However, the accuracy of the classification that results is often more variable than desired. Some suggest that human-based aerial photography will be more accurate than supervised or unsupervised classification algorithms commonly applied to satellite based remote sensing. The validity of this notion is debated by many. For forestry companies, aerial photo based photo-interpretations called Forest Resource Inventories (hereafter FRI) are the gold standard for determining tree species composition and wood volume. As many wildlife ecologists are creating models that try to track changes in forested systems driven by forestry activities, there is also a perception that such information is critical for creating good wildlife habitat models and that such models are better than those derived from satellite based classifications.

The major challenge with FRI as a product for mapping wildlife habitat is that it is very expensive to create and is updated relatively infrequently. Because of its expense it is often not available for all areas, limiting the ability of an ecologist to generate maps of wildlife habitat at broad spatial scales using FRI. One of the major benefits of CME modelling is that repeated measurements of the same covariate from two different GIS layers need not occur everywhere to include all of the available species data. CME models can be estimated using the entire sample size of species data as long as there is at least some repeated covariate data from each of the different GIS layers that is matched. CME effectively estimates the missing values when computing how species respond to environmental covariates measured with error assuming that unavailable data are missing at random (MAR).

The Forest Resource Inventory product available for the province of Alberta is known as the Alberta Vegetation Inventory (hereafter AVI). I clipped all of the polygons that intersected our 150 metre buffer and reclassified these polygons to estimate WOODED as well as the proportion of the wooded area that was CONIFER.

I evaluated how the probability of observing an Ovenbird was predicted by using the EOSD, ABMI, and AVI as the descriptor of the predictor variable WOODED. The correlation between these three GIS layers was weak: (A) EOSD and ABMI : r = 0.285, (B) EOSD and AVI : r = 0.427, (C) ABMI and AVI : r = 0.195). 13 separate analyses were attempted in this section to evaluate the effects of using a particular GIS layer and excluding those points for which GIS information was not available for one or more of the GIS layers. First, I used all the point count locations regardless of whether there was GIS data available for that point in the three layers. Second, I only used those point count locations where all three GIS layers provided estimates of WOODED. I also compared how the CME models compared when I used all three fallible predictors (EOSD, ABMI, and AVI) versus using combinations of GIS layers: (A) EOSD and ABMI; (B) EOSD and AVI; and (C) ABMI and AVI.

Table 5 - Coefficient values for regression models predicting probability of observing an Ovenbird as a function of different definitions of the predictor variable WOODED based on the EOSD, ABMI, and AVI GIS layers. The CME models describe the outcome from covariate measurement error models that combine the uncertainty in the definition of all three layers and combinations of layers to estimate the true response of Ovenbirds to WOODED. Data are reported as raw coefficients ± standard error. Using ABMI-AVI layers as the fallible measures of WOODED did not result in a model solution.

PREDICTOR	EOSD ALL	EOSD SUBSET	ABMI ALL	ABMI SUBSET	AVI SUBSET	EOSD-ABMI-FRI ALL	EOSD-ABMI-FRI SUBSET
WOODED	2.634± 0.111	2.549± 0.114	0.541± 0.092	0.269± 0.097	2.617± 0.092	8.581± 0.426	8.807± 0.458
CONSTANT	-2.734± 0.097	-2.514± 0.099	-1.057± 0.085	-0.667± 0.090	-2.519± 0.081	-7.735± 0.361	-7.765± 0.387
	EOSD-ABMI ALL	EOSD-ABMI SUBSET	EOSD-AVI ALL	EOSD-AVI SUBSET	ABMI-AVI ALL	ABMI-AVI SUBSET	
WOODED	5.915± 0.396	6.188± 0.486	8.633± 0.427	8.294± 0.412	N/A	N/A	
CONSTANT	-5.614± 0.341	-5.713± 0.418	-7.546± 0.352	-7.103± 0.338	N/A	N/A	

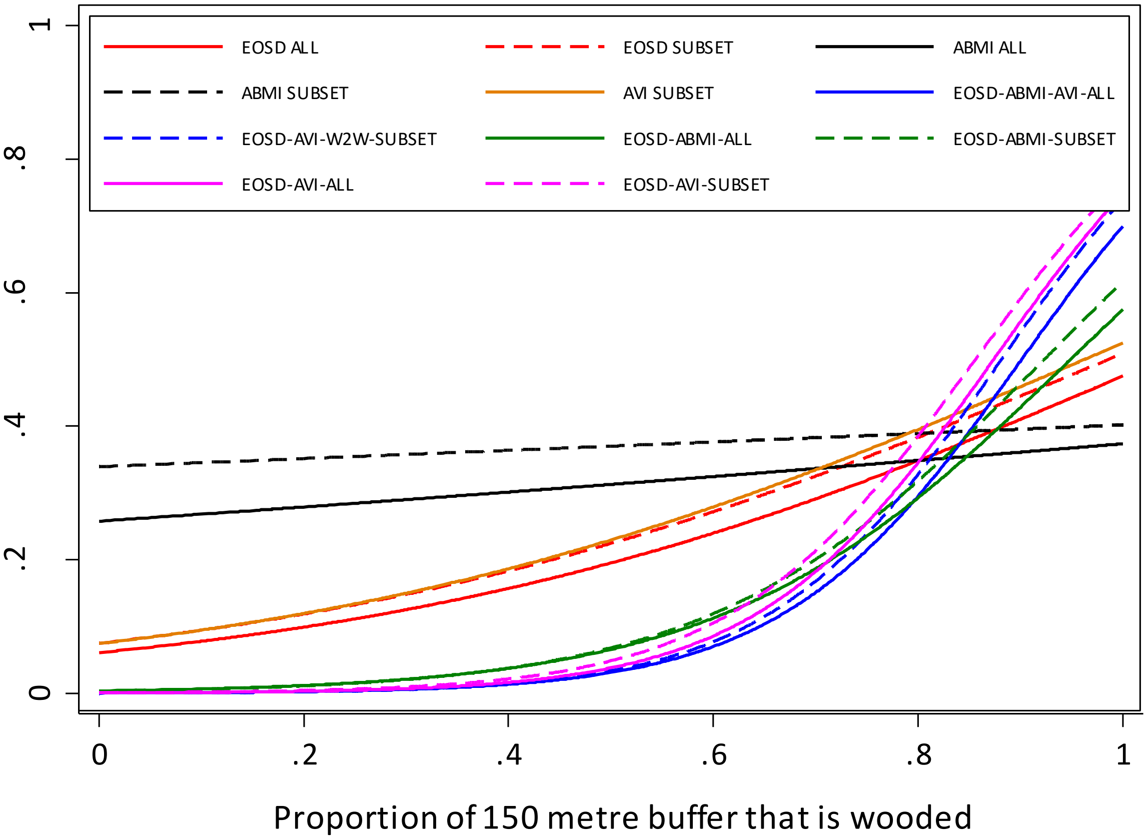

Figure 5 – Probability of observing an Ovenbird as a function of the proportion of a 150 metre buffer being WOODED. Probability curves are shown for individual logistic regressions using the EOSD, ABMI, and AVI layers as well as combinations of each layer in a CME framework which shows the estimated probability when the true estimate of WOODED is applied. Dashed lines show results when only those points where the covariate WOODED could be measured in all three layers was available (SUBSET) versus allowing for missing covariate values within one more measures of WOODED (ALL).

The AVI model was more consistent with the EOSD model than the ABMI model. This suggests that the AVI and EOSD may provide a more reliable estimate of WOODED although the correlation between the two is still quite poor. The ABMI layer as described in **Reclassification differences in GIS layers:** had the shallowest slopes, which is because of the exaggerated size of roads in this layer (roads are defined as not WOODED). In a qualitative sense, it did not really matter whether all 3 estimates of WOODED were used, whether ALL or a SUBSET of the data points were included, or whether I used pairs of WOODED estimates from the different layers. Ovenbirds had very low probability of observation when wooded was below 0.4 and increased in occurrence quickly thereafter in all CME models. However, the predicted probability of observation of Ovenbirds when using ALL of the datapoints for each layer was quite different when WOODED reached 1. For EOSD it was 0.475, ABMI 0.374,AVI 0.524, EOSD-ABMI CME 0.575, EOSD-AVI CME 0.748, and EOSD-ABMI-AVI CME 0.700. Reliability of the fallible estimates was EOSD-ABMI 0.274, EOSD—AVI 0.391, and EOSD-ABMI-AVI 0.270. The ABMI-AVI CME model failed to solve for reasons that I do not understand at this time.

This analysis demonstrates that having at least two estimates of any covariate can have dramatic effects on the nature of the species occurrence models. Adding the third did not change the shape of the curve as much but did have a fairly large influence on the predicted values near the end of the curve (WOODED = 1). Given the low reliability of all CME models that included the ABMI model it would suggest that excluding such a layer as a predictor might be a better choice that including it. However, if ABMI data was the only other measure of a fallible predictor that was being considered it seems to help generate a predictive model that is more consistent with Ovenbird natural history than ignoring covariate measurement error entirely. In this particular, application including ALL the data rather than only using the SUBSET for which WOODED could be estimated from the three layers did not have a major qualitative effect but typically resulted in models with higher estimated probability of Ovenbird observation at high values of WOODED.

CONIFER was a more reliable measure than WOODED, ranging from 0.50 to 0.64 depending on which GIS layers were used as repeated measures. As with all other analyses, the use of CME resulted in slopes that were considerably steeper than analysis that assumed that CONIFER was measured without error in each GIS layer. The inclusion of two or more GIS layers to estimate covariate measurement error resulted in some models predicting a ~ 20% different probability of observing an Ovenbird as a function of CONIFER when conifer was 0 or 1 (Table 6 & Figure 6).

Table 6 - Coefficient values for regression models predicting probability of observing an Ovenbird as a function of different definitions of the predictor variable CONIFER based on the EOSD, ABMI, and AVI GIS layers. The CME model describes the outcome model from a covariate measurement error model that combines the uncertainty in the definition of all three layers and combinations of layers to estimate the true response of Ovenbirds to CONIFER. Data are reported as raw coefficients ± standard error. WOODED is controlled for in the models and was derived from the EOSD layer.

PREDICTOR	EOSD	ABMI	AVI	EOSD-ABMI-AVI
WOODED	1.997± 0.121	2.490± 0.119	2.503± 0.117	2.080± 0.125
CONIFER	-2.407± 0.069	-1.618± 0.065	-2.554± 0.082	-3.920± 0.126
CONSTANT	-1.093± 0.109	-1.974± 0.104	-1.784± 0.102	-0.835± 0.115
	EOSD-ABMI	EOSD-AVI	ABMI-AVI	
WOODED	2.064± 0.123	1.908± 1.311	2.403± 0.129	
CONIFER	-3.248± 0.122	4.811± 0.166	-4.431± 0.174	
CONSTANT	-0.960± 0.114	-0.351± 0.127	-1.100± 0.118	

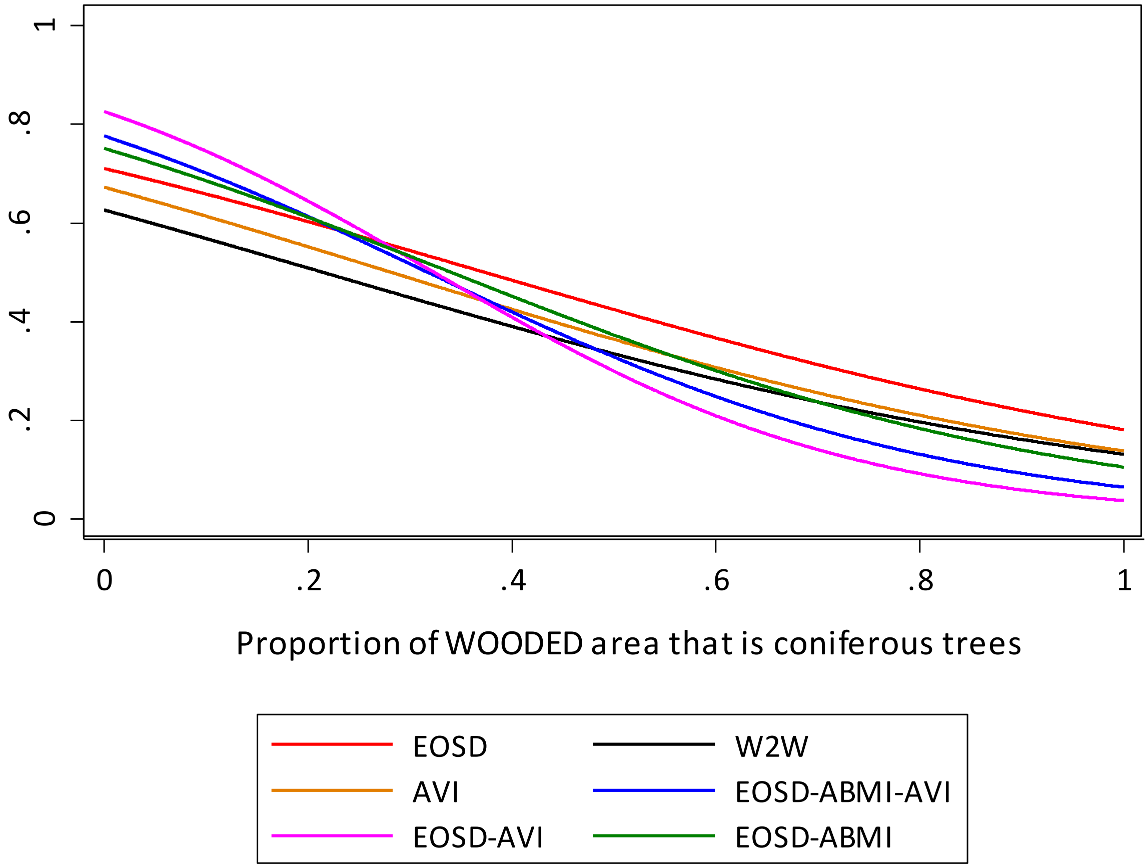

Figure 6 – Probability of observing an Ovenbird as a function of the proportion of the WOODED area that is comprised of coniferous trees within a 150 metre buffer using ABMI, EOSD, and AVI GIS layers. Probability curves are shown for individual logistic regressions using each layer as the predictor of CONIFER while controlling for WOODED-EOSD. The CME curves show the estimated probability when the true estimate of CONIFER is applied using varying combinations of GIS layers and while WOODED-EOSD is set to a value of 1.

Species location error & how this influence species - habitat models:

Typically when species – habitat models are created, the analyst buffers the point location where the observer was standing at when he recorded the species. However, using acoustic cues to detect birds means that individuals could be up to hundreds of metres away from the observer. As such, the habitat classification derived at the point where the observer was standing might not be a particularly good predictor of the habitat the birds were in. Since distance to an acoustic signal can be difficult for observers to estimate in the field, ecologists generally ignore this reality and assume that the habitat at the point of observation is consistent with where the species actually was or they create a very large buffer to ensure that the habitat the bird was in is incorporated in the predictor variable.

An alternative approach would be to assume that species was not in the buffer around the observer's location but occurred at some distance and bearing X from the observer. To incorporate the uncertainty in habitat conditions caused by uncertainty in where the species actual was, I speculate that a series of replicate buffers at *X* distance and *X* bearing from the observer could be used as replicate measurements of habitat conditions in CME models. I have not seen this approach used in the literature so more work is required to justify using such an approach.

To test this, I used the EOSD layer and generated four points 50 metres away from the point count location in each of the cardinal direction. I then created 50 metre buffers around the 5 points and measured WOODED and CONIFER in each buffer. These 5 replicate observations were then used in the CME model.

Evaluating the 5 possible "descriptions" of WOODED and CONIFER that could be created using these points, I found quite consistent models for predicting the probability of observing an Ovenbird. In other words, even though the habitat definitions varied at the point level there was a strong enough correlation between the replicate observations to not have a major influence on predictions. In part, this is caused by the fact that many of the surveys done in the BAM database were designed to sample a particular habitat strata and thus were in a relatively large area of contiguous habitat with the same classification. The further the buffers are moved from the central point however and potentially the larger the buffers the more of an influence this approach could have on estimation accuracy. This approach could also be of greater value for point counts done at the edge of two habitat classes. All that being said, the difference between the individual models and CME although small is like all other examples in this report in that the CME results in a steeper slope than models where covariate measurement error is ignored (Figure 7).

The validity of this concept needs to be evaluated through a GIS simulation where known densities of birds are modelled as a function of habitat versus non-habitat for a theoretical species where point count stations are randomly generated that include a range of percentages of habitat versus non-habitat to see how "error in spatial location" influences habitat selection models.

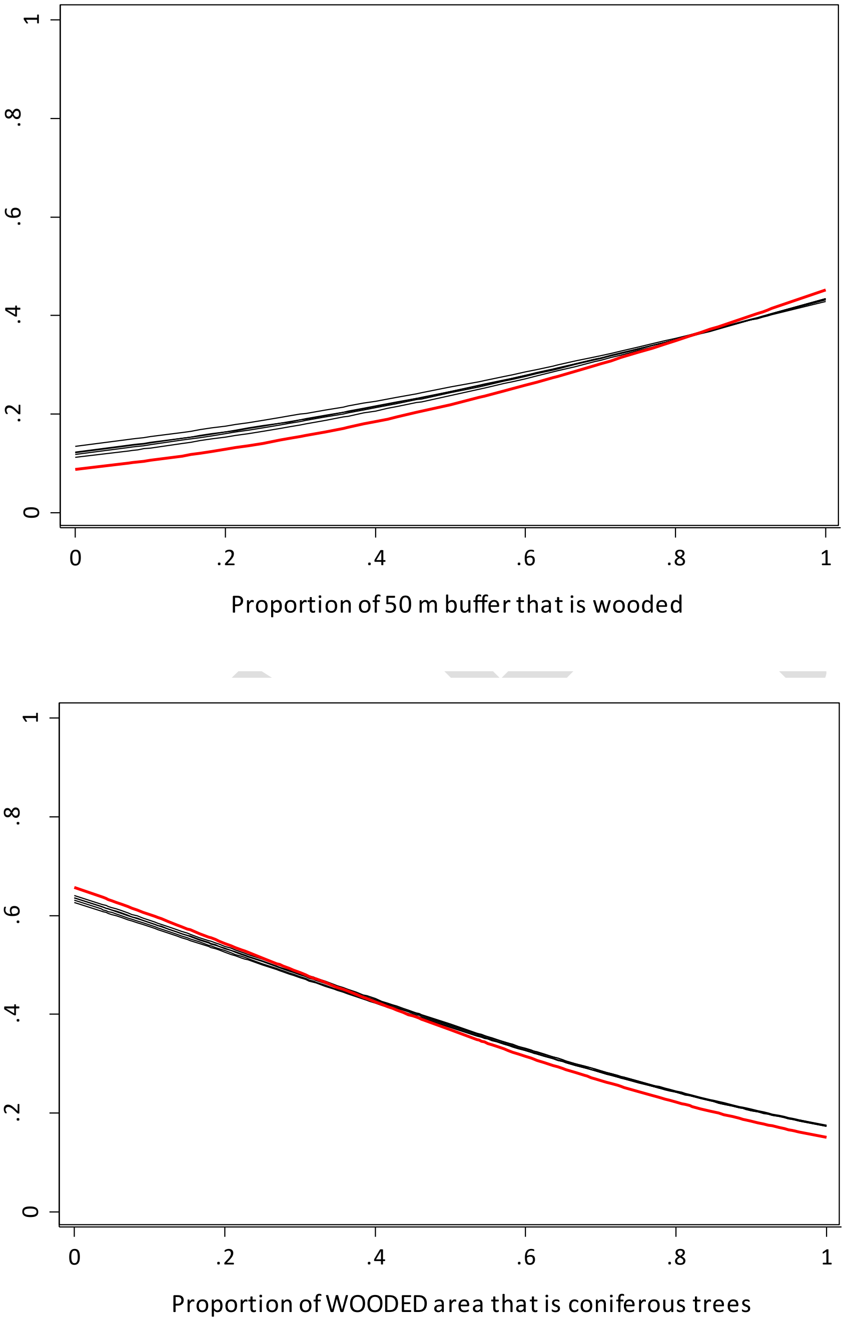

Figure 7 - Probability of observing an Ovenbird as a function of the proportion of a 50 metre buffer being WOODED (TOP) and the proportion of the WOODED area that was coniferous trees (BOTTOM). Each black line is the result from a model using 1 of 5 buffers (one centred on where observer stood and 4 in each of the cardinal directions but 50 metres from the observation point). The red line is the CME result that uses the 5 repeated measures of the covariates to account for “uncertainty” in spatial location of the Ovenbird relative to the observer.

Next Steps:

Dealing with the assumptions of CME:

The CME wrapper is limited in that it only allows the estimation of one covariate with measurement error (although multiple repeat measurements of the same covariate can be included). GLLAMM allows more than one covariate to be measured with error and to be effective in ecological modelling. I am working on learning the coding to do this which is quite complex. I am considering contacting the creator of GLLAMM and offering her authorship on a paper such as this if she can help me with learning the code.

The CME wrapper is also restrictive in that it assumes that the true covariate is assumed to have a normal distribution. I have not wrapped our heads around the implications this has for model estimation but this assumption can be relaxed in GLLAMM by using a nonparametric maximum likelihood.

Finally, there is an assumption that the repeated measurements of the covariates have identical measurement properties. This indicates that the repeated measures have the same mean (no relative bias) and the same measurement error variance. Since some of the GIS layers use “different” methods this may be false. This assumption can also be relaxed by using a model call the general congeneric measurement model which I am currently exploring how to create.

The CME wrapper approach allows the incorporation of offsets. As such I can incorporate BAM’s offset approach to deal with the nuisance variables and correct for point count methodology. This has NOT been incorporated in this analysis to reduce processing time.

Confirmation is required that the way I have estimated probability of observation via CME approach is correct.

Why does this matter?

In general, the examples I have shown result in similar qualitative patterns of Ovenbird response to WOODED and CONIFER. Thus, if the goal of the analysis is to determine there is a significant trend or pattern in relationship to a particular covariate, then use of CME may be unnecessary. However, I would argue that species – habitat models need to move past such rudimentary thinking and strive to estimate an absolute number that has meaning and value for managers who are charged with managing numbers of organisms not simple YES/ NO there is a relationship. Estimating density of organisms in a number of habitat strata is a key aspect of estimating the size of species population. Based on the models presented here our ability to do this accurately may be strongly influenced by the accuracy of our GIS layers as much or possibly even more than the approach used to measure species abundance. Future efforts need to evaluate how different stratification approaches used in such analyses influence population estimates to determine how important CME is as a tool for ecologists in accurate quantification of species abundance.

Testing alternative approaches for CME estimation:

I need to do more reading to determine how effective maximum likelihood CME estimation is relative to other techniques that deal with covariate error. I have found literature that uses the program Stata for covariate measurement error type approaches that includes:

A) Simulation Extrapolation is a simulation-based method aimed at reducing bias caused by the inclusion of covariates measured with error. Estimates are obtained by adding additional measurement error to the covariates. This resampling uncovers the trend of measurement error. Once the trend is estimated, final estimates are obtained by extrapolating back to the case of no measurement error. This could be done by computing the error between two GIS layers through a classification matrix and using that as the magnitude of measurement error to be simulated;

B) Regression Calibration – Same principle as Simulation Extrapolation but you input a reliability score into the model based on comparison of two replicates of the same covariate rather than directly estimate it using maximum likelihood;

C) Instrumental variables - This approach allows consistent estimation when the covariates are correlated with the error terms of a regression relationship. Such correlation may occur when the dependent variable causes at least one of the covariates ("reverse" causation), when there are relevant explanatory variables which are omitted from the model, or when the covariates are subject to measurement error. An instrument is a variable that does not itself belong in the explanatory equation but is correlated with the endogenous explanatory variables.

Literature Cited:

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2003). Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal*, 386-411.

Acknowledgement:

This draft report is a contribution of the Boreal Avian Modelling (BAM) Project, an international research collaboration targeting the ecology, management, and conservation of boreal birds. We acknowledge BAM's members, avian and biophysical Data Partners, and funding agencies (including the Joint Oil Sands Monitoring Program, Environment and Climate Change Canada, and the U.S. Fish & Wildlife Service), listed in full at <http://www.borealbirds.ca/index.php/acknowledgements>."