

Quantile Regression Applied to Spectral Distance Decay

Duccio Rocchini and Brian S. Cade

Abstract—Remotely sensed imagery has long been recognized as a powerful support for characterizing and estimating biodiversity. Spectral distance among sites has proven to be a powerful approach for detecting species composition variability. Regression analysis of species similarity versus spectral distance allows us to quantitatively estimate the amount of turnover in species composition with respect to spectral and ecological variability. In classical regression analysis, the residual sum of squares is minimized for the mean of the dependent variable distribution. However, many ecological data sets are characterized by a high number of zeroes that add noise to the regression model. Quantile regressions can be used to evaluate trend in the upper quantiles rather than a mean trend across the whole distribution of the dependent variable. In this letter, we used ordinary least squares (OLS) and quantile regressions to estimate the decay of species similarity versus spectral distance. The achieved decay rates were statistically nonzero ($p < 0.01$), considering both OLS and quantile regressions. Nonetheless, the OLS regression estimate of the mean decay rate was only half the decay rate indicated by the upper quantiles. Moreover, the intercept value, representing the similarity reached when the spectral distance approaches zero, was very low compared with the intercepts of the upper quantiles, which detected high species similarity when habitats are more similar. In this letter, we demonstrated the power of using quantile regressions applied to spectral distance decay to reveal species diversity patterns otherwise lost or underestimated by OLS regression.

Index Terms—Biodiversity, distance decay, environmental gradients, quantile regressions.

I. INTRODUCTION: SPECTRAL DISTANCE DECAY AND QUANTILE REGRESSION

SPECIES diversity is often the most convenient proxy for other components of biodiversity, such as genetic diversity and landscape diversity [1]. In addition to species richness (α -diversity), species complementarity (β -diversity), i.e., the amount of turnover in species composition from one site to the other [2], undertakes a key role. Given two sites with the same number of species (α -diversity), they may be either very similar in their species composition, thus showing a low species complementarity (the β -diversity), or very different. Thus, β -diversity adds to the simpler concept of α -diversity the capability of detecting spatial gradients that functionally act in determining the spatial variation in species composition, which is

one of the principal aims of ecosystem monitoring and conservation biology [3]. Species composition similarity is expected to decay while spatial distance among sites increases, with its decay rate being a straightforward measure of β -diversity [4], [5]. However, at the local scale, spatial distance may not be high enough to reveal patterns of β -diversity. In this view, Tuomisto *et al.* [6] and Rocchini [7] introduced spectral distance decay as a powerful method for detecting species composition β -diversity, considering ecological distance as measured by an optical sensor rather than a spatial one. In fact, theoretically, differences in environmental properties of different habitats should lead to differences of spectral responses, which can be detected by satellite imagery.

In most cases, distance decay models are derived by ordinary least squares (OLS) regression between species similarity as the dependent variable versus a measure of distance as the explanatory variable, herewith including spectral distance. In classical regression analysis, the residual sum of squares is minimized within a regression model for the mean of the dependent variable distribution. However, many ecological data sets, particularly those related to communities, are characterized by a high number of zeroes [8] that can add noise to the regression model. In these cases, quantile regressions can be used to evaluate trends, considering various percentage points of the distributions, thus getting a more complete picture of the set [9]–[12], i.e., by estimating the regression model on quantiles τ rather than estimating the mean for the entire cloud of points. The mean regression model can be regarded as an average across all the quantile regression models. When there is heterogeneity in decay rates, the mean regression model will fail to convey information on the lower and higher decay rates associated with lower to higher quantiles. In particular, Rocchini [7] proved that maximum decay modeling allows the characterization of the maximum diversity of habitats, which represents key information for environmental management and conservation.

The aim of this letter is to demonstrate the power of using quantile regressions applied to spectral distance decay to reveal species diversity patterns otherwise lost or underestimated by OLS regression.

II. WORKED EXAMPLE

A. Study Sites and Field Data Acquisition

The study sites are represented by four Sites of Community Importance (SCIs) of the Natura 2000 Network sampled in summer 2005 (Habitat Directive 92/43/CEE) and located in the southwestern part of Siena, Italy. These four sites were sampled

Manuscript received May 28, 2008; revised June 13, 2008. Current version published October 22, 2008.

D. Rocchini is with the Department of Environmental Science, University of Siena, 53100 Siena, Italy, and also with TerraData *environmetrics*, University of Siena, 53100 Siena, Italy (e-mail: rocchini@unisi.it).

B. S. Cade is with the U.S. Geological Survey, Fort Collins Science Center, Fort Collins, CO 80526-8818, USA (e-mail: brian_cade@usgs.gov).

Digital Object Identifier 10.1109/LGRS.2008.2001767

during the first phase of a project on evaluation and monitoring of plant species diversity of the whole Natura 2000 Network of Siena (MoBiSIC project) during June–July 2005.

These sites contain a heterogeneous assemblage of plant communities, from the more thermophile communities dominated by *Quercus ilex* and *Q. cerris* (for Ripa d’Orcia and Pignelto) to the mesic ones dominated by *Fagus sylvatica* and *Castanea sativa* (Amiata). Croplands, seminatural grasslands, and shrublands were also present, particularly at Lucciolabella, a site that is characterized by the typical biancana badlands (i.e., peculiar erosion forms generated on Plio-Pleistocene marine clay outcrops).

To quantify and monitor plant species diversity in the network of SCIs, a restricted random sampling design was applied. This sampling design was based on the sampling points used for the National Inventory of Forests and Carbon Stocks (INFC). These points were identified by following a restricted random selection. The whole national territory was divided into a grid with contiguous cells of 1×1 km. Then, one random point was selected within each kilometeric cell. Those points falling within the study SCIs were chosen as sampling points, with the number of plots roughly proportional to the SCI surface. Once each point was located with a high precision GPS, a 10×10 m sample plot was delimited. With this simple approach, we achieved a sample with a nominal density of 1 point/km² that can easily be used for spatial inference. Within each 10×10 m plot, vascular plant composition was recorded.

Overall, 48 plots were sampled. One plot, falling within free water and containing no species, was removed; thus, 47 plots were used for further distance decay models.

B. Spectral Values

An ortho-Landsat Enhanced Thematic Mapper Plus (ETM+) image (path 192, row 030, acquisition date June 20, 2000, spatial resolution 28.5 m, band from 1–5 and 7) covering the whole study area was acquired from the Global Land Cover Facility site hosted by the University of Maryland (glcfapp.umiacc.umd.edu; see [13]). For this analysis, no radiometric and atmospheric correction have been applied. Although atmospheric effects modify actual reflectance values, the spectral differences in satellite images indicate differences in the reflectance characteristics of the ground and vegetation cover [6], ensuring that ecological variability will be detected [14].

To reduce data-dimensionality and spectral noise, due to minor components with little explanatory value, an unstandardized principal components analysis was performed by reducing the original six-band Landsat ETM+ data set to three principal components, explaining a cumulative variance of 98.94%.

C. OLS and Quantile Regressions Applied to Distance Decay Models

Species composition similarity between pairs of plots was calculated by using the Sørensen coefficient C_s on the strength of its widespread use [13], which is defined as

$$C_s = \frac{2j}{a + b} \quad (1)$$

where j is the number of species shared by plots A and B, a is the total number of species in plot A, b is the total number of species in plot B, and the coefficient C_s accounts for the overlap between two species lists and ranges from 0, indicating perfect dissimilarity, to 1, indicating perfect similarity.

A semimatrix of the pairwise compositional similarity between plots was then built. At the same time, a semimatrix of pairwise spectral distances based on Euclidean distance between plots was derived. Finally, compositional similarity was plotted against spectral distance to check for a possible relation (distance decay).

Linear models were fitted using both OLS and quantile regressions. Accordingly, the decay in species compositional similarity is described as

$$S = S_0 - cd \quad (2)$$

where S is the similarity at distance d , S_0 is the initial similarity or similarity at distance 0, and c is the decay rate.

Quantile regression extends the conventional linear model to estimating all parts of the response distribution S conditional on the predictors d , providing a more comprehensive characterization of the effects than those provided by estimates of the conditional mean as made with OLS regression [11]. While it is possible to estimate all quantiles on the zero (minimum) to one (maximum) interval for a given regression model, practical reporting requirements and substantive scientific interest may focus estimation efforts on a subset of all quantiles [9]–[12]. In this letter, upper quantile thresholds were considered according to the hypothesis being tested, i.e., that maximum decay modeling allows to characterize the maximum diversity of habitats [7]. Quantile-based fitting gives different weights to positive and negative residuals, leading to an asymmetric minimization. More formally, let $\{S_1, S_2, \dots, S_n\}$ denote the similarity values of a set of points lying within the scatterplot of species similarity versus distance. OLS regression minimizes residuals by solving

$$\text{residual} = \min \sum (S_i - \hat{S}_i)^2 \quad (3)$$

where \hat{S}_i is the estimated value for each S_i .

Giving different weights to positive and negative residuals and considering absolute rather than squared residuals, (3) turns out to be

$$\text{residual} = \min \sum |S_i - \hat{S}_i|T \quad (4)$$

where T is a multiplier term that is equal to τ (the quantile value) for positive deviations (i.e., $S_i - \hat{S}_i$) and to $1 - \tau$ for negative deviations. This asymmetric minimization fits a regression model through the upper part of the response distribution for $\tau > 0.5$ and through the lower part of the distribution for $\tau < 0.5$ [9]–[12]. Quantile regression with $\tau = 0.5$ is the median regression, which can be used as a central regression line similar to the mean regression estimated with OLS regression.

Notice that the quantile minimization of residuals shown in (4) is based on absolute values rather than on squared deviations like in OLS regression, thus reducing outlier effects. We refer to Koenker and Hallock [15] for a more detailed dissertation on

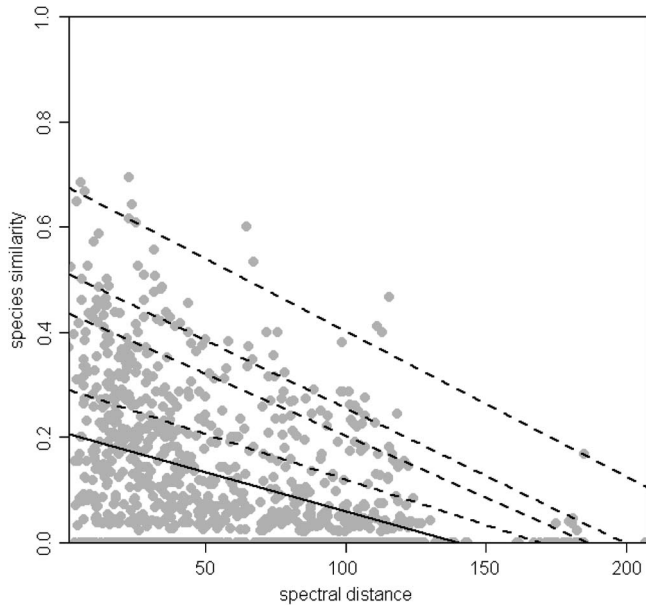


Fig. 1. Decay of species similarity versus spectral distance. (Solid line) OLS and (dashed lines) quantile regressions, considering that four different τ values (from upper to lower lines: 0.99, 0.95, 0.9, 0.75) were applied.

TABLE I
LINEAR MODELS CONSIDERING BOTH OLS AND QUANTILE REGRESSIONS
AT DIFFERENT QUANTILES τ . SAMPLE SIZE $n = 1081$

Regression type	τ	intercept (S_0)	intercept boundaries (CI 99%)	decay rate (c) $\times 10^{-5}$	decay rate boundaries (CI 99%) $\times 10^{-5}$
OLS	-	0.20 ***	0.19-0.23	149 ***	125-172
Quantile	0.75	0.29 ***	0.27-0.32	173 ***	156-224
	0.90	0.44 ***	0.41-0.48	237 ***	208-256
	0.95	0.51 ***	0.49-0.55	259 ***	220-287
	0.99	0.68 ***	0.59-0.78	276 ***	114-338

*** $p < 0.01$

the matter and to Gotelli and Ellison [16] for a brief summary of quantile regressions applied to ecological data.

A number of statistical software packages exist which perform quantile regression (see, e.g., Blossom at the U.S. Geological Survey Internet site). We used the `quantreg` package of R-software [17] on the strength of its widespread use and on simple replicability of the coded functions. Confidence intervals for quantile regression estimates were based on the rank-score test inversion approach with a localized bandwidth of quantiles, providing weights to account for distributional heterogeneity [18].

D. Results

The achieved decay rates were statistically nonzero ($p < 0.01$), considering both OLS and quantile regressions with all τ values (Fig. 1; Table I). As hypothesized by several authors, spectral distance represents a direct effect of environmental properties, thus representing a powerful tool for gradient analysis and species β -diversity comparisons [19]–[21].

OLS regression estimate of mean decay rate was only ca. half the decay rate indicated by the upper quantiles (Table I). Moreover, the intercept value, representing the similarity reached

when the spectral distance approaches zero, was very low (0.20) compared with the intercepts achieved by the upper quantiles, which detected high species similarity when habitats are more similar.

III. DISCUSSION: METHODOLOGICAL ASPECTS OF QUANTILE REGRESSIONS APPLIED TO SPECTRAL DISTANCE DECAY

We claim that one should carefully check the regression model, considering the upper quantiles, before dismissing statistically nonsignificant relations based on OLS regression estimates. For instance, in the previously cited example, a low slope may be found by OLS regression, but a higher one may be detected by stressing maximum differences in species similarity, e.g., by quantile regression, to individuate the extremes of the environmental gradients, which should control differences in species composition and richness among sites [20].

A basic question concerns the additional information content brought by spectral distances over more conventional and easily computable spatial distances. In fact, from a biological point of view, spatial distance generally acts as an ecological limiting factor accounting for the dispersal of both plant and animal species [22]. Nevertheless, methods based on distance decay do not necessarily account for environmental heterogeneity, particularly in heavily fragmented landscapes [23]. As an example, Tuomisto *et al.* [6], studying plant diversity in Amazonia, found that spatial distance accounted for only a small fraction of variance in species similarity, while environmental variation accounted for a much larger one.

One major open question concerns the main possible causes of the huge amount of noise occurring within the distance decay scatterplot shown in this letter, which has been found in other similar applications (see, e.g., [7]). This noise may derive from three main reasons: 1) the grain of the sampling units; 2) the mismatch between the grain of the Landsat ETM+ image and the field data; and 3) the difference between the time of satellite image acquisition and the field survey period. Considering the dimension of the sampling units, if the grain is small enough, one might expect that samples should share no or few species, even if their ecological properties are the same, thus provoking a high amount of points within the scatterplot with low species similarity, even when the spectral distance is small [24], [25]. Second, the pixel dimension of the Landsat ETM+ image (28.5 m) does not match that of the sampling units (10 m). Large pixels within a remotely sensed image are expected to be mixed [26]–[28], resulting in more diluted spectral information; this should inevitably impact the distance measure with a flattening effect and a reduced capability in discriminating sites within the scatterplot. Finally, the temporal gap of 5 years between images and field data may impact, in some cases, the achieved scatterplot. In fact, while forest habitats are expected to be relatively stable, some classes like cropland and badland vegetation types could show considerable human-induced disturbance. However, achieving satellite data that temporally match field data is an expensive task [29].

When this noise occurs, quantile regressions represent the most robust but straightforward approaches for modeling

the complexity of ecological data, particularly when dealing with data collected in the field [6], [18]. Of course, quantile regressions allow to estimate the decay of all portions (quantiles) of species similarity. Nonetheless, in distance decay modeling, the maximum species similarity decay is crucial, having a solid ecological foundation. In fact, as stressed by Rocchini *et al.* [20], the extremes of the environmental gradient are important in controlling differences in species composition among sites [12].

Soininen *et al.* [5], dealing with distance decay modeling, recently expressed a desire for more sophisticated analytical methods for accounting for environmental distance and decoupling it from spatial distance. This issue seems to be readily solved by applying quantile regression to spectrally based distance decay.

ACKNOWLEDGMENT

We would like to thank Editor-in-Chief W. Emery, an anonymous Associate Editor, and two anonymous referees for the improvements made to the previous draft of this letter. A. Chiarucci and G. Bacaro managed the MoBiSIC field survey project and provided useful insights.

REFERENCES

- [1] R. K. Colwell and J. A. Coddington, "Estimating terrestrial biodiversity through extrapolation," *Philos. Trans. R. Soc. Lond. B, Biol. Sci.*, vol. 345, no. 1311, pp. 101–118, Jul. 1994.
- [2] R. H. Whittaker, "Evolution and measurement of species diversity," *Taxon*, vol. 21, pp. 213–251, 1972.
- [3] P. Koleff, K. J. Gaston, and J. J. Lennon, "Measuring beta diversity for presence-absence data," *J. Anim. Ecol.*, vol. 72, no. 3, pp. 367–382, May 2003.
- [4] J. C. Nekola and P. S. White, "The distance decay of similarity in biogeography and ecology," *J. Biogeogr.*, vol. 26, no. 4, pp. 867–878, Jul. 1999.
- [5] J. Soininen, R. McDonald, and H. Hillebrand, "The distance decay of similarity in ecological communities," *Ecography*, vol. 30, no. 1, pp. 3–12, Feb. 2007.
- [6] H. Tuomisto, A. D. Poulsen, K. Ruokolainen, R. C. Moran, C. Quintana, J. Celi, and G. Cañas, "Linking floristic patterns with soil heterogeneity and satellite imagery in Ecuadorian Amazonia," *Ecol. Appl.*, vol. 13, no. 2, pp. 352–371, Apr. 2003.
- [7] D. Rocchini, "Distance decay in spectral space in analyzing ecosystem β -diversity," *Int. J. Remote Sens.*, vol. 28, no. 11, pp. 2635–2644, Jan. 2007.
- [8] H. K. Schröder, H. E. Andersen, and K. Kiehl, "Rejecting the mean: Estimating the response of fen plant species to environmental factors by non-linear quantile regression," *J. Veg. Sci.*, vol. 16, no. 4, pp. 373–382, Aug. 2005.
- [9] R. Koenker and G. Bassett, Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, Jan. 1978.
- [10] B. S. Cade and Q. Guo, "Estimating effects of constraints on plant performance with regression quantiles," *Oikos*, vol. 91, no. 2, pp. 245–254, Nov. 2000.
- [11] B. S. Cade and B. R. Noon, "A gentle introduction to quantile regression for ecologists," *Front. Ecol. Environ.*, vol. 1, no. 8, pp. 412–420, Oct. 2003.
- [12] B. S. Cade, J. W. Terrell, and R. L. Schroeder, "Estimating effects of limiting factors with regression quantiles," *Ecology*, vol. 80, no. 1, pp. 311–323, Jan. 1999.
- [13] C. J. Tucker, D. M. Grant, and J. D. Dykstra, "NASA's global orthorectified Landsat data set," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 3, pp. 313–322, 2004.
- [14] C. Song, E. Woodcock, K. C. Seto, M. P. Lenney, and S. A. Macomber, "Classification and change detection using Landsat TM data: When and how to correct atmospheric effects?" *Remote Sens. Environ.*, vol. 75, no. 2, pp. 230–244, Feb. 2001.
- [15] R. Koenker and K. Hallock, "Quantile regression," *J. Econ. Perspect.*, vol. 15, pp. 143–156, 2001.
- [16] N. Gotelli and A. Ellison, *A Primer of Ecological Statistics*. Sunderland, U.K.: Sinauer Assoc., 2004.
- [17] R. Koenker, *Quantreg: Quantile regression. R package version 4.10*, 2007. [Online]. Available: <http://www.r-project.org>
- [18] B. S. Cade, B. R. Noon, and C. H. Flather, "Quantile regression reveals hidden bias and uncertainty in habitat models," *Ecology*, vol. 86, no. 3, pp. 786–800, Mar. 2005.
- [19] H. Nagendra, "Using remote sensing to assess biodiversity," *Int. J. Remote Sens.*, vol. 22, no. 12, pp. 2377–2400, 2001.
- [20] D. Rocchini, S. Andreini Butini, and A. Chiarucci, "Maximizing plant species inventory efficiency by means of remotely sensed spectral distances," *Glob. Ecol. Biogeogr.*, vol. 14, no. 5, pp. 431–437, Sep. 2005.
- [21] G. M. Foody and M. E. J. Cutler, "Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks," *Ecol. Model.*, vol. 195, no. 1/2, pp. 37–42, 2006.
- [22] G. Chust, J. Chave, R. Condit, S. Anguilar, S. Lao, and R. Pérez, "Determinants and spatial modeling of tree β -diversity in a tropical forest landscape in Panama," *J. Veg. Sci.*, vol. 17, no. 1, pp. 83–92, 2006.
- [23] M. W. Palmer, "Distance decay in an old-growth neotropical forest," *J. Veg. Sci.*, vol. 16, no. 2, pp. 161–166, Apr. 2005.
- [24] A. Chao, R. L. Chazdon, R. K. Colwell, and T.-J. Shen, "A new statistical approach for assessing similarity of species composition with incidence and abundance data," *Ecology Lett.*, vol. 8, no. 2, pp. 148–159, Feb. 2005.
- [25] O. Steinitz, J. Heller, A. Tsoar, D. Rotem, and R. Kadmon, "Environment, dispersal and patterns of species similarity," *J. Biogeogr.*, vol. 33, no. 6, pp. 1044–1054, Jun. 2006.
- [26] P. Fisher, "The pixel: A snare and a delusion," *Int. J. Remote Sens.*, vol. 18, no. 3, pp. 679–685, Feb. 1997.
- [27] D. Rocchini, "Effects of spatial and spectral resolution in estimating ecosystem α -diversity by satellite imagery," *Remote Sens. Environ.*, vol. 111, no. 9, pp. 423–434, Dec. 2007.
- [28] C. Small, "The Landsat ETM+ spectral mixing space," *Remote Sens. Environ.*, vol. 93, no. 1/2, pp. 1–17, Oct. 2004.
- [29] S. R. Loarie, L. N. Joppa, and S. L. Pimm, "Satellites miss environmental priorities," *Trends Ecol. Evol.*, vol. 22, no. 12, pp. 630–632, Dec. 2007.