River Syst. Vol. 19/3, p. 207–224
Published online March 2011

Article

Untangling human development and natural gradients: Implications of underlying correlation structure for linking landscapes and riverine ecosystems

Yasmin Lucero*[1], E. Ashley Steel[2], Kelly M. Burnett[3] & Kelly Christiansen[3]

with 7 figures and 3 tables

Abstract: Increasingly, ecologists seek to identify and quantify relationships between landscape gradients and aquatic ecosystems. Considerable statistical challenges emerge in this effort, some of which are attributable to multicollinearity between human development and landscape gradients. In this paper, we measure the covariation between human development—such as agriculture and urbanization – and natural landscape gradients – such as valley form, climate and geology. With a dataset of wade-able streams from coastal Oregon (USA), we use linear regression to quantify covariation between human activities and landscape gradients. We show that the correlation between human development and natural landscape gradients varies dramatically with the scale of observation. Similarly, we show how the correlation varies by region, even within a scale of interest. We then use a simulation experiment to demonstrate how this inherent covariation can hinder statistical efforts to identify mechanistic links between landscape gradients and features of aquatic ecosystems. We illustrate the negative consequences of the underlying correlation structure for statistical efforts: inflated goodness-of-fit metrics and inflated error terms on key coefficients that may undermine model building. We conclude by discussing the current best statistical practices for dealing with multicollinearity as well as the limitations of existing statistical tools.

Keywords: multicollinearity, regression, scale, human impacts, streams, fish

Introduction

Statistical analysis over broad areas has proliferated in riverine ecology, motivated by diverse technological, ecological, and political drivers (Steel et al. 2010; Johnson & Host 2010). Commonly in these analyses, measures of human development (e.g., urbanization, agriculture, or roads) and measures of natural landscape gradients (e.g., elevation, geology, and climate) are used to infer mechanistic or predictive relationships with instream ecological targets (e.g., salmonid distribution, water quality, or macroinvertebrate community composition (Allan 2004a; Steel et al. 2010; Johnson & Host 2010)).

Results from such analyses have been applied in various contexts worldwide. Landscape features have been correlated with instream features to estimate the effect of landscape conditions on fish density (Creque et al. 2005), to model the effects of agriculture on macroinvertebrate communities (Lammert & Allan 1999), to prioritize conservation status of native fish species (Filipe et al. 2004), to identify restoration priorities in riparian buffers (Barker et al. 2006), and to predict potential salmonid abundance

Addresses of the authors:
[1]Northwest Fisheries Science Center, NOAA, 2725 Montlake Boulevard East, Seattle, Washington 98112 USA
[2]USDA Forest Service, Pacific Northwest Research Station, 400 North 34th Street, Suite 201, Seattle, Washington 98103 USA
[3]USDA Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, Oregon 97331 USA
*Corresponding author: yasmin.lucero@noaa.gov

www.schweizerbart.de
DOI: 10.1127/1868-5749/2011/019-0024
1868-5749/019-0024 $ 8.10

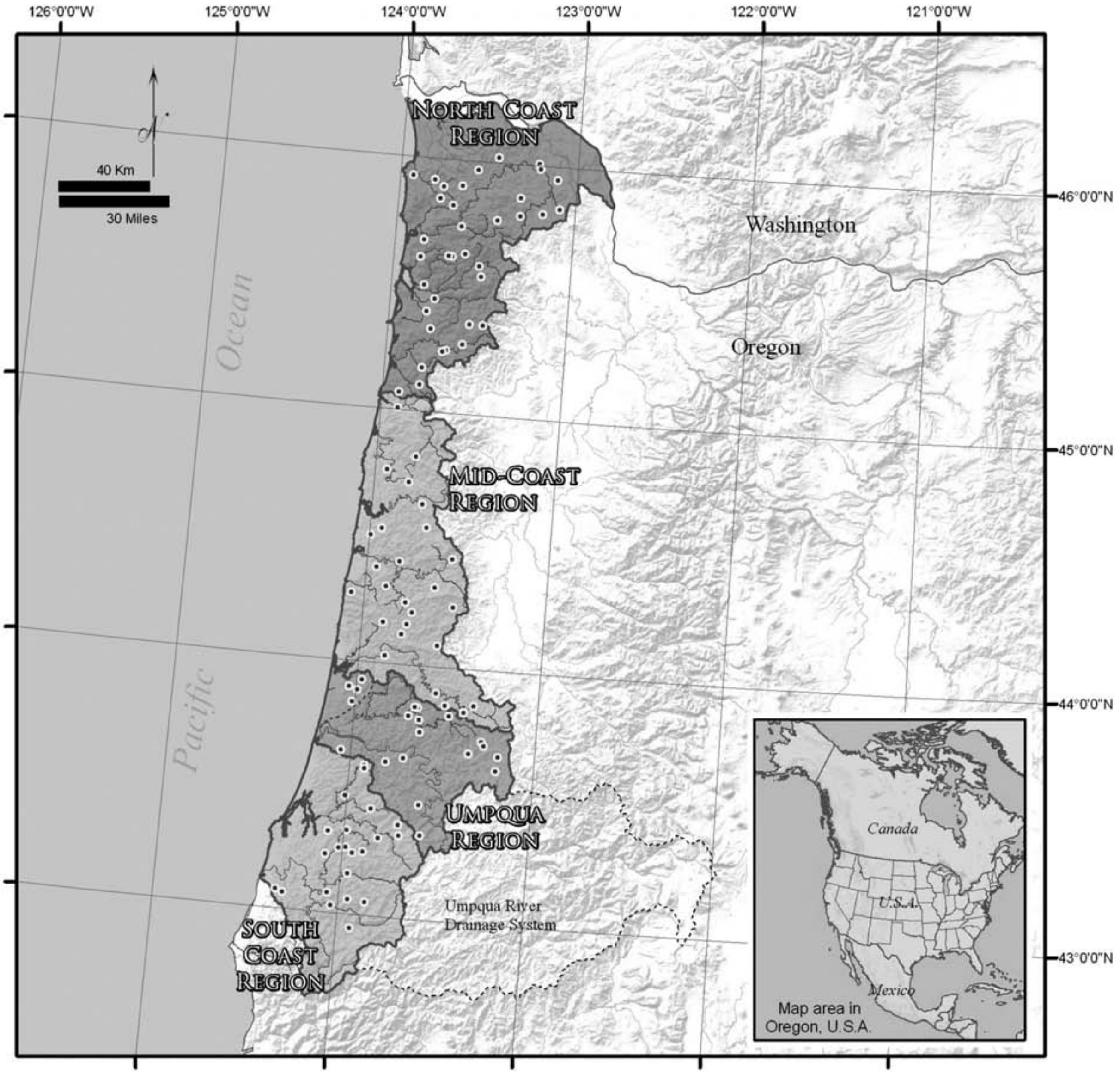

Fig. 1. The study area is the Oregon Coastal Province. The province is divided into four major management regions by the Oregon Department of Fish and Wildlife. Dots indicate study reaches.

in currently inaccessible areas (Steel et al. 2004). The evaluation of spatial patterns of land use and species distributions has also been applied to ecosystem mangement, such as the creation of conservation reserve networks (Margules & Pressey 2000; Thieme et al. 2007).

Though useful, these types of correlative analyses can be hindered by high degrees of covariation in landscape data, which can limit the ability to explain and predict conditions in stream ecosystems (Graham 2003). This problem has long been recognized in statistical analysis of streams (Johnson & Gage 1997; Allan 2004*a;* 2004*b*), but has yet to be carefully addressed. Here, we use data from the Oregon Coastal Province, USA (Fig. 1) to illustrate the complex nature of covariation in typical landscape datasets. In particular, we will demonstrate that correlations among variates are nonstationary with respect to scale and region. We will also use a simulation analysis to illustrate some of the specific consequences for building regression models. We conclude with a discussion of the implications and suggestions for interpreting models that link landscape conditions to instream responses.

Grappling with multicollinearity in landscape data

In one sense, we are discussing the well-known statistical problem of multicollinearity (Graham 2003; Legendre &

Legendre 1998). However, multicollinearity is a property of data; it does not describe the processes that produce correlation structure in data. In general, overcoming the limitations of multicollinearity depends on understanding the source of correlation structure (Grace & Bollen 2005).

Sources of multicollinearity

For this discussion, we organize the processes that produce multicollinear landscape data into three categories: redundant variates, causal links, and influential exogenous factors. In some datasets, multiple landscape variates may reflect closely related ecological features and thus are considered redundant. For example, a suite of metrics describing forest cover may be highly correlated because these are multiple measurements of a common underlying process. This category of redundant variates is well recognized in the statistical literature. Virtually any intermediate text on regression will issue the same advice for overcoming this source of multicollinearity: remove redundant variates, combine related variates into a common metric, or consider ridge regression (Mendenhall & Sincich 1996).

Causal links is our second source of multicollinear landscape data. The presence of causative relationships among ecological targets can create correlations in data. One example is spawner density and substrate size; these variates are mechanistically linked (fish select spawning habitat based in part on substrate size) and so the two are likely to be correlated (Baxter & Hauer 2000; McHugh & Budy 2004). When causal links are present, addressing multicollinearity depends on isolating and characterizing the mechanistic relationships and then modeling interactions using some form of multi-level model, such as a path analysis (Shipley 2002).

Our interest in this paper lies primarily with the third source of multicollinear landscape data, influential exogenous factors. Ecological factors excluded from the data can generate correlations among environmental variates that are included in the data. For example, in a fire dominant forest ecosystem, several metrics of tree cover are influenced by fire history. A dataset describing such a system would seldom include direct measures of fire; nonetheless the local history of fire is likely to be reflected in several of the variates, such as mean stand age and percent hardwood composition of the forest (Wimberly & Spies 2001). This shared dependence on fire history can generate correlations that are not attributable to direct mechanistic links among variates. Furthermore, the magnitude of these correlations will vary across locations with distinct fire histories, further undermining any attempt to interpret the correlations causally.

Here, we focus on human settlement as a particularly significant example of an influential exogenous factor. We posit that for rivers, many of the natural landscape gradients that influence patterns of human development also influence many aspects of aquatic ecosystems. As a result, we are at risk of observing correlations that are partially attributable to ecologically relevant mechanistic links, but are also partially attributable to the varying and influential structuring forces of human settlement and land use.

Natural gradients and human use of landscapes

The process of human settlement is shaped and informed by environmental gradients (Lyle 1999; Burgi et al. 2007). Over time, this generates correlations between natural gradients and patterns of human development and use. For example, rural residential development and agricultural land use are correlated with natural factors, such as soil type, geology, temperature, elevation, and precipitation (e.g., Kline et al. 2003; Kirch et al. 2004; Gudea et al. 2006). Instream ecological targets can be affected by these same natural landscape gradients, as well as patterns of human development and use (Hughes et al. 2006; Steel et al. 2010). For example, macroinvertebrate metrics vary in predictable ways with measures of human land use, such as urbanization and agriculture (Waite et al. 2010).

Separating anthropogenic from ecological effects may be particularly difficult in rivers, because in all but the most remote areas, rivers have long attracted humans (Postel & Carpenter 1997; Knox & Marston 2010). Thus, few rivers exist where anthropogenic and natural influences do not co-occur. To illustrate, of the reaches with the greatest natural potential to provide high-quality habitat for coho salmon (*Oncorhynchus kisutch*) in western Oregon, nearly 90 % are on private land and approximately half have adjacent riparian areas that have recently been logged or are managed for developed uses (urban, rural residential, or agriculture) (Burnett et al. 2007).

For this study, we use rural development and road density to demonstrate how human development and natural gradients can become conflated. Each of these metrics of human activity is correlated with immutable features of the landscape that also influence the formation of instream conditions, such as rock type, topography and climate.

Consequences of ignoring multicollinearity

While it is well-known that multicollinearity among predictor variates violates a core assumption of linear regression, it is nonetheless common to see multiple regression analysis using ecological datasets with extensive multicollinearity (Graham 2003; Whittingham et al. 2006).

One of our goals in this paper is to illustrate some of the consequences of ignoring multicollinearity. We conduct a simulation experiment to demonstrate that multicollinearity corrupts statistics that are frequently used in model building, such as R^2 and the standard error of coefficient estimates. Although this can be inferred from classical understanding of regression analysis (Petraitis et al. 1996; Belsley et al. 2004; O'Brien 2007), these consequences are not widely appreciated in ecology. By explicitly illustrating these problems for analysis, model fitting, and model building, we hope to encourage more cautious interpretation of regression results in landscape studies of streams.

We anticipate that regression analysis on multicollinear data will continue to be a useful, if limited, tool in applied ecology. Policy and legal mandates that apply over large spatial extents, e.g., the U.S. Endangered Species Act (16 USC section 1531, et seq. 1973) or the E.U. Water Framework Directive (OJL 327, 22 December 2000, pp. 1–73) demand broad-scale ecological assessments. As well, the relatively new field of landscape ecology has encouraged researchers in many disciplines, including river ecology, to measure, model, and predict over increasingly large areas. There are a great many ecological applications where the need for results is urgent and irrespective of the quality of the data. In most cases, regression analysis (i.e., all forms of regression, not only simple linear regression) is the best analytical method available for addressing applied questions.

Limited by both data quality and the resources available for analysis, applied ecologists will continue to proceed with imperfect regression analyses rather than no analyses. Given this practical reality, ecologists will benefit from detailed knowledge of the specific ways in which multicollinearity impacts our analyses, model fits, model building and interpretation. To this end, we illustrate some of the more subtle properties of multicollinearity in landscape data and which impact this correlation has on model fitting.

Methods

Study area

The study area is the Oregon Coastal Province on the west coast of North America (Fig. 1). The region is bordered to the west by the Pacific Ocean and to the east by the Willamette River valley and encompasses approximately 25,000 km^2. The climate is temperate maritime with mild, wet winters and warm, dry summers. The region is characterized by Coast Range sedimentary and volcanic geologies. Mountains dominate the area, except for a prominent, but geographically limited, coastal plain and interior river valleys. Montane areas are highly dissected, with drainage densities up to 8.0 km/ km^2 and elevations up to 1,250 m (Burnett et al. 2007).

Throughout the province, land cover is dominated by coniferous forests, and thus forestry is the dominant land use. Most land in the province is privately owned by the forest industry. However, one third of the province is publicly owned and managed, primarily by the state of Oregon and two federal agencies, the US Forest Service and the Bureau of Land Management. At lower elevations, agriculture is also an important land use (Ohmann & Gregory 2002; Spies et al. 2007). The province was divided into four major regions for monitoring the Oregon Plan for Salmon and Watersheds (http://www.oregon-plan.org/) (Fig. 1). The study area supports five Pacific salmonid species (*Oncorhynchus spp.*): coastal cutthroat trout (*O. clarkii*), steelhead (*O. mykiss*), and coho (*O. kisutch*), Chinook (*O. tschawytscha*), and chum (*O. keta*) salmon. Coho salmon in the province are listed as Threatened under the US Endangered Species Act (1973) and fish habitat for all species has been negatively affected human activities, including logging, channelization, log transport, and conversion of forested lands to agriculture (e.g., Thomas et al. 1993; IMST 2002).

Reach selection

This study uses data collected by the Oregon Department of Fish and Wildlife (ODF&W) through an integrated stream monitoring program (Jacobs & Cooney 1997; Firman & Jacobs 2001). The ODF&W identified monitoring reaches with two methods. In the first method, reaches were identified by a subjective selection of known high-quality spawning areas. These reaches varied between 0.8 km and 4.5 km in length (mean 1.8 ± 0.2 km). The second method used a generalized random tessellation stratified (GRTS) design to produce a spatially balanced random sample of sites to assess freshwater habitats and abundances of adult and juvenile coho salmon in wade-able streams of Oregon (Stevens & Olsen 2004). A reach was delineated around each identified site in the sample. Most reaches are approximately 1,000 m long, however reaches in small streams and outside the current distribution of coho salmon are shorter, approximately 500 m

To explore scale and region effects on correlation structure among landscape variates, we selected a subset of 281 reaches for which landscape data were available (Fig. 1). These 281 reaches are arrayed across four regions: 97 are in the North Coast region, 85 are in the Mid-Coast region, 35 are in the Umpqua region and 64 are in the South Coast region.

For the simulation analysis, we selected a subset of 122 reaches, which included only those identified by ODF&W for monitoring stream habitat characteristics. Stream habitat surveys were conducted during low-flow conditions between mid-June and late September. Specific instream measurements were obtained to describe the stream channel morphology and the physical structure of the valley and riparian areas. Detailed aquatic habitat survey methods are described by Moore et al. (2007). Reaches sampled for habitat are visited according to a rotating panel design to optimize statistical estimates for status and trend. The rotating panel design divides all reaches into four panels: those sampled annually, once every three years, once every nine years, and once every 27 years (Stevens & Olsen 2004). For the simulation analysis, we used only the reaches where habitat surveys were conducted annually or every three years. The habitat metrics included in the simulation analysis are averaged across time.

Landscape data

Our landscape variates were created from nine geospatial data layers describing topography, climate, lithology, land ownership, forest cover, timber harvest history, land use, and roads, (Table 1). These data layers are similar to those used in other riverscape studies to understand salmon habitat (Steel et al. 2004; Burnett et al. 2007; Van Sickle et al. 2008).

In a GIS (Environmental Systems Research Institute ArcMap v. 9.1), we defined an area of interest (AOI) at four spatial extents around each reach in our sample (Fig. 2). The smallest AOI is the buffer-100 m scale. This includes the area 100 *m* on either side of a study reach. The buffer is rounded at the ends of the reach. The next largest AOI is the buffer-500 m scale, which is defined in the same way, except this includes 500 m on either side of a reach. The watershed-scale AOI includes the entire upstream area draining into a reach. The final AOI is the HU scale. Hydrologic units (HU) are hierarchically nested areas used to divide and catalogue river systems in the United States (Seaber et al. 1987). The HU-scale AOI includes all of the 7th-code hydrologic units (Rickenbach et al. 2000) that are adjacent to a sampled reach. For 87 % of the 281 reaches used to explore scale and region effects, the HU-scale AOI is larger than the watershed-scale AOI.

We summarized the geospatial data layers across these AOIs. As a result, most of our variates are in units of percent-area. Some variates are in their native units such as average temperature range or number of miles of road, which are summarized as area-weighted means across the AOI (Table 1).

Variable selection and model building

We began with over 100 variates from the nine data layers in Table 1. We divided these variates into "predictor" and "response" variates, so that response variates reflect human land use activities and predictors are relatively immutable with respect to human development. From here, we aggregated landscape classes to create variates for analysis with smooth distributions and good coverage across sites. We next eliminated certain variates due to very high multicollinearity resulting from obvious functional redundancy.

We ultimately selected two response variates: percent area in rural development and road length (m), hereafter referenced as rural development and roads. Rural development was created by combining areas in land uses of agriculture and rural-residential (0.25–2.5 structures/ha) (Burnett et al. 2007). The resulting continuous variate for rural development remained highly zero-inflated, and so we transformed it to a binary variate (presence/absence) in which all AOIs with any rural development were labeled "present." We fit a logistic regression model to explain variation in rural development. The data for roads was skewed and so we log-transformed to fit a linear model with normal residuals. It should be noted that in our study area, road density is frequently related to logging activity and is thus not well correlated with rural development.

The best models for each variate were selected using forward stepwise regression with AIC. Initially, we found the best fit model with data aggregated across regions for each of the four scales for each predictor. We next found the best fit model across all regions at each scale (Tables 2–3). Forward stepwise regression is a very common method used in data mining applications but has been criticized in the context of explanatory modeling (Whittingham et al. 2006). Here, the method is appropriate because we are concerned with characterizing the existence of an underlying correlation structure among landscape variates and not uncovering explanatory relationships.

We also explored, but do not report, models for land ownership patterns and timber harvest history. The best models for land ownership overlapped substantially with our models for rural development. We were unable to find a strong model for harvest history because virtually all sites had substantial history of harvesting. All analyses were done with the R statistical computing language (R Development Core Team 2010).

Table 1. The primary dataset includes over 100 variates from 9 data layers. Of these, we selected a subset of variates to include in the forward selection process for the best fit models. We divided these variates into predictors and responses, and then found the best fit models to explain the responses variates.

Data Layer	Source	Included in model building	Range	Units	Notes
Topography	US Geological Survey 10 m DEMs	Elevation	0.6–561	m	Averaged across AOI
		Stream Gradient	0.11–13.20	% slope	Averaged across AOI
		Valley Index	1.9–100	% area	Index identifies polygons, or valleys, based on elevation in adjacent cells.
Climate	4000 m PRISM Climate data from 1961–1990 Daly et al. (1997)	Temperature range	16.4–25.8	°C	Difference between annual minimum and maximum air temperatures
		Precipitation	1077–4555	mm	Mean annual precipitation
Lithology	Walker (2003)	Mafic volcanic flows	0–100	% area	
		Sandstones	0–100	% area	
		Siltstones	0–100	% area	
Land Ownership	1:24,000 scale ODF (2004)	US Bureau of Land Management	0–100	% area	
		US Forest Service	0–100	% area	
		Private industrial forests	0–100	% area	
		Private non-industrial forests	0–100	% area	
		State	0–100	% area	
Forest Cover	30 m Ohmann & Gregory (2002)	Large Trees	0–46	% area	Predictive mapping of forest composition using direct gradient analysis and nearest neighbor imputation from satellite imagery and forest plot data. Thirty four original vegetation types were reduced and aggregated.
		Medium Trees	0–68	% area	
		Small Trees	0–48	% area	
		Hardwoods	0–90	% area	
		Remnants	0–62	% area	
Harvest History	30 m Lennartz (2005)	No timber harvest	0–100	% area	Timber harvest and burned areas were identified from satellite imagery
		Timber harvest	0–100	% area	
Land Use	30 m Burnett et al. (2007)	Agricultural	0–83	% area	
		Rural residential	0–82	% area	
		Urban	0–81	% area	
		Natural	0–38	% area	
		Forest	0–100	% area	
Roads	US Bureau of Land Management Ground Transportation Roads Publication (2009)	Road length	386–1,210,000	m	
Habitat	Oregon Department of Fish and Wildlife (2008)	Large Woody Debris	–1.3–4.7	log(m^3/m)	Volume of large wood per 100 m of channel length.

Results

We present the coefficient estimates for the best fit models for rural development (Table 2) and roads (Table 3). Although we calculated the best fit model for each unique combination of scale and region, we found no substantial differences in which predictor variates were retained. For simplicity, we present estimates for only the single overall best fit model. A subset of the coefficient estimates are also shown in Figures 3–4.

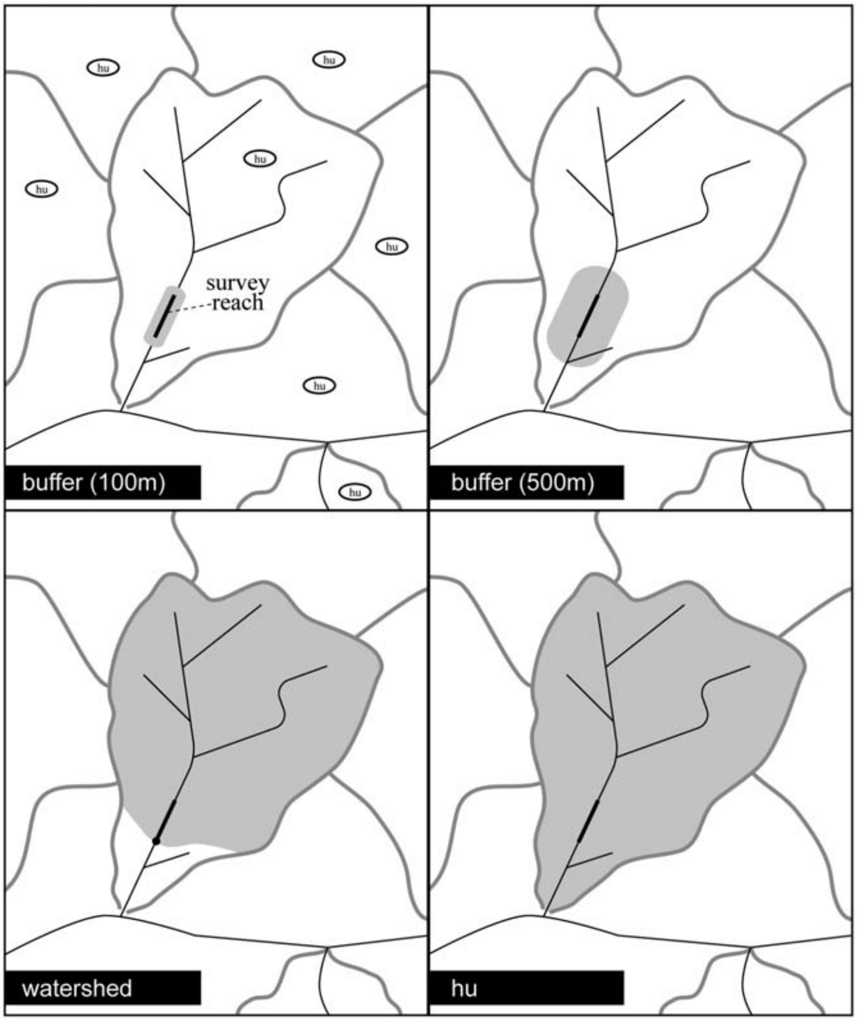

Fig. 2. Illustration of the scales of aggregation used to generate areas of interest (AOI). Landscape data are summarized across each of these areas. Some scale effects can be attributed to qualitative, as well as quantitative, differences among these areas.

Rural development

We found that the presence of rural development was predicted with reasonable accuracy – 67 % accurate on average across scales – using only a few environmental predictors: percent area in valley, annual temperature range, mean annual precipitation, and the percent area underlain by siltstone and by mafic volcanic flows (Table 2).

Overall, rural development in the Oregon Coastal Province is positively correlated with the presence of valleys and negatively correlated with precipitation and temperature; i.e., within our sample, rural development tends to occur selectively in broader river valleys and in areas that are both cooler and dryer. The best-fit model also retained some geological variates, but the correlation between rural development and geology was inconsistent across scales and regions.

While rural development and valley index are positively correlated at all scales and in all regions, the magnitude of that relationship varies considerably (Fig. 3). The relationship appears stronger when the data are aggregated at larger spatial scales. Additionally, the pattern across scales differs among the regions within our dataset. In particular, if the analysis included only the southern region, we would conclude a much stronger relationship than we would have concluded in any of the other regions (Fig. 3b).

Our final model also contained two metrics of climate: temperature and precipitation. The temperature range varies relatively little across the region (Table 1). However, mean annual precipitation ranges from quite wet to extremely wet. Initially, no pattern is apparent in these coefficients across scales, but with the data stratified by region, scale and region interact in some regions (Fig. 3d). In the North Coast region, temperature may become more significant with increasing scale, in the Mid-Coast region there appears to be no pattern, but in the South Coast and Umpqua regions, the coefficients decrease with scale. A

Table 2. Coefficient estimates and *c*-statistic for the single best fit model for rural development. Estimates are from a logistic regression of the presence of rural development on the listed variates. The *c*-statistic is a measure of how well the model describes the data. It is the expected percentage of cases where the model correctly predicts the presence or absence of rural human development; it varies between 50 % (no better than random) up to 100 % (perfect correspondence). Estimates are presented by scale and region.

							Lithology	
Scale	**Region**	**c (%)**	**Intercept**	**Valley Index**	**Precipitation**	**Temperature**	**Siltstone**	**Mafic Volcanic**
buffer 100 m	all	67	5.5	0.049	–7.4E-04	–0.0294	0.0018	–2.5E-03
buffer 500 m		67	8.6	0.172	–9.3E-04	–0.0428	0.0049	–4.1E-03
watershed		67	2.1	0.682	–7.7E-04	–0.0231	–0.0090	–4.8E-03
hu		65	8.9	0.461	–7.3E-04	–0.0463	0.0098	–9.3E-03
buffer 100 m	North	91	64.5	0.316	–2.1E-03	–0.3371	–0.0829	1.9E-02
	Mid-coast	67	5.7	0.056	–6.1E-04	–0.0320	–0.0029	–1.7E-02
	Umpqua	60	–11.7	0.024	2.2E-03	0.0320	–1.9206	NA
	South	75	5.3	0.052	–3.8E-03	–0.0077	0.0135	–3.1E-01
buffer 500 m	North	78	20.0	0.233	–1.8E-04	–0.1119	–0.0054	–3.3E-04
	Mid-coast	66	4.9	0.167	–4.9E-04	–0.0300	0.0168	–1.8E-02
	Umpqua	63	–2.6	0.163	7.3E-04	–0.0021	–0.4707	1.2E+00
	South	73	11.9	0.251	–4.2E-03	–0.0361	0.0089	3.6E-02
watershed	North	77	6.9	0.880	–3.0E-05	–0.0625	–0.0226	–5.9E-03
	Mid-coast	65	1.3	0.388	–1.3E-03	–0.0065	0.0328	–5.9E-03
	Umpqua	62	–5.6	0.191	9.9E-04	0.0110	–0.2846	1.9E+00
	South	70	2.9	0.794	–2.5E-03	–0.0149	–0.0039	6.5E-03
hu	North	73	13.4	0.644	–6.3E-04	–0.0757	0.0077	–8.4E03
	Mid-coast	58	7.4	0.176	–2.1E-04	–0.0356	–0.0190	–1.3E-02
	Umpqua	68	7.3	0.612	–1.2E-03	–0.0427	0.0286	3.7E+00
	South	85	80.7	2.031	–1.1E-02	–0.3222	0.0139	4.8E+01

similar pattern is evident in the coefficients for precipitation. What appears to be a clear scale effect in the valley index is not repeated in the climate predictors.

Similarly, if we examined these coefficients at only one scale of aggregation, we could conclude that there was a regional effect. Although, whether we found a regional effect that increases or decreases with latitude would depend on which scale of aggregation we used to examine the data.

Roads

Our best fit model for predicting roads explains some of the variance in roads, but the proportion varies from scale to scale (Table 3). At the watershed scale, we can explain a sizable portion of variation in roads, but at the other scales, these predictors explain very little of the variation. In some regions, roads are more closely tied to these predictors. For example, if we aggregate the data at the watershed scale and look only at the Umpqua region, we can explain much of the variation in roads.

Nearly the same predictors that proved valuable for explaining rural development were retained in the model for roads. However, we retained siltstone in the rural development model and sandstone in the roads model. Initially, this seems appropriate because the presence of rural development and roads might be related in the landscape. Although the location of some roads in the region may be determined by the presence of rural development, a great many are related to the forestry industry and thus occur in the absence of rural development. Additionally, when we examine the coefficients, the predictors function differently in this model than in the model for rural development.

The presence of roads is generally positively correlated with the valley index, precipitation, temperature, sandstone and mafic volcanic flows; in contrast, rural development was negatively correlated to precipitation. The

Table 3. Coefficient estimates and goodness-of-fit for the single best fit model for roads (m). Results are from a linear regression of the log of roads on the listed variates. Estimates are presented by scale and region. The coefficient estimate for mafic volcanic is unavailable for the Umpqua region at the buffer-100m scale due to lack of data.

							Lithology	
Scale	**Region**	**R^2**	**Intercept**	**Valley Index**	**Precipitation**	**Temperature**	**Sandstone**	**Mafic Volcanic**
buffer 100 m	all	0.07	6.1	4.5E-03	2.4E-04	0.0394	1.0E-03	–2.5E-04
buffer 500 m		0.06	8.0	–8.9E-04	1.5E-04	0.0233	1.3E-03	7.5E-04
watershed		0.33	1.0	4.5E-01	3.3E-04	0.2698	2.5E-03	1.9E-03
hu		0.05	10.7	1.1E-02	8.1E-05	0.0063	2.4E-03	4.1E-03
buffer 100 m	North	0.22	6.0	–1.3E-03	2.4E-04	0.0562	4.1E-03	2.9E-05
	Mid-coast	0.03	7.5	3.1E-03	2.3E-05	–0.0091	1.1E-03	–1.3E-03
	Umpqua	0.04	4.9	–4.9E-04	4.4E-04	0.0881	1.9E-03	NA
	South	0.20	3.4	1.1E-02	7.5E-04	0.1084	1.1E-03	2.0E-03
buffer 500 m	North	0.14	8.1	–1.6E-03	2.6E-04	0.0042	1.9E-03	8.8E-04
	Mid-coast	0.01	9.0	3.0E-03	–3.7E-05	–0.0136	1.7E-03	1.0E-03
	Umpqua	0.09	5.1	4.7E-03	7.3E-04	0.1185	6.4E-04	–3.8E-02
	South	0.21	6.3	6.4E-03	2.8E-04	0.0855	1.5E-03	–1.0E-03
watershed	North	0.43	–1.1	5.4E-01	6.9E-04	0.2896	7.3E-04	2.3E-03
	Mid-coast	0.45	1.5	6.1E-01	1.1E-04	0.2067	7.6E-03	1.7E-02
	Umpqua	0.50	–14.4	–1.4E-01	4.0E-03	0.8529	3.2E-04	4.8E-02
	South	0.26	–1.5	3.7E-01	1.5E-03	0.3278	–5.3E-04	–8.6E-03
hu	North	0.18	10.1	6.8E-03	2.3E-04	0.0144	–1.1E-06	4.3E-03
	Mid-coast	0.14	9.7	8.8E-02	2.2E-04	–0.0140	7.7E-03	8.6E-03
	Umpqua	0.18	12.3	–1.5E-02	2.1E-04	–0.0357	–2.9E-03	1.0E-02
	South	0.04	10.7	4.6E-03	–1.3E-04	0.0261	2.1E-03	6.6E-03

sign of the coefficients are not entirely consistent across scale and region. When examining the coefficients more closely, we are tempted to conclude a scale effect. In particular, there appears to be something unique about the watershed scale (Fig. 4).

Simulation experiment

Wood volume is one of several features that are routinely measured in surveys of stream habitat (Moore et al. 1997). Large wood is valuable to rearing juvenile salmon in streams because it provides cover, shade and refuge from rapid currents (Beechie et al. 2000; Roni & Quinn 2001). We chose wood volume as an example of an instream target of ecological significance that may be modeled from landscape predictors. Using a dataset describing habitat features in Oregon streams, we developed a simulation framework (Fig. 5) to demonstrate how traditional landscape analyses can be susceptible to multicollinearity and may produce misleading results.

Steps 1–2: We use regression to measure the correlation structures that exist within our data. We must measure both the explanatory relationship between our variate of interest (wood volume) and its predictors, as well as the relationship among the predictors.

We find that with simple linear regression, we can explain a significant portion of the variation ($R^2 = 0.37$) in wood volume (logLWDvol) using only three explanatory variates: gradient, elevation and the presence of rural development

$$\text{logLWDvol} = 1.92 + 0.50{*}\text{logGradient} + 0.0016{*}\text{Elevation} - 0.63{*}\text{RuralDevelopment} + \text{error} \quad (1)$$

These coefficients comprise the vector **β** (Fig. 5). Traditionally, an analysis would conclude by interpreting this result. However, a correlative relationship exists among these predictors for wood volume. The Pearson's correlation coefficients between log-gradient and elevation is 0.27, between log-gradient and rural development is 0.30, and between elevation and rural development is 0.16.

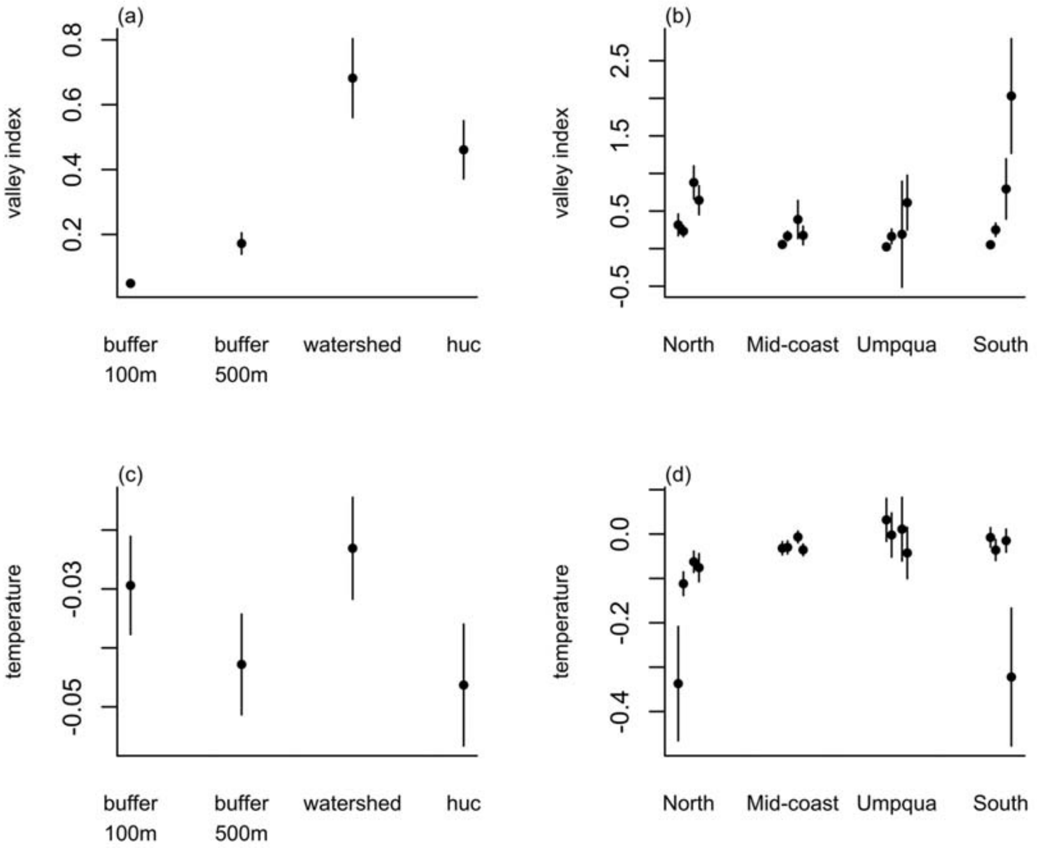

Fig. 3. Selected coefficient estimates with standard errors from the best fit model for the presence of rural development. In the left column, we show the estimates as a function of scale for all regions. In the right column, we show the scale by region estimates. The scales are given in the same order in both columns, but not labeled in the right column.

These correlations are small enough that most analysts would ignore them.

Our original data describes the percent area in rural development, however this variate is very zero inflated and so we convert it to the binary variate, presence of rural development. Using a logistic regression, we can predict the presence of rural development using only gradient and elevation:

$$\Pr\{\text{Rural Development} = 1\} = \text{logit}^{-1}(2.07 - 2.13{*}\text{logGradient} - 0.0062{*}\text{Elevation}) \quad (2)$$

Note that Equation (2) is estimated from the subset of reaches that were sampled for habitat metrics, and thus the result differs from those presented in Table 2, which are estimated from the complete set of reaches. This model correctly predicts the presence of rural development 73 % of the time. These logistic regression coefficients comprise the vector **Θ** (Fig. 5).

Before we proceed, we center and scale our data by subtracting the means and dividing by one standard deviation for each variate. This facilitates our ability to systematically simulate coefficient values across predictors. We also re-estimate **Θ** and **β** given the centered and scaled variates.

Step 3. We clone our independent variates, gradient and elevation, using the method of moments based on a gamma distribution (Hogg & Craig 1959). In the method of moments, we fit a probability density function to a histogram of data. We then take random samples from the fitted gamma distribution to generate cloned data with the same distributional characteristics as the original data. This also eliminates any correlation structure that exists empirically between gradient and elevation. In the simulation, all of the correlation structure will be contained in the intermediate variate, rural development.

Step 4. **Θ** describes the observed correlation structure among our predictors, and so we use these values along with a tuning parameter, α, to simulate new values for rural development. We select new coefficients ($\tilde{\Theta}$) stochastically by taking a random draw from a multinomial distribution with weights **Θ** and constrain them to sum to

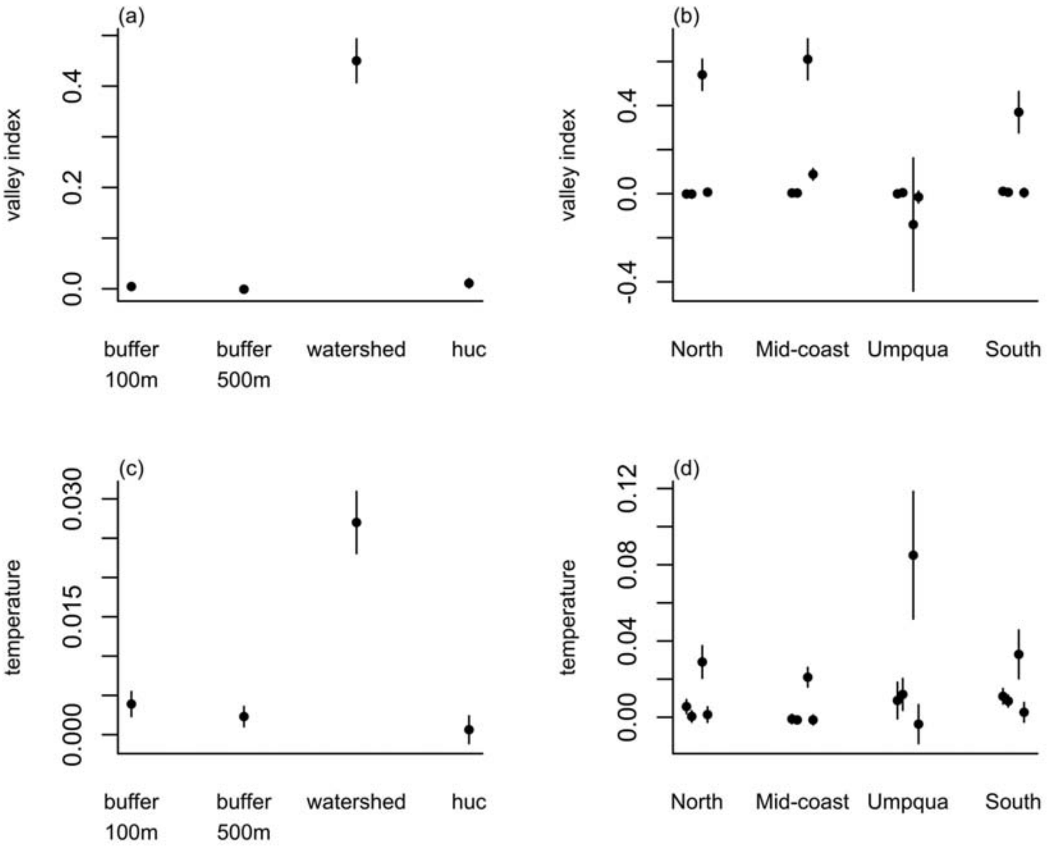

Fig. 4. Selected coefficient estimates with standard errors from the best fit model for the log of roads. In the left column, we show the estimates as a function of scale for all regions. In the right column, we show the scale by region estimates. The scales are given in the same order in both columns, but not labeled in the right column.

α. Throughout this section we use the tilde symbol (~) to distinguish simulated from empirical values. We add a noise term by taking a random draw from a standard normal distribution and multiplying this by $(1-\alpha)$.

Simulating new values for the presence of rural development is a two-step process. First, we simulate the probability of observing rural development (p_r) given the predictors:

$$p_r = \text{logit}^{-1}\,[\Sigma\,\tilde{\theta}_0 + \tilde{\theta}_1\tilde{x}_1 + \tilde{\theta}_2\tilde{x}_2 + (1-\alpha)\varepsilon]$$

Next, we use the resulting probability to simulate an outcome for rural development by taking a random draw from a Bernoulli distribution:

$$\tilde{x}^{\alpha} \sim Bern(p_r)$$

Step 5. To simulate new values for our response variate, wood volume, we again select new coefficients at random by drawing from a multinomial distribution with weights $\boldsymbol{\beta}$. We use these values in conjunction with our simulated predictors to predict simulated values for our response variate ($\tilde{y}$).

$$\tilde{y} = \Sigma\tilde{\beta}_0 + \tilde{\beta}_1\tilde{x}_1 + \tilde{\beta}_2\tilde{x}_2 + \tilde{\beta}_3\tilde{x}_3^{\alpha} + \tilde{\beta}_4\varepsilon$$

where

$$\varepsilon \sim \text{Norm}(0,1) \qquad \text{and} \qquad \sum_{i=0}^{4}\tilde{\beta}_1 = 1$$

and x_1 is the simulated value for log-gradient, x_2 is the simulated value for elevation, and $\tilde{x}_3^{\alpha}$. is the simulated value for presence of rural development.

Step 6. We have now fully simulated new values for all of our variates with the desired correlation structure. We next estimate new coefficient values using simple linear regression.

$$\tilde{y} \sim x_1 + x_2 + x_3^{\alpha} \Rightarrow \tilde{\boldsymbol{\beta}} \qquad (3)$$

Step 7. Finally, we compare the model fit between the empirical relationship and the relationships that emerge

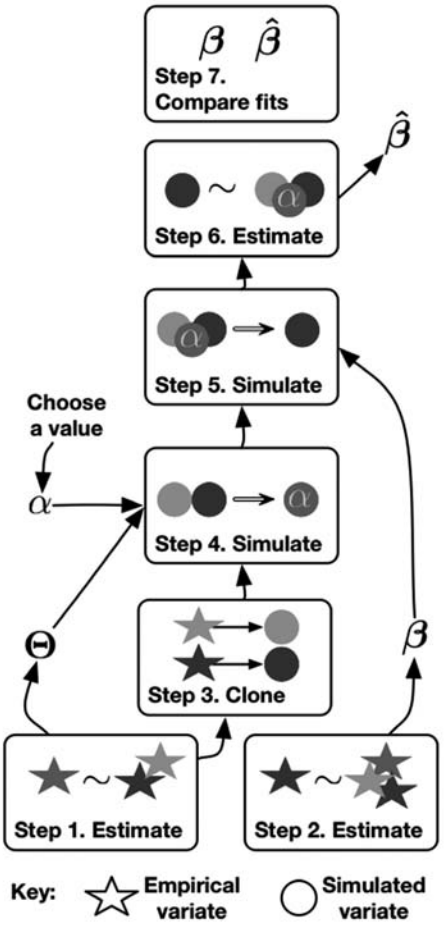

Fig. 5. Diagram of the simulation experiment. Stars represent our data for four variates: wood volume, presence of rural development, gradient and elevation. Circles represent simulated clones of these variates. The empirical correlation structure in our data is used to create a simulated dataset with similar correlation structure but with the addition of a tuning parameter (α). This allows us to control the amount of correlation structure among the predictors and observe how this affects our ability to estimate our response, wood volume.

when we manipulate the correlation structure by tuning α. Any change in the fit of $\hat{\boldsymbol{\beta}}$ with α is inappropriate because α is unrelated to the relationship between wood volume and its predictors; it is only the tuning parameter, α, that influences the relationship among the predictors.

To illustrate the effect of α on overall model fit, we also vary the amount of noise in the model predicting wood volume. We define a parameter, β_ε, to be the proportion of variation in wood volume that is not explained by the predictors. The remaining variation in wood volume $(1–\beta_\varepsilon)$ is distributed among the predictors in proportion to their weight within $\boldsymbol{\beta}$.

Simulation results

As expected, we find that problems are introduced by increasing the correlation structure among the predictors. First, while the coefficient estimates are unbiased, the standard errors of those estimates increase as correlation structure increases (Fig. 6). This increase in estimated parameter variance has been observed before and is a familiar consequence of multicollinearity (Belsley et al. 2004; O’Brien 2007).

We also see that the standard errors increase most in the estimates for rural development and least in the estimates for log-gradient. This reflects the relative magnitudes of the standardized values of $\boldsymbol{\beta}$ that were used to generate the correlation structure. It seems that the more a variate is involved in the correlation structure, the more its coefficient estimates are inflated due to multicollinearity.

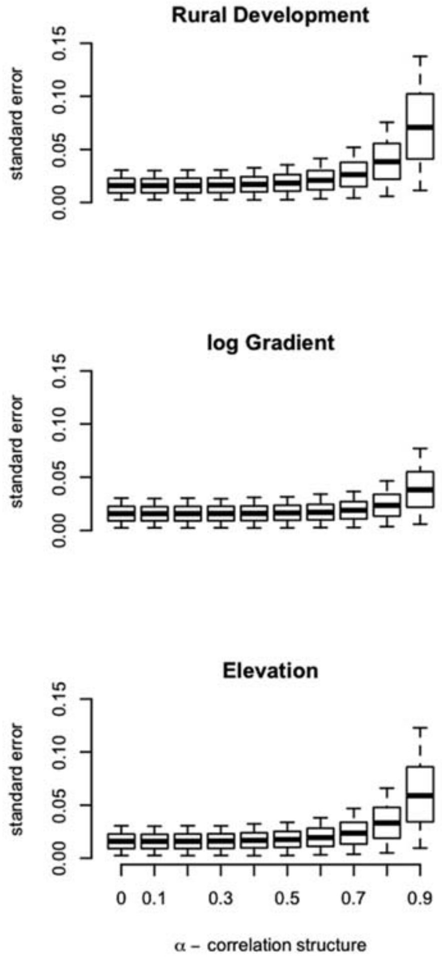

Fig. 6. Standard errors of coefficients from wood volume model as a function of correlation structure (α). The presence of rural development is correlated to gradient and elevation via α.

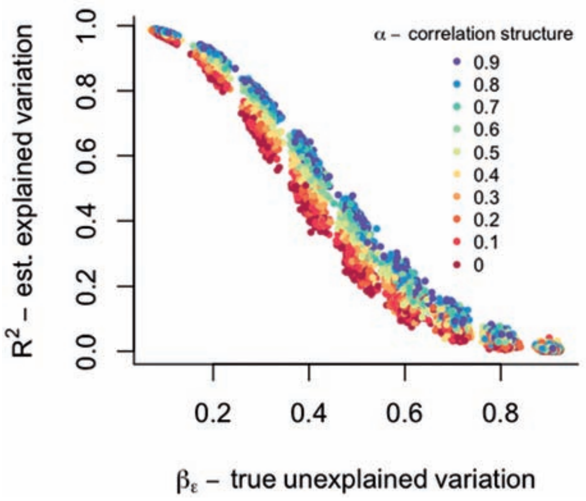

Fig. 7. Goodness-of-fit (R^2) of the wood volume model given true unexplained variation (β_ε) as a function of the correlation structure among the predictors (α). When $\alpha = 0$, we observe the correct inverse relationship between R^2 and β_ε. As correlation structure increases, we begin to overestimate R^2.

The second problem we encounter is that our overall goodness-of-fit estimates become inflated as correlation structure increases. In Figure 7, we show the relationship between β_ε, which is the true noise in the relationship between wood volume and its predictors, and the estimated explained variation, R^2. When there is no correlation structure among the predictors ($\alpha = 0$), we see the correct inverse relationship between these quantities. However, as α increases, we see an incorrect increase in R^2; i.e., our perceived goodness-of-fit is inflated due to the presence of multicollinearity. Essentially, we are over-fitting. This pattern falls apart when β_ε is very high; this is because the correlation among the predictors matters less when the predictors overall explain only very little variation in the wood volume.

In practice, it is the intermediate values for β_ε and α that are most representative of a typical model building process in ecology. For intermediate values of β_ε between 0.35 and 0.45 and an $\alpha = 0.5$, the largest inflation of R^2 is from 0.35 to 0.50. This increase in R^2 is likely large enough to influence model building decisions, i.e. we would be likely to retain any variate whose inclusion caused a similar change to R^2.

Discussion

We have demonstrated that, in the Oregon Coastal Province, immutable landscape features can predict human use at the landscape scale surprisingly well. Furthermore, these correlations between natural and human features vary by both scale and region. Although we model data from coastal Oregon, we suggest that such correlation structures exist for most areas of the world, though the character likely varies with the history of human development and the environment. Human impact on the landscape is widespread (DeFries et al. 2008), and therefore we expect that correlations between human and natural features are common. Human development is rarely, if ever, purely random within a heterogenous landscape. These spatial correlation structures may shift abruptly at political borders or ecological features. For example, we have shown that in Oregon, rural development tends to occur in valleys. However, nothing is inherent to rural development that requires it to occur in valleys; in other regions of the world, rural development occurs preferentially at high elevation – e.g., New Mexico, Bolivia, Himalaya (Brush 1982; Guillet et al. 1983; Julyan 2006). The structuring factor is the history of human settlement, something that is likely to vary with landscape features and regions. These correlation structures would also be expected to shift over time as technological or sociological changes enable new types of areas to be preferentially developed.

We also demonstrated, via simulation study, that these correlations cause the well-known problems of inflated coefficient standard errors and overly optimistic estimates of model fit associated with more traditional types of multicollinearity. We confirmed that the magnitude of statistical problems associated with this form of multicollinearity is, at least in part, driven by the strength of the true relationships between the individual predictors and the response. Though they present statistical challenges, these correlation structures are part of the functioning of the ecosystem. They are important to study for their own sake as well as for their ability to undermine statistical estimation.

Particular complexities of the spatial ecology of rivers

While spatial data may co-vary in many ways, spatial patterns sub-sampled by rivers introduce bias in spatial patterns that is rarely considered. Rivers themselves are not distributed randomly across the landscape but occur at low points in the landscape, flow from higher to lower elevations, and flow through particular lithologies. This relationship between rivers and the landscapes through which they flow has a number of implications for studying riverine spatial ecology. In particular, it may magnify problems associated with spatial covariance of landscape data. Samples of landscape data that are defined by and centered on rivers represent a biased sampling of the landscape.

As rivers are studied at increasingly broad scales, we do not take a larger, random sample of the landscape through which the river flows. Because of the way rivers cut through landscapes, larger scales of landscape aggregation include systematically different areas. For example, larger scales generally include more area at higher elevations and a smaller proportion of riparian forest. In our study, the smaller scales, the buffer scales, generally included a greater proportion of area in river valleys than did the broader scales. In expanding to the larger watershed and HU scales, we added steeper hillslopes, higher elevations, and older forests but little additional valley. Even simply increasing the length or location of sampled river reaches, introduces systematic changes in the landscape through which the river flows. As the reach length increases in, for example a downstream direction, the reach generally gets wider, carries a larger volume of water, and is of lower gradient (Vannote et al. 1980). Moving downstream may also take us from higher elevations and alpine forests to lower elevations and coastal forests, as well as across land ownership gradients. And, there are climatic gradients. In coastal Oregon, downstream areas generally have lower rainfall and increasing fog. The systematic effects of increasing scale can motivate a mechanistic explanation for some of the scale effects we observed. However, we also observed that scale effects differ across regions, even within our relatively homogenous study area; coefficient estimates were non-stationary with respect to both scale and location and there appeared to be an interaction between scale and region.

Spatial correlation has been widely considered in the spatial ecology literature, although with a greater emphasis on terrestrial systems (Legendre 1993). However, these methods from spatial statistics are concerned with characterizing the auto-correlation within a variate across space, which is a slightly different sense of correlation than our focus on the co-occurrence and overlap of multiple processes.

Some instream features of river structure have begun to be addressed in the statistical literature. For example, the unidirectional flow and network structure of river systems creates unique patterns of spatial correlation in water chemistry (ver Hoef et al. 2006; Peterson et al. 2007). New analytical methods have been developed to characterize these spatial correlations (Garreta et al. 2009; Peterson & ver Hoef 2010). Also, a new literature is emerging to understand the special properties of population dynamics in dendritic networks, such as riverine systems (Campbell Grant et al. 2007; Brown & Swan 2010).

Implications for landscape-scale riverine research and monitoring

1. The interpretation of spatial summaries is dynamic

We showed that full information signified by a particular variate is particular to the scale and region of the data. The ultimate consequence of this result is that the very significance of our data is dependent on underlying correlation structures and, therefore, is dynamic. This is a very difficult thing to accept in practice. Intuitively, we expect tangible variables, like ‘precipitation,’ to have fixed interpretations. Instead, a metric of precipitation may indirectly provide information about elevation or rural residential housing development. To further complicate matters, this indirect information may be present only in one region or at one spatial scale. By working to clearly describe the shifting interpretation of data in different contexts, we make it easier to understand how even simple variates must be interpreted with great care and consideration of context.

2. Expect multicollinearity

Ecologists are often guilty of not doing the most basic step, looking for the multicollinearity that is present in their data (Graham 2003). Our most vehement advice is that there should be a shift in expectations. When using spatial datasets and, in particular, when linking spatial data to riverine systems, we should expect underlying correlation structures to lead to data with dynamic interpretation.

3. Increase the use of exploratory data analysis

A combination of maps and graphs can provide insights about how variables are related across space and scale. These visualization tools do not demand high-level statistics, but rather, careful thinking about the dynamic nature of the datasets and about how variables are related to one another. Interactions among the spatial data should be explored beyond simple spatial autocorrelation of one variable or stationary multicollinearity that results from redundant metrics. We also advocate the liberal use of exploratory graphical, cartographic and statistical analyses. Spatial datasets have many subtleties that can be uncovered by thorough exploratory analysis before they obscure the results of more complex statistical analyses.

4. Increase interdisciplinary research

Because data describing, for example, precipitation, may inadvertently contain information about rural residential development, it is wise to include social scientists in the search for better models and better modeling techniques. While much of riverine landscape ecology has relied on the implicit assumption of a random distribution of humans across landscapes (Steel et al. 2010), other disciplines have developed methods to model and predict human settlement. Geography, human ecology and history each can contribute to our understanding of how humans have chosen to use landscapes and how, in turn, they have shaped aquatic ecosystems. Unraveling the complexities of how human land use affects instream processes would be facilitated by combining insights from geography, statistics, landscape ecology, human ecology and other scientific disciplines.

5. Resist mechanistic interpretations

It is natural to interpret coefficients mechanistically; this is particularly problematic in riverine landscape research. When data and analyses become complex, the need to simplify is at odds with the cognitive discipline required to distinguish cause from correlation. Observed patterns should be considered just that, patterns. We can compare patterns across scale and region (including underlying correlation structures) and we can generate theories from observed patterns. But, to demonstrate mechanism, we need more than correlative results.

6. Use advanced statistical tools

Latent variable analysis is a growing and important avenue for analyzing unobserved variables (Graham 2003; Shipley 2002; Tomer & Pugesek 2003). While terms vary widely across methods, a large class of statistical methods relies on a latent variable framework.

A latent variable is an entity that is not directly observed, such as our influential exogenous factors. Latent variables may be usefully divided into two types: hidden or hypothetical. Hidden variables are features that can be measured in principle, but for some reason are not. For example, we often rely on model predictions for stream surface temperature. While it is certainly possible to measure temperature, it is impractical to do so for all sites in a watershed and a latent variable model provides a reliable approximation (Daly et al. 1997). A hypothetical variable represents an entity which cannot be directly measured, usually an abstraction. A good example would be habitat quality. This is a complex and multi-faceted concept that may be modeled from a large number of directly measureable entities (Rose 2000; Johnson 2007).

Latent variable models are any method that relies on modeling these unobserved factors. We include in this category structural equation modeling, hierarchical or multilevel linear models, factor analysis, covariance structure analysis, and path analysis, as well as other methods (Graham 2003; Lee 2007). Many of these methods have a long history of wide use in the social sciences (Mueller 1997; Bullock et al. 1994), where several software tools have been developed to facilitate the use of latent variable models.

Latent variable methods have also been thoroughly pioneered in ecology. For example, covariance structure analysis was used by Infante et al. (2006) to estimate the dependence of fish assemblages on channel shape given several intermediary reach scale variates and again by Zorn et al. (2006) to estimate fish biomass from habitat and landscape variates collected at different scales. Pugesek et al. (2003) present a thorough introduction to structural equation modeling. And, Qian et al. (2010) illustrate the utility of multilevel linear models by predicting nitrous oxide emissions from fertilized soil despite confounding factors at multiple scales.

7. Do not extrapolate beyond the correlation structure

It is standard advice to avoid out-of-sample prediction when faced with irreducible multicollinearity (Mendenhall & Sincich 1996). This is a very serious problem because, as applied ecologists, we require the ability to generalize from regions where we have data to regions where we have none. Unfortunately, the dynamic significance of coefficient estimates bounds our interpretation of relationships and models to the correlation structure of the original data. When we build a model, we must respect that it is inherently limited to the scale and region of the data used to build the model; we cannot routinely generalize the results of our analyses to new regions or new systems. We can make predictions within the correlation structure (e.g., Steel et al. 2004), or we can compare observations across correlation structures. But, we must be clear that when we compare patterns across scales or regions, we are comparing across correlation structures as well.

8. Measure and monitor those areas where correlation structures fall apart

Often, the ultimate goal is to move beyond data with "dynamic interpretation" and to understand the mechanistic roles of our variates. Understanding correlative structures helps us do this, because it allows us to then look for the places where these structures fall apart. For example, in Oregon there may be very few valleys without rural development, but those few valleys may be particularly informative in untangling the mechanistic roles of valleys versus rural development. Monitoring programs can be designed specifically to include these types of high information sites.

In summary, our results suggest an increase in both caution and skepticism when building models that quantify relationships between what happens across the landscape and riverine responses. We are forced to acknowledge that variates can be highly correlated without representing redundant ecological information or reflecting a mechanistic link among the variates. We also see encouragement for interdisciplinary teams to further investigate how best to find useable information from these relatively new, and now ubiquitous, spatial datasets (Steel et al. 2010). And, this work opens the door to considering how spatial correlation structures themselves might be informative in designing statistical methods, in setting up monitoring programs, and in understanding how landscape conditions drive what happens in streams and rivers.

Acknowledgements

We would like to thank Eli Holmes, and other internal reviewers for providing useful comments and feedback. Thanks are due to Blake Feist for help with metadata and to Kara Anlauf for providing the well-curated habitat data. We are grateful to the Oregon Watershed Enhancement Board (OWEB), the Integrated Status Effectiveness and Monitoring Program (ISEMP) and the National Research Council Research Associate Program for providing funding to support this research.

References

Allan, J. (2004a): Influence of land use and landscape setting on the ecological status of rivers. – Limnetica **23**: 187–198.

Allan, J. (2004b): Landscapes and riverscapes: the influence of land use on stream ecosystems. – Annual Review of Ecology and Systematics **35**: 257–284**.**

Barker, L.S., Felton, G.K. & Russek-Cohen, E. (2006): Use of Maryland biological stream survey data to determine effects of agricultural riparian buffers on measures of biological stream health. – Environmental Monitoring and Assessment **117**: 1–19.

Baxter, C.V. & Hauer, F.R. (2000): Geomorphology, hyporheic exchange, and selection of spawning habitat by bull trout (*Salvelinus confluentus*). – Canadian Journal of Fisheries and Aquatic Sciences **57**: 1470–1481.

Beechie, T., Pess, G., Kennard, P., Bilby, R. & Bolton, S. (2000): Modeling recovery rates and pathways for woody debris recruitment in northwestern Washington streams. – North American Journal of Fisheries Management **20**: 436–452.

Belsley, D., Kuh, E. & Welsch, R. (2004): Regression diagnostics: Identifying influential data and sources of collinearity. – Wiley Online Library ISBN: 0471058564.

Brown, B.L. & Swan, C.M. (2010): Dendritic network structure constrains metacommunity properties in riverine ecosystems. – Journal of Animal Ecology **79**: 571–580.

Brush, S. (1982): The natural and human environment of the central Andes. – Mountain Research and Development **2**: 19–38.

Bullock, H., Harlow, L. & Mulaik, S. (1994): Causation issues in structural equation modeling research. – Structural Equation Modeling: A Multidisciplinary Journal **1**: 253–267.

Burgi, M., Hersperger, A., Hall, M., Southgate, E. & Schneeberger, N. (2007): Using the past to understand the present land use and land cover. – A Changing World **8**:133–144.

Burnett, K., Reeves, G., Miller, D., Clarke, S., Vance-Borland, K. & Christiansen, K. (2007): Distribution of salmon-habitat potential relative to landscape characteristics and implications for conservation. – Ecological Applications **17**: 66–80.

Campbell Grant, E., Lowe, W. & Fagan, W. (2007): Living in the branches: population dynamics and ecological processes in dendritic networks. – Ecology Letters **10**: 165–175.

Creque, S., Rutherford, E. & Zorn, T. (2005): Use of GIS-derived landscape-scale habitat features to explain spatial patterns of fish density in Michigan rivers. – North American Journal of Fisheries Management **25**: 1411–1425.

Daly, C., Taylor, G. & Gibson, W. (1997): The PRISM approach to mapping precipitation and temperature. – In Preprints, 10th Conf. on Applied Climatology, Reno, NV, Amer. Meteor. Soc, 10–12.

DeFries, R., Foley, J. & Asner, G. (2008): Land-use choices: balancing human needs and ecosystem function. – Frontiers in Ecology and the Environment **2**: 249–257.

Filipe, A., Marques, T., Seabra, S., Tiago, P., Ribeiro, F., Da Costa, L., Cowx, I. & Collares-Pereira, M. (2004): Selection of priority areas for fish conservation in Guadiana River Basin, Iberian Peninsula. – Conservation Biology **18**: 189–200.

Firman, J.C. & S.E. Jacobs. (2001): A survey design for integrated monitoring of salmonids. Proceedings of the first international symposium on geographic information systems (GIS) in fishery science. Pages 242–252 in T. Nishida, P.J. Kailola, and C.E. Hollingworth, editors. Fishery GIS Research Group, Saitama, Japan.

Garreta, V., Monestiez, P. & Ver Hoef, J. (2009): Spatial modelling and prediction on river networks: up model, down model or hybrid? – Environmetrics **21**: 439–456.

Grace, J. & Bollen, K. (2005): Interpreting the results from multiple regression and structural equation models. – Bulletin of the Ecological Society of America **86**: 283–295.

Graham, M. (2003): Confronting multicollinearity in ecological multiple regression. – Ecology **84**: 2809–2815.

Gudea, P.H., Hansen, A.J., Rasker, R., & Maxwell, B. (2006): Rates and drivers of rural residential development in the Greater Yellowstone. – Landscape and Urban Planning **77**: 131–151.

Guillet, D., Godoy, R., Guksch, C., Kawakita, J., Love, T., Matter, M. & Orlove, B. (1983): Toward a Cultural Ecology of Mountains: The Central Andes and the Himalayas Compared. – Current Anthropology **24**: 561–574.

Hogg, R.V. & Craig, A.T. (1959): Introduction to mathematical statistics. – Macmillan, New York, New York

Independent Multidisciplinary Science Team (IMST). (2002): Recovery of wild salmonids in western Oregon lowlands. – Technical Report 2002-1 to the Oregon Plan for Salmon and Watersheds.Governor's Natural Resources Office, Salem, Oregon, USA.

Infante, D., Wiley, M., Seelbach, P., Hughes, R., Wang, L. & Seelbach, P. (2006): Relationships among channel shape, catchment characteristics, and fish in Lower Michigan streams. – In American Fisheries Society Symposium, vol. 48. American Fisheries Society, 5410 Grosvenor Ln. Ste. 110 Bethesda MD 20814-2199 USA,

Jacobs, S.E. & Cooney, C.X. (1997): Oregon coastal spawning surveys, 1994 and 1995. – Oregon Department of Fish and Wildlife, Information Reports (Fish) 97-5. Portland, OR, USA.

Johnson, L. & Gage, S. (1997): Landscape approaches to the analysis of aquatic ecosystems. – Freshwater Biology **37**: 113–132.

Johnson, L. & Host, G. (2010): Recent developments in landscape approaches for the study of aquatic ecosystems. – Journal of the North American Benthological Society **29**: 41–66.

Johnson, M. (2007): Measuring habitat quality: a review. – The Condor **109**: 489–504.

Julyan, R. (2006): The mountains of New Mexico. – In Place Names of New Mexico. University of New Mexico Press, Santa Fe, New Mexico.

Kirch, P.V., Hartshorn A.S., Chadwick, O.A., Vitousek, P.M., Sherrod, D.R., Coil, J., Holm, L., & Sharp, W.D. (2004): Environment, agriculture, and settlement patterns in a marginal Polynesian landscape. – Proceedings of the National Academy of Sciences **101**: 9936–9941.

Kline, J., Azuma, D. & Moses, A. (2003): Modeling the spatially dynamic distribution of humans in the Oregon (USA) Coast Range. – Landscape Ecology **18**: 347–361.

Lammert, M. & Allan, J. (1999): Assessing biotic integrity of streams: effects of scale in measuring the influence of land use/cover and habitat structure on fish and macroinvertebrates. – Environmental Management **23**: 257–270.

Lee, S. (2007): Handbook of latent variable and related models. Volume 1 of Handbook Series on Computing and Statistics with Applications. – North-Holland/Elsevier, Amsterdam, The Netherlands.

Legendre, P. (1993): Spatial autocorrelation: trouble or new paradigm? – Ecology **74**: 1659–1673.

Legendre, P. & Legendre, L. (1998): Numerical ecology. – No. 20 in Developments in Environmental Modelling. Elsevier Science, 2nd edition. Amsterdam, The Netherlands.

Lyle, J. (1999): Design for human ecosystems: Landscape, land use, and natural resources. – Island Press, Washington, D.C.

Margules, C.R. & Pressey, R.L. (2000): Systematic conservation planning. – Nature **405**: 243–253.

McHugh, P. & Budy, P. (2004): Patterns of spawning habitat selection and suitability for two populations of spring Chinook salmon, with an evaluation of generic versus site-specific suitability criteria. – Transactions of the American Fisheries Society **133**:89–97.

Mendenhall, W. & Sincich, T. (1996): A second course in statistics: regression analysis, sixth edition. – Prentice Hall, Upper Saddle River, New Jersey.

Moore, J., Schindler, D.E., Carter, J., Fox, J., Griffiths, J. & Holtgrieve, G. (2007): Biotic control of stream fluxes: Spawning salmon drive nutrient and matter export. – Ecology **88**: 1278–1291.

Moore, K., Jones, K. & Dambacher, J. (1997): Methods for stream habitat surveys. – Information Report 2007-01, Oregon Department of Fish and Wildlife.

Mueller, R. (1997): Structural equation modeling: back to basics. – Structural Equation Modeling: A Multidisciplinary Journal **4**: 353–369.

O'Brien, R. (2007): A caution regarding rules of thumb for variance inflation factors. – Quality and Quantity **41**: 673–690.

Ohmann, J. & Gregory, M. (2002): Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. – Canadian Journal of Forest Research **32**: 725–741.

Peterson, E.E., Theobald, D. & Ver Hoef, J.M. (2007): Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. – Freshwater Biology **52**: 267–279.

Peterson, E.E. & Ver Hoef, J.M. (2010): A mixed-model moving-average approach to geostatistical modeling in stream networks. – Ecology **91**: 644–651.

Petraitis, P., Dunham, A. & Niewiarowski, P. (1996): Inferring multiple causality: the limitations of path analysis. – Functional Ecology **4**: 421–431.

Postel, S. & Carpenter, S. (1997): Freshwater ecosystem services. – In G.C. Daily (ed.), Nature's services. Island Press, Washington D.C. pp. 392.

Pugesek, B., Tomer, A. & von Eye, A. (2003): Structural equation modeling: applications in ecological and evolutionary biology. – Cambridge University Press, Cambridge, U.K.

Qian, S.S., Cuffney, T.F., Alameddine, I., McMahon, G. & Reckhow, K.H. (2010): On the application of multilevel modeling in environmental and ecological studies. – Ecology **91**: 355–361.

R Development Core Team (2010): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rickenbach, Z., Clarke, S. & Burnett, K. (2000): Sixth/Seventh Field Hydrologic Units for the CLAMS Area Geospatial Data Presentation Form: vector digital data. – URL http://www.fsl.orst.edu/clams/cfsl0233.html.

Roni, P. & Quinn, T.P. (2001): Effects of wood placement on movements of trout and juvenile coho salmon in natural and artificial stream channels. – Transactions of the American Fisheries Society **130**: 675–685.

Rose, K. (2000): Why are Quantitative Relationships between Environmental Quality and Fish Populations so Elusive? – Ecological Applications **10**: 367–385.

Seaber, P.R., Kapinos, F.P. & Knapp, G.L. (1987): Hydrologic Unit Maps: U.S. Geological Survey Water-Supply Paper 2294, p. 63.

Shipley, B. (2002): Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. – Cambridge University Press. Cambridge, U.K.

Spies, T., McComb, B., Kennedy, R., McGrath, M., Olsen, K. & Pabst, R. (2007): Potential effects of forest policies on terrestrial biodiversity in a multi-ownership province. – Ecological Applications **17**: 48–65.

Steel, E., Feist, B., Jensen, D., Pess, G., Sheer, M., Brauner, J. & Bilby, R. (2004): Landscape models to understand steelhead (*Oncorhynchus mykiss*) distribution and help prioritize barrier removals in the Willamette basin, Oregon, USA. – Canadian Journal of Fisheries and Aquatic Sciences **61**: 999–1011.

Steel, E.A., Hughes, R.M., Fullerton, A.H., Schmutz, S., Young, J.A., Fukushima, M., Muhar, S., Poppe, M., Feist, B.E. & Trautwein, C. (2010): Are We Meeting the Challenges of Landscape-Scale Riverine Research? A Review. – Living Reviews in Landscape Research **4**.

Stevens, Jr. D. & Olsen, A. (2004): Spatially balanced sampling of natural resources. – Journal of the American Statistical Association **99**: 262–278.

The Water Framework Directive. Directive 2000/60/EC of the European Parliament and the Council of 23 October 2000 establishing a framework for Community action in the field of water policy – OJL 327, 22 December 2000, pp. 1–73.

Thieme, M., Lehner, B., Abell, R., Hamilton, S.K., Kellndorfer, J., Powell, G. & Riveros, J.C. (2007): Freshwater conservation planning in data-poor areas: An example from a remote Amazonian basin (Madre De Dios River, Peru and Bolivia). – Biological Conservation **135**: 484–501.

Thomas, J. et al. (1993): Forest ecosystem management: an ecological, economic, and social assessment. Report of the forest ecosystem management assessment team (FEMAT), US Forest Service, Bureau Land Management, National Marine Fisheries Service, US Fish and Wildlife Service, National Park Service, Environmental Protection Agency, Portland, Oregon.

Tomer, A. & Pugesek, B. (2003): Guidelines for the implementation and publication of structural equation models. – In: Structural Equation Modeling: Applications in Ecological and Evolutionary Biology. Cambridge University Press, Cambridge, United Kingdom.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R. & Cushing, C.E. (1980): The River Continuum Concept. – Canadian Journal of Fisheries and Aquatic Sciences **37**: 130–137.

Van Sickle, J., Baker, J., Herlihy, A., Bayley, P., Gregory, S., Haggerty, P., Ashkenas, L. & Li, J. (2008): Projecting the biological condition of streams under alternative scenarios of human land use. – Ecological Applications **14**: 368–380

Ver Hoef, J.M., Peterson, E. & Theobald, D.M. (2006): Spatial statistical models that use flow and stream distance. – Environmental and Ecological Statistics **13**: 449–464.

Waite, I.R., Brown, L.R., Kennen, J.G., May, J.T., Cuffney, T.F., Orlando J.L. & Jones, K.A. (2010): Comparison of watershed disturbance predictive models for stream benthic macroinvertebrates for three distinct ecoregions in western US. – Ecological Indicators **10**: 1125–1136.

Walker, G., MacLeod, N., Miller, R., Raines, G. & Connors, K. (2003): Spatial digital database for the geologic map of Oregon. – Open-File Report 03–67, US Geological Survey.

Whittingham, M., Stephens, P., Bradbury, R. & Freckleton, R. (2006): Why do we still use stepwise modelling in ecology and behaviour? – Journal of Animal Ecology **75**: 1182–1189.

Wimberly, M.C. & Spies, T.A. (2001): Influences of environment and disturbance on forest patterns in coastal Oregon watersheds. – Ecology **82**: 1443–1459.

Zorn, T., Wiley, M., Hughes, R., Wang, L. & Seelbach, P. (2006): Influence of landscape characteristics on local habitat and fish biomass in streams of Michigan's lower peninsula. – In: American Fisheries Society Symposium, vol. 48. American Fisheries Society, 5410 Grosvenor Ln. Ste. 110 Bethesda, Maryland, USA.

Received 1 October 2010
Modified version received 9 February 2011
Accepted 10 February 2011