ABSTRACT

DIRENZO, JOSEPH ANTHONY. What’s All the Fuzz About? Effects of FIA Spatial Precision on the Performance of Species Distribution Models. (Under the direction of Ross Mentemeyer).

The Forest Inventory and Analysis (FIA) program of the United States Forest Service (USFS) provides geographically extensive measurements of forest structure and composition across the continental United State of America. FIA data are often used in species distribution models (SDMs) to relate species occurrence or abundance at known locations with environmental and/or spatial characteristics. However, in compliance with the Food Security Act of 1985 (H.R. 2100) as amended by the Department of the Interior and the Related Agencies Appropriations Act (H.R. 3423) in 2000, the United States Forest Service is required to protect the location of FIA plots and identity of the land owners who’s land may house an FIA plot. Therefore, spatial coordinates provided in the publically available FIA data are “fuzzed”. In this study, our objective was to determine whether fuzzed coordinates affected SDM results and assess the extent to which the spatial grain of environmental predictors matters (100 m to 1000 m, in 100 m increments). We modeled the occurrence and abundance of six species with small to broad environmental niches with environmental covariates extracted to either the true or fuzzed plot coordinates. For all species, the environmental covariates at each sample site did not significantly differ between the use of fuzzed and true data coordinates within grain size. There was little variability in prediction accuracy between fuzzed and true coordinate models of species occurrence or abundance across grain sizes for each species. The grain of environmental covariate data influenced precision but not accuracy, where many of the environmental covariates shared the same mean but not variance. With this research, we found that there are negligible downsides to using the publicly available fuzzed coordinate data instead of the true
coordinates in SDMs of large tree species and that future researchers can use finer scale (100m) grains with confidence.
What’s All the Fuzz About? Effects of FIA Spatial Precision on the Performance of Species Distribution Models.

by
Joseph Anthony DiRenzo

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science

Natural Resources

Raleigh, North Carolina

2016

APPROVED BY:

______________________________  ______________________________
Ross Meentemeyer                James McCarter
Committee Chair

______________________________
Robert Abt
BIOGRAPHY

Joe was born in Connecticut to a Colombian mother, Gloria Elena, and an Italian father, Federico Joseph, in 1988. A love of all things living inspired him to graduate from Southern Connecticut State University with a B.S. in Biology in 2011. Joe has volunteered saving sea turtles in Costa Rica, studied lupin in Iceland, and hugged oak trees in California. In the summer of 2015, while pursuing his graduate degree, he married Emily, the love of his life. Together they enjoy pondering the meaning of life, gazing at the countless stars in the night sky, and exploring the earth.
ACKNOWLEDGMENTS

I’d like to thank Ross Meentemeyer, my advisor, for taking a chance on. It was a pleasure working with him in the CAGIS lab at UNC at Charlotte and the Center for Geospatial Analytics at NCSU. I’d like to thank Jim McCarter and Bob Abt, for being on my committee and for the all the help and support. I’d like to thank my good friend Whalen and my sister Grace for all the help and support. Finally, I’d like to thank my mother Gloria, my brother Johnpaolo, and my loving wife Emily, for all the support you all have offered and given me over the past few years.

I couldn’t have done it without all of you. Thank you so much!
# TABLE OF CONTENTS

BIOGRAPHY ............................................................................................................................... ii

ACKNOWLEDGMENTS .................................................................................................................. iii

LIST OF TABLES ............................................................................................................................... v

LIST OF FIGURES ............................................................................................................................. vii

INTRODUCTION ............................................................................................................................... 1

METHODS ........................................................................................................................................ 3
  STUDY AREA ................................................................................................................................. 3
  SPECIES ......................................................................................................................................... 3
  OBSERVATIONAL DATA ................................................................................................................. 4
  PREDICTOR SURFACES ............................................................................................................... 4

EXPLORATORY ANALYSIS OF PREDICTORS .................................................................................... 4

MODELING ...................................................................................................................................... 5
  VARIABLE SELECTION .................................................................................................................. 5
  CROSS VALIDATION .................................................................................................................... 6

ACCURACY ASSESSMENT ................................................................................................................... 6

RESULTS ......................................................................................................................................... 7
  COMPARISON OF PARAMETER ESTIMATES BETWEEN TRUE AND FUZZ COORDINATES .................. 7
  ACCURACY OF PREDICTED SPECIES OCCURRENCE ................................................................. 8
  ACCURACY OF PREDICTED SPECIES ABUNDANCE .................................................................... 8

DISCUSSION ................................................................................................................................... 8

TABLES AND FIGURES ..................................................................................................................... 11

REFERENCES ................................................................................................................................... 19
LIST OF TABLES

Table 1. Description of environmental variables used in species distribution models........14

Table 2. Results of the paired t-tests between estimated parameter values of environmental covariates from true and fuzzed coordinates for each species. Each mini-table represents the species, the rows contain grain sizes, and the columns list the environmental variables. The cells within the tables list the corresponding p-values. Asterisks (*) indicate a statistically significant difference (* p<0.1, ** p<0.05, *** p<0.01). The abbreviations in the table follow as: PPT, mean total yearly precipitation (1981-2010); PSI, potential solar irradiance (calculated for the 183rd day (July 2nd) of 2014); TMIN, monthly mean of minimum daily temperature (1981-2010); TWI, topographic wetness index. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus Agrifolia; QUDO, Quercus douglasii; SESE, Sequoia semprevirens; UMCA, Umbellularia californica. .................................................15

Table 3. Results of the paired t-tests between 100 m true coordinate models (assumed to be the most accurate) and all other grain sizes and coordinate types for each species. Black text indicates comparisons between the 100 m true coordinate model output compared to other grain sizes using true coordinates. Red text indicates comparisons between the 100 m true coordinate model output compared to the output of models using fuzzed coordinates. Each mini-table represents the species, the rows contain grain sizes, and the columns list the environmental variables. The cells within the tables list the corresponding p-values. Asterisks (*) indicate a statistically significant difference (* p<0.1, ** p<0.05, *** p<0.01). Abbreviations in the table follow as: PPT, mean total yearly precipitation (1981-2010); PSI, potential solar irradiance (calculated for the 183rd day (July 2nd) of 2014); TMIN, monthly mean of minimum daily temperature (1981-2010); TWI, topographic wetness index. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica. .................16

Table 4. Accuracy metric means and standard deviation for models using true and fuzz coordinate data models at each resolution for each species. Black indicates TRUE model. Red indicates FUZZ model. Bold indicates greater accuracy. Asterisks (*) indicate a statistically significant difference (* p<0.1, ** p<0.05, *** p<0.01) in means of the accuracy metric between TRUE and FUZZ models for the species at the resolution. Abbreviations follow as: Occurrence accuracy metrics: AUC, area under curve; COMMISSION, false positive rate (type I error); OMISSION, false negative rate (type II error); TSS, true skill statistic. Abundance accuracy metrics: COR, Pearson correlation coefficient; RMSE, root mean square error, units in square meters (m²). PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus Agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica. ............................................................................17
Table 5. Accuracy metric means for true and fuzz coordinate models at each resolution for each species. Black indicates TRUE model. Red indicates FUZZ model. Bold indicates greater accuracy. Asterisks (*) indicate a statistically significant difference (* p<0.1, ** p<0.05, *** p<0.01) in means of the accuracy metric between the 100 m TRUE model and the corresponding TRUE or FUZZ model for the species at the resolution. Please refer to Table 3 for standard errors associated with accuracy metrics. Abbreviations follow as:

Occurrence accuracy metrics: AUC, area under curve; COMMISSION, false positive rate (type I error); OMISSION, false negative rate (type II error); TSS, true skill statistic.

Abundance accuracy metrics: COR, Pearson correlation coefficient; RMSE, root mean square error, units in square meters (m²). PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus Agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.
LIST OF FIGURES

Figure 1. Estimated parameter values for each environmental variable (along the rows) for each species (along the columns) presence data used with true and fuzz coordinate. Yellow represents model run with fuzz coordinated. Blue represents models run with true coordinates. Boxplots represent the interquartile range of estimated parameters around the line, which indicates the median parameter estimate. Points represent outlier values. The abbreviations in the figure follow as: PPT, mean total yearly precipitation (1981-2010), units in mm; PSI, potential solar irradiance (calculated for the 183rd day (July 2nd) of 2014), units in watt hours per square meter (WH/m$^2$); TMIN, monthly mean of minimum daily temperature (1981-2010), units in °C; TWI, topographic wetness index, unitless. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUOD, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica. ..............................................................11

Figure 2. Accuracy metrics for occurrence models with 10 fold cross validation. Species are along the columns, and accuracy metrics are along the rows. Yellow represents fuzzed coordinates and blue represents true coordinates. Boxplots represent the interquartile range of estimated parameters around the line, which is the median value. Points represent outlier values. Abbreviations in the figure are as follows: AUC, area under curve; COMMISSION, false positive rate (type I error); OMISSION, false negative rate (type II error); TSS, true skill statistic. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUOD, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica. ..................................................................................................................12

Figure 3. Accuracy metrics for abundance models with 10 fold cross validation. Species are along the columns, and accuracy metrics are along the rows. Yellow represents models run with fuzzed coordinates and blue represents models run with true coordinates. Boxplots represent the interquartile range of estimated parameters around the line, which is the median value. Points represent outlier values. Abbreviations in the figure follow as: COR, Pearson correlation coefficient; RMSE, root mean square error, units in square meters (m$^2$). PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUOD, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica. ........13
INTRODUCTION

The Forest Inventory and Analysis (FIA) program of the United States Forest Service (USFS) is a national treasure. The program provides geographically extensive measurements of forest structure and composition across the continental United States. Forests are divided into 6000-acre hexagonal areas (Brand 2004) with long-term survey plots established at random locations within each area (Reams et al. 2005). Plots are typically re-measured at five-year intervals to detect the emergence of forest health problems (e.g. invasive pests and pathogens) and monitor demographic changes in growth, mortality, and recruitment of new individuals. The broad geographic extent of FIA data make it especially applicable for species distribution models.

Species distribution models (SDMs) relate species occurrence or abundance at known locations with the environmental and/or spatial characteristics of those locations (Elith & Leathwick 2009). They are used for aquatic or terrestrial plants or animals. SDMs have a broad spectrum of implications, including predicting species occurrence (Iverson & Prasad 1998) and species abundance (Fei & Steiner 2007, Harris 1999, Iverson & Prasad 1998) at unmeasured locations, the potential spread of an exotic species into a new spatial domain (Morin 2003), and potential shifts in species distributions under projected climate change scenarios (Iverson & Prasad 2002, Woodall et al. 2009, Gibson 2013, Serra-Diaz et al. 2016).

To effectively model species distributions, it is important to consider what environmental factors constrain the species. Environmental variables may include precipitation, temperature, elevation, light, distance to water, distance to forest edge, intensity of urbanization or agriculture, and many more, including variations of these. Spatial scale, including extent (the spatial domain of the study) and grain (the grid-cell size of the explanatory or predictor data) is integral to developing SDMs, and should be chosen based on the goals of the study and the nature of the data. Ideally, the grain of the explanatory data should align with the spatial accuracy and precision of the observed species data (Dungan et al. 2002, Tobalske 2002), however, this may not always possible.

In compliance with the Food Security Act of 1985 (H.R. 2100) as amended by the Department of the Interior and the Related Agencies Appropriations Act (H.R. 3423) in 2000, the USFS is required to protect the location of FIA plots and identity of the landowners who’s land may house an FIA plot (www.fia.fs.fed.us). If precise plot coordinates are available, and overlaid on a surface containing parcel data and landowner information, it is possible to identify landowners to plots. Therefore, the precise coordinates of subplots are reported as existing up to 805m (½ mile) away from the true location of the plot in order to protect landowner privacy. In addition, attribute data from up to 20% of privately owned subplots are also substituted with attributes from similar subplots. These processes of moving coordinates and substituting attributes are termed “fuzzing” and “swapping”, respectively (www.fia.fs.fed.us). Fuzzing and swapping only occurs within county or supercounty (an aggregation of two or more adjacent counties in the same state) limits in order to maintain reliability of regional studies.

Since subplot coordinates are fuzzed up to 805m, scientists have hypothesized that using
environmental data of at least 800m resolution (or coarser) would not likely express the situational differences existing between true and fuzzed locations. Some analyses, however, may require the use of finer scale environmental data to be used in order to answer certain questions. Franklin et al. (2013) examined the differences of predicting climate refugia using fine and coarse grain predictors, and suggested that some climate refugia captured at finer scales may be missed using coarser scale data. In the same respect, studies such as that of Serra-Diaz et al. (2015), examining range shifts under climate change also require finer scale environmental data in order to adequately address the issues at hand. While FIA data is commonly used in SDM applications, it remains unclear if, how, and to what extent the use of fuzzed coordinates and correspondingly large resolution environmental data might affect analyses.

Several studies have examined the effects of fuzzed coordinates. Coulston et al. (2004) examined the influence of fuzzed plot locations on the accuracy of ordinary kriging and residual kriging estimates of forest biomass and found no statistically significant difference ($\alpha = 0.05$) between the accuracy of models developed with the true FIA plot locations and models developed with fuzzed plot locations. The study did note, however, that accuracy was related to the spatial autocorrelation of the data. In a later study, Coulston et al. (2006) systematically examined the effects of fuzzed coordinates of simulated data with varying degrees of spatial autocorrelation on ordinary kriging and linear regression across different resolutions (30m, 250m, 500m, 1000m, 2000m). Their study showed that fuzzed plot locations did not significantly influence the accuracy of kriging estimates, but in some situations linear regression model accuracy was significantly influenced (decrease in $R^2$) as grain size of predictors and SAC decreased. These results suggest that unless the independent variable has high spatial autocorrelation, only coarse spatial resolution data should be used to develop linear regression models (Coulston et al. 2006).

McRoberts et al. (2005) conducted a design-based study for circular areas of increasing radii centered at either precise or fuzzed coordinates, and then compared the two. Their results suggest the effects of using fuzzed coordinates are negligible at any radii greater than approximately 32km. Most recently, Gibson (2013) found no statistically significant differences in accuracy between using precise or fuzzed coordinates with predictor variables at 1km resolution.

The general objective of this study is to determine the best resolution to model FIA data for use in species distribution modeling when predicting either occurrence or abundance of a tree species. We systematically address key issues to evaluate:

1) Do predictor variables extracted from fuzzed coordinates significantly differ from those extracted from precise coordinates and does the spatial grain of the variables matter?
2) Do fuzzed data differently affect the accuracy of species occurrence and abundance models and do accuracies vary with spatial grain?
3) Are some plant life strategies affected by fuzzed plot coordinates more than others?
We hypothesize that:

a) On average, as the grain size of the environmental data used decreases, predictor values will differ more between those extracted at TRUE coordinates vs. those extracted at FUZZ coordinates.

b) On average, occurrence models built with TRUE coordinates will increase in accuracy as grain size of environmental data decreases; models built with FUZZ data will exhibit little to no change in accuracy as grain size of environmental data decreases.

c) On average, abundance models built with TRUE coordinates will increase in accuracy as grain size of environmental data decreases; models built with FUZZ data will exhibit little to no change in accuracy as grain size of environmental data decreases.

d) On average, models predicting occurrence or abundance for a specialist species will be more sensitive to grain size and coordinate fuzzing than models predicting occurrence or abundance for generalist species.

METHODS

STUDY AREA

This study focuses on the states of California and Oregon. California and Oregon comprise a combined area of 262,075 sq. mi. and house most of the California Floristic Province (CFP) which is home to a great number of endemic species. This area, with its Mediterranean climate, is a hotspot for plant biodiversity (Myers & Mittermeier, 2000) and its land use changes (Thorne et al., 2009) and fire regimes (Dolanc et al., 2014) make it an excellent area of interest to study changes in species occurrence and abundance. The topographically heterogeneous nature of the study area also lends itself well to examining the effects of grain size on environmental variables.

SPECIES

Three conifer tree species (*Pinus ponderosa*, *Pseudotsuga menziesii*, and *Sequoia sempervirens*) and three broad-leaved tree species (*Quercus Agrifolia*, *Quercus douglasii*, and *Umbellularia californica*) were chosen, representing specialist and generalist strategies across a broad geographical region. Of the three conifers, *P. ponderosa* and *P. menziesii* are generalist species, occurring throughout much of the study area with their true ranges extending well beyond the borders of California and Oregon. *S. sempervirens* is a specialist species, occurring along the coast from Monterey County in California to southern Oregon. Of the three broad leaf species, *Q. douglasii* is deciduous, and *Q. Agrifolia* and *U. californica* are evergreen. *S. sempervirens*, *Q. Agrifolia*, and *Q. douglasii*, and *U. californica* are endemic to the study area.
OBSERVATIONAL DATA

The FIA database was queried for plot, species, true coordinates, fuzzed coordinates, and species cumulative basal area data. These species occurrences were used as known presences. Any plot within 25km of a known species presence where the species did not occur was used as a known species absence (VanDerWal et al. 2009). This was done separately for true and fuzzed coordinates since the number of absences may change depending on the distance each plot is fuzzed. These dissolved 25km buffers around known species presences set the modeling extent for each species.

PREDICTOR SURFACES

Four predictor surfaces were used in the analysis. Table 1 describes each of these predictors. Precipitation (PPT) and mean minimum monthly temperature (TMIN) were developed by the Parameter-elevation Regressions on Independent Slopes Model (PRISM) group (Daly et al. 1994) at 800m resolution. These data were downsampled to 100m resolution by Alan Flint using methods described in Flint & Flint (2012). A 10m digital elevation model (DEM) of California and Oregon was acquired from the National Elevation Dataset (Gesch 2007, Gesch et al. 2002) from the United States Geological Survey (USGS). This DEM was resampled to 100m resolution and used to develop potential solar irradiance (PSI) and topographic wetness index (TWI) surfaces. PSI was calculated using the algorithm of Dubayah (1994) and methods developed by Rich et al. (Rich 1990, Rich et al. 1994) and further developed by Rich and Fu (2000, 2002). The formula is

$$PSI = \cos\theta\cos S + \sin\theta\cos(\Phi - A)$$

where $\theta$ is the solar zenith angle, $\Phi$ is the solar azimuth, $S$ is the slope gradient of the terrain, and $A$ is the slope aspect.

TWI was calculated using methods developed by Beven & Kirby (1979). The formula is

$$TWI = \ln\left(\frac{a}{\tan B}\right)$$

where $a$ is the specific catchment area ($a = A/L$, catchment area $A$ divided by contour length $L$) and $B$ is the local slope.

Using R (R Core Team 2013) and the raster package (Hijmans 2012), these 100m surfaces were each repeatedly resampled using bilinear interpolation to produce surfaces at resolutions of 200m, 300m, 400m, 500m, 600m, 700m, 800m, 900m, and 1000m.

EXPLORATORY ANALYSIS OF PREDICTORS

Each plot has TRUE and FUZZ coordinates associated with it. For each known species presence, predictor values were extracted from each of the four predictor surfaces at each of
the 10 resolutions at TRUE and FUZZ coordinates. Using a two tailed paired T test, the distributions of predictor values were compared for a statistically significant differences.

MODELING

Species distribution data were modeled in two ways. Species occurrence was modeled using a generalized linear model (GLM) with a binomial distribution assumed for the response variable. Species abundance was modeled as cumulative plot basal area using a GLM. The responses for each species were analyzed for spatial autocorrelation (SAC) in the model residuals using Moran’s I as an indicator for SAC.

In order to account for SAC in species occurrence, an autocovariate was calculated as a function of the response variable and was incorporated as a predictor variable. This was done with the spdep package (Bivand 2005) using methods developed by Augustin et al. (1996) and Gumpertz et al. (1997). By including an autocovariate, the GLM equation is altered from its usual form to

\[ Y = X\beta + \rho A + \epsilon \]

where \( \beta \) is a vector of coefficients for the intercept and explanatory variables \( X \), and \( \rho \) is the coefficient of the autocovariate \( A \). \( A \) as a weighted average can be calculated as

\[ A = \frac{\sum_{j \in k_i} w_{ij} y_j}{\sum_{j \in k_i} w_{ij}} \]

where \( y_j \) is the response value of \( y \) at site \( j \) among site \( i \)’s set of \( k_i \) neighbors; and \( w_{ij} \) is the weight given to site \( j \)’s influence over site \( i \). An inverse distance-weighting scheme was used to weight neighbors within a 25km radius.

To account for SAC associated with species abundance, a simultaneous autoregressive (SAR) error model was used. In this model, a term representing spatial structure is added to the normal linear equation and takes the form

\[ Y = X\beta + \lambda W \mu + \epsilon \]

where \( \lambda W \mu \) represents the spatial structure \( \lambda W \) in the spatially dependent error term \( \mu \), and \( \lambda \) is the spatial autoregression coefficient.

VARIABLE SELECTION

Each species response, using each set of coordinates at each resolution was modeled with all the available variables using the corresponding model (GLM with autocovariate for occurrence; SAR for abundance). Using the MuMIn package (Barton 2014) in R, all possible models were fit and ranked using the Akaike information criterion (AIC). Since a changing in AIC of two or less can be considered negligible, all models with a change in AIC of two or
less from the best model were averaged together (Duong 1984, Brockwell 1996), and those predictors present in the averaged model were the predictors selected for inclusion.

CROSS VALIDATION

Once the relevant variables were selected using all the available data, the accuracy of SDM predictions was assessed using a 10 fold cross validation technique. For each species response at each combination of coordinates and predictor resolution, the data was randomly partitioned into 10 groups. Nine of the groups (the training set) were used to fit the model, and the remaining group (the testing set) was used to assess the accuracy and error of the model. This was done 10 times, with a different group held aside each time as the testing set.

ACCURACY ASSESSMENT

Four accuracy metrics were used to assess occurrence models: the true skill statistic (TSS), area under the curve (AUC), commission rate (COMIS), and omission rate (OMIS).

The area under the receiver operator curve (AUC) is often used as a threshold-independent measure of model performance because it has been shown to be independent of prevalence (Manel, Williams, & Ormerod 2001; McPherson, Jetz, & Rogers 2004). In presence/absence data, an AUC value of 1 means that the data was perfectly classified by the model; therefore, an AUC score of 0.5 means that model performs as well as a random guess (Philips et al. 2006).

A popular measure for presence-absence SDMs is Cohen’s Kappa; Kappa ranges from -1 to +1. A statistic of +1 indicates perfect agreement between model predictions and observations, whereas a statistic of zero or less indicates performance no better than random (Cohen 1960). The kappa statistic also accounts for both omission (false negative rate) and commission (false positive rate) errors (Manel, Williams, & Ormerod 2001). Kappa, however, has a tendency to be dependent on prevalence, which can bias its measure of accuracy (Cicchetti & Feinstein 1990; Byrt, Bishop & Carlin 1993; Lantz & Nebenzahl 1996). The true skill statistic (TSS) corrects for this dependence while keeping all of the advantages of kappa (Allouche, Tsoar & Kadmon 2006). It is calculated as

\[
TSS = \frac{a}{a + c} + \frac{d}{b + d} - 1
\]

where \(a\) is the number of cells for which presence was correctly predicted by the model, \(b\) is the number of cells for which the species was not found but the model predicted presence, \(c\) is the number of cells for which the species was found but the model predicted absence, and \(d\) is the number of cells for which absence was correctly predicted by the model.

Since TSS accounts for both omission and commission already, it is a good indicator of overall model performance. Omission and commission were still calculated and compared in order to possibly assess where more of the model error or decrease in accuracy sourced from.
Two accuracy metrics were used to assess abundance models: Pearson’s correlation coefficient (COR), and root mean square error (RMSE). Pearson’s correlation coefficient is an indicator of how closely the observed and predicted values agree in relative terms (Potts & Elith 2006). It ranges from +1 (perfect positive correlation) to -1 (perfect negative correlation). A perfect correlation, however, does not indicate exact predictions. All predictions may be biased in a consistent direction.

Root mean square error (RMSE) is a commonly used metric that measures the difference between predicted values and observed values. RMSE depends on sample size and the agreement between the observed and predicted values. It is calculated as

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}
\]

where \(n\) is the sample size, \(y_i\) are the observed values, and \(\hat{y}_i\) are the predicted values. It is in the units of the observed and predicted values.

For each species, at each grain size, using each set of coordinates, all accuracy metrics were calculated (TSS, AUC, OMIS, and COMIS for occurrence models; COR and RMSE for abundance models). Accuracy metrics from models using true coordinates were compared to accuracy metrics using fuzzed coordinates. A t-test was used to test for a statistically significant difference between true and fuzzed metrics at each grain sizes for each species, for each accuracy metric. T-tests were also used to compare accuracy metrics calculated from the models built with environmental data extracted at 100m grain using the true coordinates (most accurate data and models used in this study) to all other grain sizes and coordinates for that combination of species and metric.

RESULTS

COMPARISON OF PARAMETER ESTIMATES BETWEEN TRUE AND FUZZ COORDINATES

The estimated parameter values for environmental covariates for species (presence data) differed among grain sizes, but they did not differ between the use of fuzz and true data coordinates within grain size (Figure 1; Table 2). Both PSI and TWI had narrower estimated parameter values as grain size increased for all species, whereas the estimated parameter values for both PPT and TMIN remained constant across grain sizes for all species. Meanwhile, across all species median parameter values for all environmental variables remained fairly constant (Figure 1).

Assuming that the parameter values for environmental covariates was the most accurate at 100 m (the finest grain size used) using the true coordinate data, we compared all other grain sizes for each species and environmental variable to that standard. We found few significant


differences between environmental parameter estimates of other grain sizes or those using fuzz coordinates (Table 3).

**ACCURACY OF PREDICTED SPECIES OCCURRENCE**

We found little variability in the accuracy metrics (TSS, AUC, COMIS, OMIS) of predicted species occurrence across grain sizes for each species (Figure 2; Table 4). There was higher precision in accuracy of true coordinate models than the fuzz models (i.e., wider confidence intervals; Figure 2), but no large differences in median accuracy metrics between true and fuzz coordinate models.

Again assuming that species occurrence predictions are the most accurate at 100 m (the finest grain size used) using the true coordinate data, we compared all other grain sizes for each species and accuracy metric to that standard. We found almost no significant difference between the accuracy of the 100m true coordinate model to all other grain sizes and coordinates for that species, for that metric (Table 5). The exception was *U. californica*, which showed a significant difference between the 100m true coordinate model and models at 500, 900, and 1000 m resolutions under two of the four accuracy test statistics (OMIS and TSS; Table 5).

**ACCURACY OF PREDICTED SPECIES ABUNDANCE**

We found little variability in the accuracy (COR, RMSE) of predicted species abundance across grain sizes for each species using different accuracy metrics (Figure 3; Table 4). We found no significant differences between the predicted species abundance between true and fuzz coordinate models based on either of the accuracy metrics (Table 4).

Assuming predictions are the most accurate at 100 m (the finest grain size used) using the true coordinate data, we compared all other grain sizes for each species against that standard. We found almost no significant difference between the accuracy of the 100m true coordinate model to all other grain sizes and coordinates for that species, for that metric (Table 5). The exception was *P. menziesii*, which showed a significant difference in the COR accuracy statistic between the 100 m true coordinate model and predictions at 700 m and 1000 m resolution.

**DISCUSSION**

We found that the effects of using publicly available fuzzed coordinates of plant species presence data to develop species distribution models at resolution of 100 m to 1000 m (in 100 m increments) did not produce significantly different results than if the true coordinates were used to construct models. These results verify that scientific integrity is not lost even when precise geographic coordinates are not provided, and is congruent with the results of Gibson *et al*. 2015. We found that if scientists are interested in the effect of environmental covariates on species presence, then there are slight differences in parameter precision but not accuracy.
Meanwhile, if scientists are interested in the predictive capabilities of occurrence or abundance models using fuzzed data, we found that models performed equally as well across grain sizes. To our knowledge, this is the first study to systematically examine the effects of using true versus fuzzed FIA data in SDMs with varying grain sizes of environmental predictor variables.

The estimated values for environmental covariates of each species (presence data) did not differ between the use of fuzzed and true data coordinates within grain size, likely because environmental variability follows laws of biogeography, where closer sites are more similar and further sites are more dissimilar. Fuzzed coordinate data at most differ to true coordinate data by 805 m, making it likely that environmental covariates, such as rainfall, temperature, and humidity, do not differ significantly among nearby locations. Differences in environmental covariates are likely even less of a problem at larger grain sizes, where fine scale environmental variation is averaged and lost (Franklin et al. 2013).

Both PSI and TWI had narrower estimated parameter values as grain size increased for all species due to the aggregation of cells into larger grain sizes. These two predictor surfaces were developed scaling a 10 m DEM from USGS to larger grain sizes. Meanwhile, the precipitation and temperature predictors were acquired as downscaled versions of 800 m PRISM data, so models run with finer resolutions (< 800 m) had aggregations of cells with similar covariate values. Because of this, we don’t really know what the true local (100 m) measurement for temperature or precipitation would be. Interestingly, both true and fuzzed coordinate models responded similarly to the scaling of covariates across grain sizes, suggesting that modifications to environmental covariate resolution on grain size is more important than geographic coordinate accuracy on grain size. Since many of the environmental covariates shared the same mean but not variance, the resolution of environmental covariate data influenced model precision more than accuracy.

There was little variability in the accuracy of predicted species occurrence and abundance across grain sizes for each species between fuzzed and true coordinate models. This may be due to the small variation of the environmental covariates across models, resulting in little variation in model outputs, and therefore small differences in accuracy. The second reason model accuracy did not vary across grain sizes for each species between fuzzed and true coordinate models is that the spatial autocorrelation, even in a heterogeneous landscape, may be sufficient even at the 100 m grain size to not exhibit significant changes in model output. From our data and modeling approach, we do not support the hypothesis that only environmental data of 800 m grain size or larger should be used in order to suppress the effects of fuzzed coordinates.

For the first time, we have demonstrated in a systematic way that the parameter estimates and model accuracy for occurrence and abundance data across several plant species do not differ between models using fuzzed or true coordinate FIA data. We found that there are negligible negative effects in terms of model accuracy from using publicly available fuzzed coordinates data instead of the true coordinates for SDMs of tree species. This is good news for the many
research projects conducted using publically available FIA data with the fuzzed coordinates. Our results suggest that fine resolution predictors (i.e., 100 m) can and should be used confidently. This study was performed using six adult tree species with varying life-histories. Further research should include investigating the effects of fuzzed and true coordinates across grain sizes for different life stages (i.e., seeds, seedlings, sub-adults) and additional species of plants (i.e., herbaceous, under-story, etc.) with other life-history included in the FIA database. Another interesting approach to this would be to assess the difference in total area predicted at different scales. Additional modeling approaches should also be investigated to determine if outcomes from different methods are consistent. In general, this study indicates that researchers can rest assured that their studies using coarse grain predictors and FIA data are ok, and future researchers can use fine scale (100 m) surfaces with confidence.
Figure 4. Estimated parameter values for each environmental variable (along the rows) for each species (along the columns) presence data used with true and fuzz coordinate. Yellow represents model run with fuzz coordinated. Blue represents models run with true coordinates. Boxplots represent the interquartile range of estimated parameters around the line, which indicates the median parameter estimate. Points represent outlier values. The abbreviations in the figure follow as: PPT, mean total yearly precipitation (1981-2010), units in mm; PSI, potential solar irradiance (calculated for the 183rd day (July 2nd) of 2014), units in watt hours per square meter (WH/m²); TMIN, monthly mean of minimum daily temperature (1981-2010), units in °C; TWI, topographic wetness index, unitless. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.
Figure 5. Accuracy metrics for occurrence models with 10 fold cross validation. Species are along the columns, and accuracy metrics are along the rows. Yellow represents fuzzed coordinates and blue represents true coordinates. Boxplots represent the interquartile range of estimated parameters around the line, which is the median value. Points represent outlier values. Abbreviations in the figure are as follows: AUC, area under curve; COMMISSION, false positive rate (type I error); OMISSION, false negative rate (type II error); TSS, true skill statistic. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziessii; QUAG, Quercus agrifolia; QUOD, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.
Figure 6. Accuracy metrics for abundance models with 10 fold cross validation. Species are along the columns, and accuracy metrics are along the rows. Yellow represents models run with fuzzed coordinates and blue represents models run with true coordinates. Boxplots represent the interquartile range of estimated parameters around the line, which is the median value. Points represent outlier values. Abbreviations in the figure follow as: COR, Pearson correlation coefficient; RMSE, root mean square error, units in square meters (m$^2$). PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.
Table 1. Description of environmental variables used in species distribution models.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPT</td>
<td>30-year average of monthly mean precipitation (1980-2010)</td>
<td>mm</td>
</tr>
<tr>
<td>PSI</td>
<td>potential solar irradiance</td>
<td>WH/m$^2$</td>
</tr>
<tr>
<td>TMIN</td>
<td>30-year average of monthly mean minimum daily temperature (1980-2010)</td>
<td>°C</td>
</tr>
<tr>
<td>TWI</td>
<td>topographic wetness index</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Results of the paired $t$-tests between estimated parameter values of environmental covariates from true and fuzzed coordinates for each species. Each mini-table represents the species, the rows contain grain sizes, and the columns list the environmental variables. The cells within the tables list the corresponding $p$-values. Asterisks (*) indicate a statistically significant difference (* $p<0.1$, ** $p<0.05$, *** $p<0.01$). The abbreviations in the table follow as: PPT, mean total yearly precipitation (1981-2010); PSI, potential solar irradiance (calculated for the 183rd day (July 2nd) of 2014); TMIN, monthly mean of minimum daily temperature (1981-2010); TWI, topographic wetness index. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus Agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Species = PIPO</th>
<th>Species = PSME</th>
<th>Species = QUAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>PPT</td>
<td>PSI</td>
<td>TMIN</td>
</tr>
<tr>
<td>200</td>
<td>0.72</td>
<td>0.94</td>
<td>0.61</td>
</tr>
<tr>
<td>300</td>
<td>0.58</td>
<td>0.80</td>
<td>0.62</td>
</tr>
<tr>
<td>400</td>
<td>0.95</td>
<td>0.60</td>
<td>0.62</td>
</tr>
<tr>
<td>500</td>
<td>0.68</td>
<td>0.72</td>
<td>0.54</td>
</tr>
<tr>
<td>600</td>
<td>0.67</td>
<td>0.51</td>
<td>0.02 **</td>
</tr>
<tr>
<td>700</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>800</td>
<td>0.89</td>
<td>0.78</td>
<td>0.60</td>
</tr>
<tr>
<td>900</td>
<td>0.64</td>
<td>0.75</td>
<td>0.89</td>
</tr>
<tr>
<td>1000</td>
<td>0.53</td>
<td>0.96</td>
<td>0.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Species = QUDO</th>
<th>Species = SESE</th>
<th>Species = UMCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>PPT</td>
<td>PSI</td>
<td>TMIN</td>
</tr>
<tr>
<td>200</td>
<td>0.78</td>
<td>0.55</td>
<td>0.98</td>
</tr>
<tr>
<td>300</td>
<td>0.76</td>
<td>0.08 *</td>
<td>0.97</td>
</tr>
<tr>
<td>400</td>
<td>0.91</td>
<td>0.10 *</td>
<td>0.81</td>
</tr>
<tr>
<td>500</td>
<td>0.93</td>
<td>0.69</td>
<td>0.48</td>
</tr>
<tr>
<td>600</td>
<td>0.84</td>
<td>0.11</td>
<td>0.77</td>
</tr>
<tr>
<td>700</td>
<td>0.31</td>
<td>0.63</td>
<td>0.96</td>
</tr>
<tr>
<td>800</td>
<td>0.75</td>
<td>0.60</td>
<td>0.83</td>
</tr>
<tr>
<td>900</td>
<td>0.40</td>
<td>0.24</td>
<td>0.84</td>
</tr>
<tr>
<td>1000</td>
<td>0.75</td>
<td>0.08 *</td>
<td>0.39</td>
</tr>
<tr>
<td>1000</td>
<td>0.95</td>
<td>0.14</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Table 3. Results of the paired $t$-tests between 100 m true coordinate models (assumed to be the most accurate) and all other grain sizes and coordinate types for each species. Black text indicates comparisons between the 100 m true coordinate model output compared to other grain sizes using true coordinates. Red text indicates comparisons between the 100 m true coordinate model output compared to the output of models using fuzzed coordinates. Each mini-table represents the species, the rows contain grain sizes, and the columns list the environmental variables. The cells within the tables list the corresponding $p$-values. Asterisks (*) indicate a statistically significant difference (** $p<0.01$, *** $p<0.05$, *** $p<0.001$).

Abbreviations in the table follow as: PPT, mean total yearly precipitation (1981-2010); PSI, potential solar irradiance (calculated for the 183$^{rd}$ day (July 2$^{nd}$) of 2014); TMIN, monthly mean of minimum daily temperature (1981-2010); TWI, topographic wetness index. PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus agrifolia; QUOD, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Species = PIPO</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPT</td>
<td>PSI</td>
<td>TMIN</td>
<td>TWI</td>
</tr>
<tr>
<td>100</td>
<td>0.72</td>
<td>0.94</td>
<td>0.61</td>
<td>0.15</td>
</tr>
<tr>
<td>200</td>
<td>0.60</td>
<td>0.62</td>
<td>0.65</td>
<td>0.17</td>
</tr>
<tr>
<td>300</td>
<td>0.53</td>
<td>0.63</td>
<td>0.67</td>
<td>0.25</td>
</tr>
<tr>
<td>400</td>
<td>0.45</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
<tr>
<td>500</td>
<td>0.38</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
<tr>
<td>600</td>
<td>0.35</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
<tr>
<td>700</td>
<td>0.32</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
<tr>
<td>800</td>
<td>0.30</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
<tr>
<td>900</td>
<td>0.28</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
<tr>
<td>1000</td>
<td>0.26</td>
<td>0.68</td>
<td>0.68</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Species = QUAG</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPT</td>
<td>PSI</td>
<td>TMIN</td>
<td>TWI</td>
</tr>
<tr>
<td>100</td>
<td>0.73</td>
<td>0.66</td>
<td>0.62</td>
<td>0.51</td>
</tr>
<tr>
<td>200</td>
<td>0.60</td>
<td>0.74</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>300</td>
<td>0.53</td>
<td>0.77</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>400</td>
<td>0.46</td>
<td>0.78</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>500</td>
<td>0.40</td>
<td>0.80</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>600</td>
<td>0.35</td>
<td>0.83</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>700</td>
<td>0.30</td>
<td>0.86</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>800</td>
<td>0.28</td>
<td>0.90</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>900</td>
<td>0.26</td>
<td>0.94</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>1000</td>
<td>0.24</td>
<td>0.99</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Species = SESE</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPT</td>
<td>PSI</td>
<td>TMIN</td>
<td>TWI</td>
</tr>
<tr>
<td>100</td>
<td>0.82</td>
<td>0.96</td>
<td>0.51</td>
<td>0.85</td>
</tr>
<tr>
<td>200</td>
<td>0.70</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>300</td>
<td>0.58</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>400</td>
<td>0.46</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>500</td>
<td>0.35</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>600</td>
<td>0.26</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>700</td>
<td>0.20</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>800</td>
<td>0.14</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>900</td>
<td>0.10</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>1000</td>
<td>0.05</td>
<td>0.79</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Table 4. Accuracy metric means and standard deviation for models using true and fuzz coordinate data models at each resolution for each species. Black indicates TRUE model. Red indicates FUZZ model. Bold indicates greater accuracy. Asterisks (*) indicate a statistically significant difference (* p<0.1, ** p<0.05, *** p<0.01) in means of the accuracy metric between TRUE and FUZZ models for the species at the resolution. Abbreviations follow as: Occurrence accuracy metrics: AUC, area under curve; COMMISSION, false positive rate (type I error); OMISSION, false negative rate (type II error); TSS, true skill statistic. Absence accuracy metrics: COR, Pearson correlation coefficient; RMSE, root mean square error, units in square meters (m²). PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus Agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellularia californica.
Table 5. Accuracy metrics mean for true and fuzz coordinate models at each resolution for each species. Black indicates TRUE model. Red indicates FUZZ model. Bold indicates greater accuracy. Asterisks (*) indicate a statistically significant difference (* p<0.1, ** p<0.05, *** p<0.01) in means of the accuracy metric between the 100 m TRUE model and the corresponding TRUE or FUZZ model for the species at the resolution. Please refer to Table 3 for standard errors associated with accuracy metrics. Abbreviations follow as:

**Occurrence accuracy metrics**: AUC, area under curve; COMMISSION, false positive rate (type I error); OMISSION, false negative rate (type II error); TSS, true skill statistic.

**Abundance accuracy metrics**: COR, Pearson correlation coefficient; RMSE, root mean square error, units in square meters (m²). PIPO, Pinus ponderosa; PSME, Pseudotsuga menziesii; QUAG, Quercus Agrifolia; QUDO, Quercus douglasii; SESE, Sequoia sempervirens; UMCA, Umbellaria californica.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Species + PIPO</th>
<th>Species + PSME</th>
<th>Species + QUAG</th>
<th>Species + QUDO</th>
<th>Species + SESE</th>
<th>Species + UMCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.28 ± 0.02</td>
<td>0.26 ± 0.02</td>
<td>0.26 ± 0.02</td>
<td>0.23 ± 0.01</td>
<td>0.26 ± 0.02</td>
<td>0.24 ± 0.02</td>
</tr>
<tr>
<td>200</td>
<td>0.26 ± 0.02</td>
<td>0.25 ± 0.02</td>
<td>0.26 ± 0.02</td>
<td>0.23 ± 0.01</td>
<td>0.25 ± 0.02</td>
<td>0.24 ± 0.02</td>
</tr>
<tr>
<td>300</td>
<td>0.25 ± 0.02</td>
<td>0.25 ± 0.02</td>
<td>0.26 ± 0.02</td>
<td>0.23 ± 0.01</td>
<td>0.25 ± 0.02</td>
<td>0.24 ± 0.02</td>
</tr>
<tr>
<td>400</td>
<td>0.25 ± 0.02</td>
<td>0.25 ± 0.02</td>
<td>0.26 ± 0.02</td>
<td>0.23 ± 0.01</td>
<td>0.25 ± 0.02</td>
<td>0.24 ± 0.02</td>
</tr>
</tbody>
</table>

...
REFERENCES


Bivand, R. 2005. spdep: spatial dependence: weighting schemes, statistics and models. - R package version 0.3- 17.


