Impact of compositional and configurational data loss on downscaling accuracy

Amy E. Frazier[1]* and Peter Kedron[2]

[1] Oklahoma State University, Stillwater, Oklahoma, USA
[2] Ryerson University, Toronto, Ontario, Canada

*Corresponding author: amy.e.frazier@okstate.edu

Abstract

Datasets collected at widely varying spatial scales are often merged to address questions related to global environmental change. Integrating data requires aggregation, which reduces data quality and introduces statistical biases collectively known as the Modifiable Areal Unit Problem (MAUP). These biases result from different forms of compositional and configurational data loss that occur during aggregation, but little is known about the relationship between data loss and MAUP biases for downscaling. This study uses the well-established process of landscape and surface metric scaling to examine how uncertainties related to the composition and configuration of land cover patterns propagate across scales when data are aggregated and ultimately impact downscaling results. Results suggest a link between compositional data loss and downscaling accuracy, particularly in the patch-based landscape paradigm. Further work is needed to determine if relationships exist between compositional and configurational data loss measures and downscaling error in the surface paradigm.

Keywords: scale and scaling; landscape ecology, spatial pattern metrics, heterogeneity, remote sensing

I. Introduction

In the age of data-driven science, diverse datasets collected at widely varying spatial scales are increasingly being merged to address questions related to global environmental change. However, integrating data collected at different spatial scales requires aggregation, which reduces data quality and introduces statistical bias. These biases, collectively known as the Modifiable Areal Unit Problem (MAUP), result from different forms of compositional and configurational data loss that occur during aggregation. MAUP biases can be minimized when aggregated data have a lower degree of heterogeneity (Holt et al. 1996; Steel and Holt 1996), but beyond this recognition, studies on the relationship between forms of data loss and MAUP biases remain limited.

Building on recent theoretical and methodological advancements in spatial science, remote sensing, and landscape ecology, we investigate how MAUP biases propagate across resolutions and whether the spread of bias, in the form of compositional and configurational data loss, can be used to forecast downscaling accuracy. Spatial pattern scaling provides a useful platform from which to examine these issues because research has established that several land cover pattern metrics exhibit consistent and robust scaling relationships across resolutions (Wu 2004). Yet, when these scaling relationships are extrapolated (i.e., downscaled) to predict metrics at a finer resolution, large errors typically result (Frazier 2014), likely from MAUP-driven aggregation biases (Frazier 2014, 2015a).

Until recently, examination of this hypothesis was limited to hard-classified landscape images, but the emergence of sub-pixel remote sensing classification techniques that preserve greater heterogeneity than their traditional, pixel-based counterparts have improved our ability to quantify data loss and preserve landscape heterogeneity, thereby opening the door for renewed investigations. Simultaneously, the proliferation of surface metrics in landscape ecology provides a means from which to compare this hypothesis across two different landscape paradigms: patch-based and surface. We use the well-established process of landscape and surface metric scaling (Turner et al. 1989; Wu 2004; Frazier 2015b) to examine how uncertainties related to the composition and configuration of land cover patterns propagate across scales when data are aggregated and ultimately impact downscaling results. Predetermination of the impact of these biases on downscaling may eventually allow assessment of whether a landscape is a satisfactory candidate for downscaling.

II. Methods

Data were collected through a geographically stratified sampling of forest land cover in four forested ecoregions in the eastern United States (Omernik 1987). Within each ecoregion, we randomly sampled 125 20x20km plots from a continuous grid (Fig. 1a,b). We removed grid squares comprising urbanized areas greater than 500,000 people. We then clipped the national land cover map (NLCD), aggregated to parent classes including a 'forest' class, and the tree canopy cover (TCC) product, both 30m resolution, to sample boundaries. Each NLCD and TCC plot was then aggregated to 60, and the 60m raster aggregated to 120, 180, 240, 420, and 480m using majority rules (NLCD) or mean (TCC) aggregation (Fig. 1c).

We computed analogous patch-based and surface metrics to measure downscaling accuracy (Table 1) and a suite of metrics from each paradigm that measure landscape composition and configuration (McGarigal et al. 2009). Patch-based metrics were computed using Fragstats (McGarigal et al. 2012) and surface metrics using SPIP software (Image Metrology).

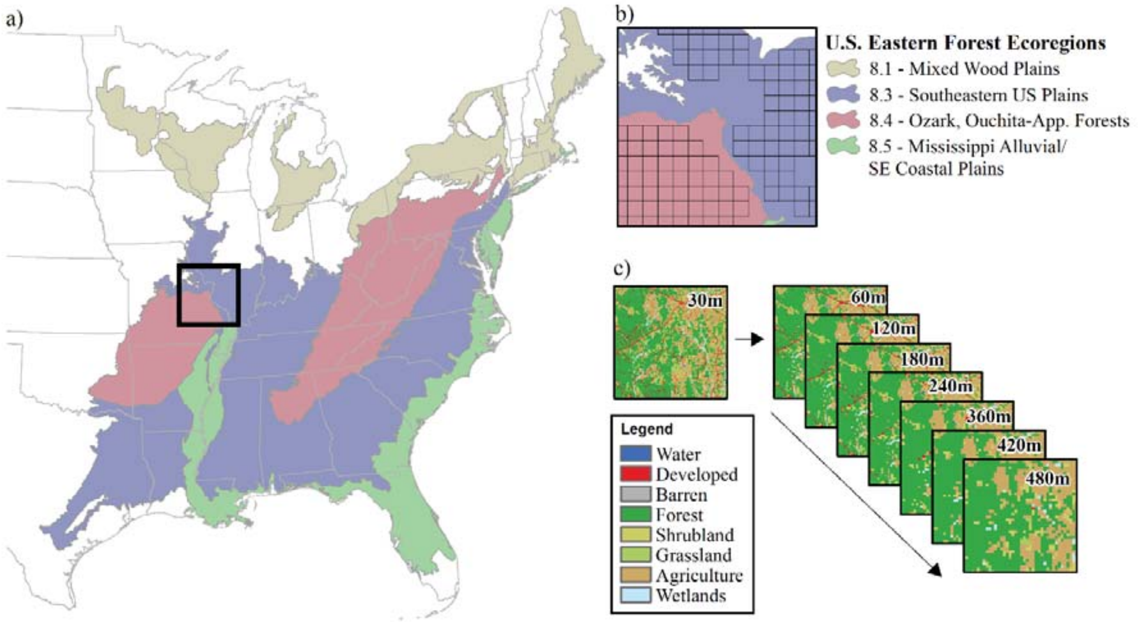
a)
b)
U.S. Eastern Forest Ecoregions
8.1 - Mixed Wood Plains
8.3 - Southeastern US Plains
8.4 - Ozark, Ouchita-App. Forests
8.5 - Mississippi Alluvial/ SE Coastal Plains
c)
30m
60m
120m
180m
240m
360m
420m
480m
Legend
Water
Developed
Barren
Forest
Shrubland
Grassland
Agriculture
Wetlands

Figure 1. (a) Eastern U.S. forest ecoregions, (b) grid sampling scheme, and (c) NLCD aggregation.

Reserving the 30m raster, a scalogram was plotted for PD and Sds for the seven coarse resolutions for each plot, and various scaling functions fit to the scalogram (i.e., power law, first-, second-, and third-order polynomials). Wu (2004) and Frazier (2015b) demonstrate these curve types accurately model metric behavior across scale. To generate a measure of MAUP-induced error, those curves were downscaled to predict each metric value at 30m and those predicted values compared to true 30m PD and Sds values through a measure of relative error (Eq. 1), where M_p is the predicted metric from the scaling function, and M_t is the true metric value. Low E_{rel} values indicate the scaling function accurately predicts the metric at a finer resolution. The curve type producing the lowest average E_{rel} for each metric was selected as the candidate model for further examination.

Table 1. Analogous metrics for measuring downscaling accuracy and data loss in the two paradigms.

Patch-Based	**Surface**
Downscaling metrics	
Patch Density (PD)	Peak Density (Sds)
Compositional data loss metrics	
Total area of forest (AREA)	--
Largest patch index (LPI)	Maximum peak height (Sp)
Configurational data loss metrics	
Edge Density (ED)	Peak-Peak range height (Sy)
Percent of Like Adjacencies (PLADJ)	Moran's I index (Smi)
Mean Fractal Dimension (FRAC)	Surface fractal dimension (Sfd)

To capture compositional data loss within the patch-based paradigm, we computed the linear rate of change of the data loss metrics (Table 1) across the seven coarse resolutions. Finally, we used conventional, ordinary least squares regression to measure the influence of data loss on downscaling accuracy. Expressed as Equation 2, $\hat{y}_i$ is the estimated value of the dependent valiable E_{rel} for plot i, β_0 represents the intercept, and $\beta_{config(k)}$ and $\beta_{comp(j)}$ are coefficients for the independent variable x_{ik} and x_{ij}, our measures of either compositional or configurational data loss, and ε_i represents the error term. Independent regressions were completed for each ecoregion for each paradigm and a combined regression for all regions for each paradigm.

$$E_{rel}(\%) = \left|(M_p - M_t)/M_t\right| * 100 \quad (1)$$

$$\hat{y}_i = \beta_0 + \sum_k \beta_{config(k)}\, x_{ik} + \sum_k \beta_{comp(j)}\, x_{ij} + \varepsilon_i \quad (2)$$

III. Results and Discussion

Across the four ecoregions, a third-order polynomial model performed best for downscaling PD, and power law performed best for Sds, which is consistent with recent findings (Frazier 2015a). PD mean R^2 values were consistently 0.97. Model fit improved with the surface paradigm, and R^2 values were >0.99 for Sds (Table 2). Relative error statistics (Table 2) show average downscaling accuracies ranged from 28.3 to 39.8 for PD, and 27.7 to 39.3 for Sds. Thus, there is little correlation between model fit (R^2) and downscaling accuracy (E_{rel}), but results confirm prior findings that even when scaling relationships are strong and model fits are high, metric downscaling of aggregated data is not particularly accurate (Frazier 2014).

Table 2. Scaling model fit (R^2) and downscaling relative error (E_{rel}) across ecoregions and paradigms.

	PD				**Sds**			
Ecoregion	**8.1**	**8.3**	**8.4**	**8.5**	**8.1**	**8.3**	**8.4**	**8.5**
Mean R^2	0.97	0.97	0.97	0.97	0.998	0.996	0.997	0.998
Std. Dev. R^2	0.01	0.01	0.02	0.01	0.002	0.004	0.005	0.002
Mean E_{rel}	28.3	31.8	39.8	28.8	27.7	39.3	34.4	28.1
Std. Dev. E_{rel}	10.7	9.85	14.9	7.94	8.7	13.1	16.1	10.3
Sample (n)	124	125	125	125	124	125	125	125

Tables 3 and 4 summarize the results of OLS regression for the two paradigms. For the patch paradigm, the rate of AREA loss was consistently a significant predictor of relative downscaling error. ED was also significant for most models. Across all ecoregions, PLADJ was also significant. Model fit values ranged from 0.41 to 0.596. For the surface paradigm, models were poorly fit. No measures of compositional or configurational data loss were significant across all regions. Variance inflation factors indicated possible collinearity of variables. These findings suggest further examination of alternative metrics is necessary. Of the two models with reasonable fit, the rate of loss of maximum peak height (Sp), a measure of compositional loss, was the best predictor of downscaling accuracy.

Table 3. Impacts of data loss on downscaling accuracy of PD (patch paradigm).

Ecoregion	8.1	8.3	8.4	8.5	Total
Intercept	18.174***	34.948***	40.101***	36.102***	40.802***
AREA	0.666*	1.250***	2.243**	1.914***	1.361***
LPI	-0.243	0.678**	-0.644	-0.602	0.029
ED	-2.566***	1.351***	0.984	0.934**	1.609***
PLADJ	0.625	1.173***	2.745	0.558	1.184***
Diagnostics					
R^2	0.478	0.596	0.467	0.410	0.454
Sample (n)	124	125	125	125	499

*p < 0.10, **p<0.05, ***p<0.01; FRAC was not significant in any models

Table 4. Impacts of data loss on downscaling accuracy of Sds (surface paradigm).

Ecoregion	8.1	8.3	8.4	8.5	Total
Intercept	34.375***	52.849***	41.33***	30.780***	39.288***
Sp	1.083	14.762***	10.790***	1.251	5.371***
Sy	-0.281	-6.769	-0.335	-0.114	-1.093
Sfd	-1.414***	-2.969***	0.361	-0.716	-1.140***
Diagnostics					
R^2	0.064	0.246	0.264	0.024	0.112
Sample (n)	124	125	125	125	499

*p < 0.10, **p<0.05, ***p<0.01; Smi was not significant in any models

IV. Conclusions

Results suggest a link between compositional data loss and downscaling accuracy, particularly in the patch-based paradigm—the greater the rate of AREA loss during aggregation, the more difficult it becomes to predict metric values at a finer resolution. Further work is needed to determine if any relationships exist between compositional and configurational data loss measures and downscaling error in the surface paradigm. An initial step would be to further establish correspondences between patch and surface metrics within these and other ecoregions. Future work should also address whether ecological characteristics impact the success these measures.

Acknowledgements: This work is supported by a grant to the authors from the U.S. National Science Foundation (#1561021) for “Data Complexity and Spatial Scaling: Prediction Accuracy and Implications for Emerging Landscape Paradigms”.

References:

Frazier, A. E. (2014) A new data aggregation technique to improve landscape metric downscaling. *Landscape Ecology*, *29*(7), 1261-1276

Frazier, A.E. (2015a) Landscape heterogeneity and scale considerations for super-resolution mapping. *International Journal of Remote Sensing,*

Frazier, A. E. (2015b) Surface metrics: scaling relationships and downscaling behavior. *Landscape Ecology*, 1-13

Holt, D., Steel, D.G., Trammer, M., & Wrigley, N. (1996) Aggregation and Ecological Effects in Geographically Based Data. *Geographical Analysis* 28(3):244-261

McGarigal, K., Tagil, S., & Cushman, S. A. (2009). Surface metrics: an alternative to patch metrics for the quantification of landscape structure. Landscape Ecology, 24(3), 433-450

McGarigal, K., SA Cushman, and E Ene. 2012. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available at the following web site: http://www.umass.edu/landeco/research/fragstats/fragstats.html

Omernik, J.M. (1987) Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77(1), 118-125.

Turner, M. G., O'Neill, R. V., Gardner, R. H., & Milne, B. T. (1989). Effects of changing spatial scale on the analysis of landscape pattern. Landscape ecology, 3(3-4), 153-162

Steel, D. G., & Holt, D. (1996) Analysing and adjusting aggregation effects: the ecological fallacy revisited. *International Statistical Review/Revue Internationale de Statistique*, 39-60

Wu, J. (2004) Effects of changing scale on landscape pattern analysis: scaling relationships. *Landscape Ecology,* 19:125-138.