PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

River density and landscape roughness are universal determinants of linguistic diversity

Jacob Bock Axelsen and Susanna Manrubia

Proc. R. Soc. B 2014 **281**, 20133029, published 16 April 2014

PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Axelsen JB, Manrubia S. 2014 River density and landscape roughness are universal determinants of linguistic diversity. *Proc. R. Soc. B* **281**: 20133029. http://dx.doi.org/10.1098/rspb.2013.3029

Received: 19 November 2013
Accepted: 18 March 2014

Subject Areas:
ecology, theoretical biology, environmental science

Keywords:
linguistic diversity, language–area relationship, river systems, landscape roughness, environment, biogeography

Author for correspondence:
Susanna Manrubia
e-mail: scmanrubia@cab.inta-csic.es

†Present address: Department of Zoology, Tel Aviv University, Tel Aviv, Israel.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2013.3029 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing

River density and landscape roughness are universal determinants of linguistic diversity

Jacob Bock Axelsen† and Susanna Manrubia

Centro de Astrobiología, INTA-CSIC, Ctra. de Ajalvir km 4, Madrid, Torrejón de Ardoz 28850, Spain

Global linguistic diversity (LD) displays highly heterogeneous distribution patterns. Though the origin of the latter is not yet fully understood, remarkable parallelisms with biodiversity distribution suggest that environmental variables should play an essential role in their emergence. In an effort to construct a broad framework to explain world LD and to systematize the available data, we have investigated the significance of 14 variables: landscape roughness, altitude, river density, distance to lakes, seasonal maximum, average and minimum temperature, precipitation and vegetation, and population density. Landscape roughness and river density are the only two variables that universally affect LD. Overall, the considered set accounts for up to 80% of African LD, a figure that decreases for the joint Asia, Australia and the Pacific (69%), Europe (56%) and the Americas (53%). Differences among those regions can be traced down to a few variables that permit an interpretation of their current states of LD. Our processed datasets can be applied to the analysis of correlations in other similar heterogeneous patterns with a broad spatial distribution, the clearest example being biological diversity. The statistical method we have used can be understood as a tool for cross-comparison among geographical regions, including the prediction of spatial diversity in alternative scenarios or in changing environments.

1. Introduction

Primal linguistic diversity (LD) may stem from a myriad of environmental and cultural factors. Several quantitative analogies with ecological diversity patterns [1,2] are an indirect evidence in support of the pivotal role likely played by the environment in explaining the distribution of linguistic groups. Conspicuous examples are the increase in language richness with sampling area [3]—in a fashion equivalent to the species–area relationship for biological species [4]—and the latitude diversity gradient, which was early described for species [5] and much later identified for the density of cultures [6] and that of languages in North America [7] and Africa [8].

Some studies have attempted to quantify the direct relationship between linguistic or cultural diversity and the environment. One of the first investigations related LD to climatic variation at the scale of whole countries [9], concluding that a key determinant of LD is the subsistence strategies of human groups. Later, a positive correlation between cultural diversity and temperature and rainfall patterns was described [6], and evidence that mean annual temperature is only relevant at the macro scale was presented. Also landscape elevation was shown to positively correlate with LD [10]. There is thus abundant evidence of the effect of environmental variables on LD, though a consensus on which mechanisms dominate the build up of LD at the local scale has not been achieved [11]. The fact that several parameters are correlated with LD points to a complex underlying process significantly affected by a number of different variables and perhaps dependent on historical, economical and cultural facts, as several previous analyses have discussed [7,9,6].

Some of the causes underlying language demise are more clearly understood, because this is an ongoing process severely affected by cultural practices.

A comparison of the extinction risk of birds and mammals with languages reveals that the latter (especially small languages) are more threatened than species [10]. At present, around 7000 descendants [12,13] of an estimated peak number of about 30 000 simultaneous coexisting languages [12] remain. The process of language death often involves a competition whereby speakers are forced to or spontaneously abandon their native language in favour of a foreign language [14,15]. Historically, this process has affected different world regions at different times and with varying strength, causing an unequal preservation of linguistic stocks. Nowadays, high LD concentrates in Africa, Asia and the Pacific, while the Americas and Europe have already suffered severe losses: Africa, for instance, with an area almost twice that of South America, harbours five times more languages.

In this contribution, we focus our analysis on the relationship between environmental variables and LD, with the additional consideration of human population density. Our working hypothesis is that the remarkable heterogeneity observed in current LD should contain essential information on which environmental variables are universal and which others are unique in explaining the composition of different world regions. At present, LD is mapped to an unprecedented degree of detail [16], as is the planet and its environmental state and dynamics. Data obtained from geographical information systems have quantified variations in rain, temperature or productivity, e.g. in time and in space with very high precision. We take advantage of all these data to revisit the question of which are the factors that mostly affect global LD. To this end, we apply an unbiased statistical method that yields the relative weight contributed by each variable and takes into account the cross-dependencies within the set of variables. We begin with 14 variables and identify a reduced set (precipitation, temperature, landscape roughness and rivers) with which language abundance is highly correlated. Among those tested, the density of rivers and landscape roughness appear as key factors, not explicitly considered before, in explaining high LD. The method allows a quantitative comparison between large world regions that bear highly dissimilar abundances and leads to the identification of variables that have been particularly relevant in specific regions, permitting an informed analysis of the differences. We conjecture that the information so gathered may assess future losses of LD at a spatially explicit scale.

2. Material and methods

(a) Language database

The global language diversity has been collected by SIL International (see http://www.ethnologue.com/), mapped by Global Mapping International into the World Language Mapping System (http://www.gmi.org/wlms/index.htm) and compiled in an extensive database called Ethnologue [13]. Nowadays, it includes standardized ISO codes for about 6900 extant languages and 700 extinct ones.

(b) Mapping of data

Electronic supplementary material, figure S1 presents an example of the high-resolution polygons corresponding to the territory spanned by the speakers of a language. In order to study the variation of LD across the world, care should be taken to analyse equal-size areas independently of latitude and longitude. This problem has been solved by using an area-preserving projection (electronic supplementary material, figure S2). To avoid insularity effects which may confound the natural spread of a language over a territory, we included in our study only those languages found within the 100 largest landmasses on the Earth (see electronic supplementary material, figure S3). The choice of an optimal spatial scale to represent data and quantify the correlations between variables and LD has been based on a maximum entropy principle (see electronic supplementary material, figure S4), which settled that scale at 222×222 km^2 cells (figure 1). Only one previous study used a comparably high resolution [8], but only for Africa.

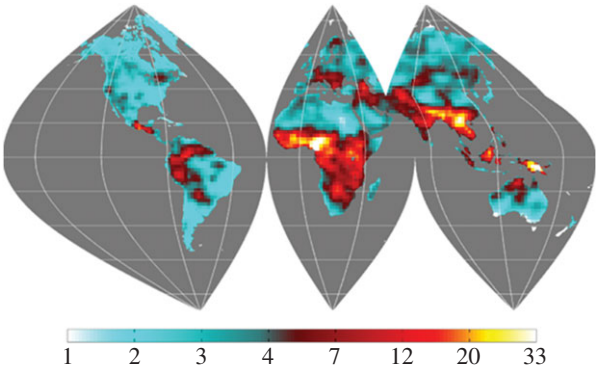

Figure 1. Current heterogeneity of LD. The 100 largest landmasses are depicted in this map, which is a coarse-grained representation of actual linguistic domains in squared areas of size 222×222 km^2 (see electronic supplementary material, figure S3). The sinusoidal projection is area-preserving, maintaining, in particular, the circumference of the Earth at each latitude (see electronic supplementary material, figure S2).

(c) Environmental variables

We selected 14 (non-independent) environmental variables (table 1) and downloaded the corresponding data from several different sources. Detailed explanations for all datasets are provided as electronic supplementary material, S1.

(i) Vegetation

As a measure of seasonal vegetation dynamics, we used the remote sensed 'fraction of absorbed photosynthetically active radiation' (or FAPAR [17]), a relative quantity at a high temporal and spatial resolution that takes values between 0 and 1. Data were downloaded from the European Commission (http://fapar.jrc.ec.europa.eu/Home.php). Processed data are represented in electronic supplementary material, figure S5.

(ii) Precipitation and temperature

The CRU TS v. 2.0 climate dataset [18] from the Climate Research Unit, University of East Anglia (UK), formed the basis of the climate data layer. As for vegetation, we considered average, maximum and minimum values of precipitation (electronic supplementary material, figure S6) and temperature (electronic supplementary material, figure S7) inside each cell.

(iii) Rivers and lakes

The Global Self-consistent, Hierarchical, High-resolution Shoreline dataset [19], available from the National Oceanic and Atmospheric Administration USA, formed the basis of our river density data layer. The number of river branches within each cell yielded the density of rivers (electronic supplementary material, figure S8), while distances to lakes larger than 1000 km^2 yielded a map of proximity to large masses of fresh water (electronic supplementary material, figure S9).

Table 1. Summary of regression results for fits to continental regions. If the factor is found insignificant, the symbol '*' is used; significant factors with either positive or negative weight are indicated with '+' or '−', respectively. We define a fundamental set of factors containing the variables in italic: precipitation, temperature, roughness and density of rivers. Origin and processing of data are described in Material and methods; world maps can be found as electronic supplementary material, figures S5–S10, S12 and S13. A quantitative measure of the significance of each variable in the four regions studied has been obtained by measuring the cumulative values of the corresponding kernel densities (as shown in electronic supplementary material, figures S19–S22; significance values are compiled in electronic supplementary material, table S2).

factors	Africa	extended Asia	America	Europe
average chlorophyll	*	*	*	*
minimum chlorophyll	*	*	+	*
maximum chlorophyll	+	*	*	*
average precipitation	+	+	*	+
minimum precipitation	−	−	+	*
maximum precipitation	−	*	+	−
average temperatures	+	+	*	*
minimum temperatures	−	−	*	*
maximum temperatures	+	*	*	*
roughness	+	+	+	+
altitude	*	*	+	−
density of rivers	+	+	+	+
distance to lakes	*	−	*	+
population density	*	−	−	+

(iv) Landscape elevation

Land elevation was obtained from a dataset of global bathymetry and elevation data at high resolution called SRTM30_Plus [20] by the Scripps Institution of Oceanography, UCSD, USA (electronic supplementary material, figure S10).

(v) Landscape roughness

Calculation of roughness required substantial data processing of the land elevation dataset. Roughness was quantified as a function of the Hurst exponent [21] at each point in the SRTM30 database (every 1 km), calculated in 100 km length transects (to minimize correlations between adjacent cells) and averaged over eight different directions (see electronic supplementary material, figure S11). The resulting roughness map, where each cell contains an average over 1600 points, is represented in electronic supplementary material, figure S12.

(vi) Population density

The population density index was obtained from the Center for International Earth Science Information Network, Columbia University, USA; and from Centro Internacional de Agricultura Tropical, 2005. Gridded Population of the World v. 3 (GPW v. 3): Population Density Grids. Palisades, NY: Socioeconomic Data and Applications Center, Columbia University (http://sedac.ciesin.columbia.edu/gpw/). Data are represented in electronic supplementary material, figure S13.

(d) Correlations between datasets

The environmental variables considered are not mutually independent. In order to control for spurious correlations and to identify variables with a high explanatory power, we quantified the cross-dependencies between every pair of factors. First, the histograms of all variables have been rescaled so as to maximize their meaningfulness (electronic supplementary material, figure S14). The global and partial correlations between all possible pairs of rescaled datasets have been calculated from the corresponding covariance matrix (see electronic supplementary material, S2 and figure 2), and a principal components analysis has been performed (electronic supplementary material, figure S15). As an additional example, the particular case of partial correlations for Africa and the Americas is shown in electronic supplementary material, figures S16 and S17.

(e) Bayesian statistical model

The core of the method to determine the relevance of each variable consists of the application of a Bayesian model with linear terms using Markov chain Monte Carlo sampling (see electronic supplementary material, S3). This procedure returns the estimated weights of a linear combination of the environmental factors that yield the observed language richness (see electronic supplementary material, figure S18 for the implementation of the method, table S1 for a convergence assessment and figures S19–S22 for the kernels obtained). The fitting protocol has been repeated for four continental regions: Europe (west of the Ural Mountains), the American continent, Africa and extended Asia (Asia, the Pacific and Australia). There is a natural separation among those regions caused by mountain ranges and oceans, as well as important historical differences. Beyond these causes, the division into four regions is supported by a quantitative divide arising from the study of languages–area diversity curves (see below).

(f) Spatial correlations

In order to assess to what extent space is relevant to explain LD at the scale of our basic cells, we have quantified spatial correlations in the residuals (difference between logarithms of data and fit values according to the method above, see electronic supplementary material, figure S23 and table S3) by means of Moran's I [22]. Relevant definitions are included as the electronic supplementary material, S4.

(g) Languages–area curves

We have calculated languages–area curves in a manner analogous to the estimation performed in ecology for species–area curves [4], through nested areas of increasing size. Results are represented in figure 3, where each data point stands for the average number of languages $\langle N(a)\rangle$ found within land plots of area size a. In addition to the average number of languages, we have calculated the distribution $p(N)$ of abundances for each value of a (data not shown). The calculation has been repeated

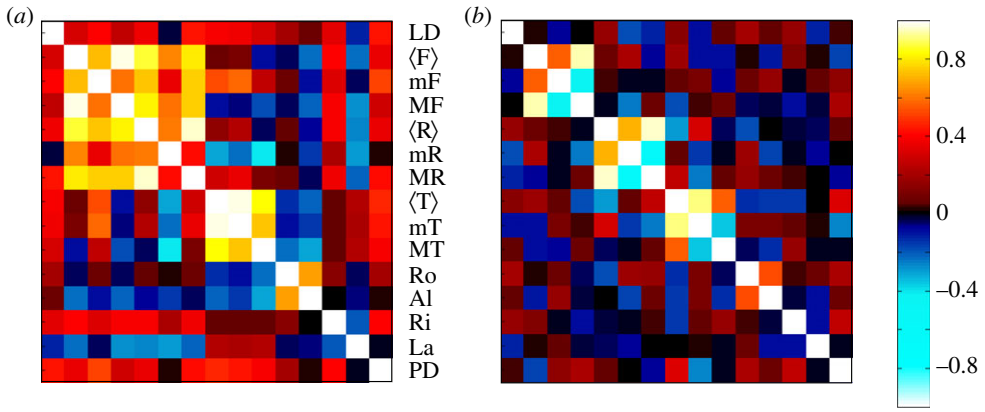

Figure 2. Correlation matrices for all data layers and all grid cells (size 222 × 222 km^2). The entries are Pearson correlations, values as shown in the colour bar. LD, linguistic diversity; ⟨F⟩, average FAPAR; mF, minimum FAPAR; MF, maximum FAPAR; ⟨R⟩, average precipitation; mR, minimum precipitation; MR, maximum precipitation; ⟨T⟩, average temperature; mT, minimum temperature; MT, maximum temperature; Ro, roughness; Al, altitude; Ri, river density; La, distance to nearest lake; PD, population density. Note that the linguistic diversity (LD) is located in the first column and the first row. *p*-Values above 5% significance level are black. Panel (*a*) shows global correlations and panel (*b*) shows partial correlations.

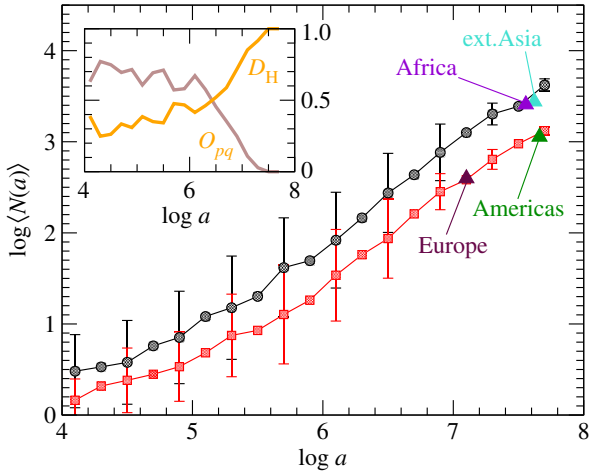

Figure 3. The languages–area relationship. Curves in the main plot represent the average number of languages ⟨*N*(*a*)⟩ encountered within areas *a* of various sizes. Black circles correspond to the LD in regions of Africa and extended Asia (Asia, the Pacific and Australia); red squares stand for the joint region of Europe and Americas. The standard deviation is shown for representative cases. The total richnesses of Africa, extended Asia, Europe and the Americas, are represented as coloured triangles. The inset quantifies the similarity between the distributions *p*(*N*) for the two diversity curves through the overlap O_{pq} between two distributions and Hellinger's distance D_H, which is a measure of dissimilarity (electronic supplementary material, section S5).

for several different continental regions, eventually permitting a global divide between high- and low-diversity regions. For a given value of *a*, thus, two distributions *p*(*N*) are obtained, one for each of the curves in figure 3. These distributions are compared through two different metrics (the overlap between distributions and Hellinger's distance between curves) to quantify their degree of similarity (see electronic supplementary material, S5 and inset of figure 3).

3. Results

(a) Correlations between variables and global linguistic diversity

Mutual dependencies between all possible pairs of variables have been quantified through the covariance matrix. In this approach, LD is treated as an additional variable. A summary of results averaged over all possible 222 × 222 km^2 cells is represented in figure 2.

Overall, seasonally varying factors are strongly correlated internally, as visualized through the formation of clusters along the diagonal in figure 2*a*. Expected correlations, as those between average vegetation and minimum and maximum vegetation, or rainfall and temperature, are recovered by the analysis. This indicates not only a correlation in time (seasons) but also a dependence of the parameters describing, for example, the average vegetation versus the minimum vegetation. Roughness and altitude are robustly related because rougher terrains are typically located at higher altitude [23]. LD displays high positive correlations with most variables measured, with the exception of minimum precipitation and altitude. It shows also a negative correlation with distance to lakes.

As direct dependencies between LD and environmental variables are masked by cross-correlations within the whole set, we resort to partial correlations in order to uncover *bona fide* dependencies. Partial correlations between all pairs of variables are shown in figure 2*b*. We observe that the number of large correlations declines drastically, most notably the correlation between vegetation and rainfall. Furthermore, the correlations between LD and everything else also drop significantly. The dependence of LD on vegetation and population density seems to vanish, which suggests that the dependence with the latter two factors apparent in figure 2*a* should be explained by a more basic subset of the remaining variables. Global LD maintains significant positive correlations with average precipitation (in agreement with [6]), but also with river density and roughness.

To simplify further the matrices in figure 2, we have calculated their principal components (electronic supplementary material, figure S15). This analysis shows that population density and rivers are the two most proximate points for both total and partial correlations. Together with the result on partial correlations, this is a robust indication that dense river systems are fundamentally related to high LD.

(b) Correlation between variables and regional linguistic diversity

The analysis above has been repeated for several major continental regions, revealing significant differences between

Table 2. Table of cross-correlations obtained by applying the fitted weights w using all data layers β from one region to another, as indicated. The diagonal stands for the explainability of LD in each region with the variables used in this study. The uncertainty in the reported numbers is below 1% in all cases.

applied to and fit result	w_{Africa}	$w_{ext.Asia}$	w_{Europe}	$w_{Americas}$
β_{Africa}	*0.80*	0.72	0.37	0.64
$\beta_{ext.Asia}$	0.58	*0.69*	0.37	0.52
β_{Europe}	0.46	0.39	*0.56*	0.42
$\beta_{Americas}$	0.45	0.45	0.14	*0.53*

environmental factors explaining diversity in the Americas, Europe, Africa and a large region that we label extended Asia and comprises Asia, Australia and the Pacific.

As an example, the analysis of partial correlations between all pairs of variables for Africa and the Americas is displayed in the electronic supplementary material, figure S16. A clearer representation of the variables affecting LD in these two regions is shown in the electronic supplementary material, figure S17, where it can be seen that only river density and roughness are positively correlated to LD in both regions. Other significant factors are exclusive of either region. In the case of precipitations, correlations with LD are significant, but of fully different sign. The different dependence on the remaining variables can be interpreted as measurable evidence of the different regimes where the two regions are with respect to the dynamics of their LD. A clear indication is found in the reciprocate response for rainfall. In the Americas, high diversity spots correspond to territories where indigenous cultures remain. Apparently, these are found in mountainous areas or deep forests with less fortunate rainfall patterns. In Africa, more rainfall is correlated to high LD on average, but minimum and maximum precipitation levels have adverse effects on LD (despite its overall enhancing effect on population density).

The analysis of significant factors for the four regions is summarized in table 1 (see also the electronic supplementary material, table S2), and further details can be found in the electronic supplementary material, figures S19–S22. Among the environmental data analysed, only roughness and rivers are found universally significant. Beyond roughness and rivers, the combination of variables that explains LD in different regions seems to be unique. Africa is the region where environmental variables seem to play the largest role in explaining LD: the fit of model-to-data reaches 80%. The explainability of the other regions is not as high, decaying to 53% for the Americas (table 2).

The explainability of African LD at the spatial scale studied is maintained if, instead of the whole set of variables, we use a subset formed by rainfall, temperatures, roughness and rivers. This fundamental set of factors explains 66% of the language variability in extended Asia, 53% in Europe and 48% in America. Using all factors only gives a slightly better fit (compare table 2 and the electronic supplementary material, table S4).

(c) Spatial autocorrelations

Our previous analysis has not taken into account the spatial structure of LD for the sake of simplicity and in order to permit a straight identification of the role played by each of the environmental variables studied. Including spatial correlations would significantly increase the number of parameters of the model, likely conferring it a higher predictive capacity but at the same time confounding the effect of main variables. This nonetheless, it is important to quantify the relevance of spatial autocorrelations. The difference between the observed LD and the prediction of the model (i.e. the absolute residuals) has been calculated for each individual 222×222 km^2 cell (see electronic supplementary material, figure S23). Correlations between neighbouring cells appear locally, especially in areas of high LD and typically around large rivers (Niger, Ganges and Rhine). Significant negative autocorrelations also remain in a few areas, as in the Amazon basin and the Argentinian Patagonia, indicating simultaneously an LD below expectations. These correlations are short-ranged, indicating that the elementary cell we have selected closely coincides with the typical scale at which spatial correlations are lost.

In agreement with the above, spatial autocorrelations in the residuals at the continental scale are non-significant. Measures of Moran's I and the z-score for each of the four main regions considered demonstrate that spatial autocorrelations in the residuals are not significant at the 5% level (electronic supplementary material, table S3).

(d) Mutual explainability of different continental regions

To investigate the universality of the weights w, we looked at how much of the, for example, European LD would be explained by using the weights obtained for African LD. In table 2, we show the mutual explainability for the w's including all variables. Table 2 should be read as follows: when using the weight factors found by regression of LD in, for example, Africa to form expectations of the LD in, for example, Europe we see that our expectations match the data at 46%. The rows have a maximum in the diagonal because the maximal degree of explanation is achieved by fitting the local LD to the corresponding local environmental factors. This explains that, for example, $w_{Americas}$ applied to African LD gives a relatively high number, but the result is much below the actual fit for Africa. The most universal weight factors resulted from the fit to the African LD, as the column labelled 'Africa' shows the overall highest correlations between expectations and observations. The case of European diversity is the opposite, because w_{Europe} have the lowest explanatory power when applied to other regions.

These results are qualitatively robust if we perform our fits with the reduced set of variables (rainfall, temperature, roughness and rivers) identified as significant for Africa in our previous analysis, as summarized in the electronic supplementary material, table S4.

(e) Languages–area curves

The application of regression weights from Africa to Asia and *vice versa* yields the highest mutual explainability among all possible pairs of regions (table 2). This agrees with Africa and Asia presenting the highest LD, hosting 2564 and 2762 living languages, respectively. Europe and the Americas are

not that similar regarding their mutual explainability, with Europe appearing as the most singular region. They have a comparable LD (Europe hosts 396 languages and the Americas 1132), with a density that is about threefold lower on the average than that of the rich continental regions. This fact can be visualized in a different way if we represent the languages–area relationship for Africa and extended Asia and that for Europe and the Americas separately (figure 3). A significant difference in LD occurs for all values of the area, but for $a > 10^6$ km^2 the two distributions $p(N)$ become fully separated. The difference in average richness is consistently maintained for all values of a: a region of a fixed size holds about a threefold larger LD in Africa and extended Asia than in Europe and America. The calculation was repeated for the Americas versus the Eastern hemisphere, with no significant difference (data not shown).

(f) Current divide and future diversity

The high LD in Africa and extended Asia, together with the remarkable explainability of their LD on the basis of environmental factors suggests that those regions might be closer to a past state of peak LD. The LD of the Americas and Europe has been severely affected by colonization and demography, which among others should contribute to explain present LD in those regions. In a sense, the LD of the Americas and Europe could be interpreted as the future of currently rich regions in a scenario where LD will continue decreasing.

A rough way of inferring the future global LD is to use the regression weights obtained for the Americas and apply them to the remaining continents (electronic supplementary material, figure S24): a global drop in LD is expected overall. This analysis can be made quantitative by adding the number of languages predicted from the languages–area curve to occupy each cell. The statistical model yields 3700 languages remaining (half of present diversity). This prediction is in reasonable agreement with the most accepted common estimate of the future number of languages, which states that 3500 languages will be extinct within the century [24].

4. Discussion

We have quantified the correlation between a set of 14 environmental variables and global LD at a meaningful and high spatial resolution. We have used data with the largest available precision and evaluated mutual and absolute correlations by means of an unbiased statistical procedure that relies on a Markov chain Monte Carlo model. The method allows a straight incorporation of new data and can be simply used to evaluate differences between regions and to infer alternative scenarios by means of the crossed use of fitted weights.

Our results support that LD can be explained to a good extent by means of environmental variables. The clearest example is that of Africa, where up to 80% of its LD is explained by a subset of the environmental variables used in this study. Population density, in particular, plays no role in African LD. In other regions, such as the Americas, low population density, high altitude and low vegetation are correlated to LD. We interpret these correlations as a result of the population centres created by speakers of European languages (found mostly in coastal areas, where population density is high), which have either replaced or purged native languages. Indigenous peoples of America may have mostly survived in low vegetation highlands. Population density is thus the proximate cause explaining the LD of a continent that has been under recent heavy demographical influence from the outside. Though to a lesser extent, also the LD of Asia, Australia and the Pacific is negatively correlated to population density. Europe is an exception in this respect, because its LD (significantly lower than that of Africa and Asia, and comparable to that of the Americas) is positively correlated to population density. This points to demographic evolution as a relevant variable in explaining European LD. This nonetheless, Europe and the Americas are the less explainable regions on the basis of environmental variables. Additional variables and effects not considered in this study will be needed to fully explain their LD patterns. On the one hand, we have shown that spatial autocorrelations are non-significant at the continental level, though their introduction should improve the predictive power of the model at the local scale. On the other hand, cultural variables are expected to be of relevance, as previous studies suggest. The spread of linguistic groups, for instance, has been shown to depend on political complexity [25], this being a variable that would be worth including in an extended analysis of global LD. Other recent historical factors such as warfare and colonization, and ecological variables such as pathogens may have been relevant in shaping LD in particular regions. Factors not considered in this study should explain up to 50% of current LD, and in the light of our results those meaningful ones might be even specific to a time and a region.

Among the variables analysed, only two of them play a universal role in explaining the LD of different continental regions: landscape roughness and river density. Though this appears as an intuitively sensible result, the significance of both variables had not been quantitatively demonstrated in any previous study. It is highly likely that roughness and rivers are at the very origin of linguistic diversification. As with speciation, it can be argued that roughness promotes linguistic isolation and enhances diversification through a mechanism analogous to parapatric speciation. In some instances, also rivers act as physical barriers to group dispersal and similarly cause speciation [26]. But rivers also provide many services such as water, food, protection and means of transportation, thus seemingly mitigating the separating effect of rough landscapes. Water, food or protection can be also ensured near lakes, but distance to lakes does not consistently appear as a relevant factor to explain LD. We conjecture that the key difference between lakes and rivers, and the importance of the latter to promote linguistic diversification, relies in transportation. Water courses have been instrumental in the dispersion of humans [27], of which the fast European settlement process in North America [28] is a prominent example. In ancient times, regions where river branches meet were regions where human groups met and settled. Those contacts have left their imprints in archaeological sites such as the Three Gorges in China or, more explicitly, in actual river trade networks such as that which was still operating in mediaeval Russia [29]. Hence, regions of high river density have acted as social hubs, promoting the interaction among different linguistic groups. It is conceivable that rivers may have boosted LD through a process analogous to genetic recombination [30,31]. This scenario is indirectly supported by observations of how frequent contacts between speakers of different languages may result in

new hybrid languages within a few generations. It has been put forward that rapidly emerging contact languages may then have played a significant role in language evolution [32]. In the scenario here discussed, this mechanism might have turned regions rich in waterways into cradles of LD. If true, this could in part explain the hot spots that we have identified in all continents linked to fluvial density. At present, diversity is conspicuous along the basins of the Niger River (Nigeria), the Ganges River (Bangladesh), the Red River delta (Vietnam), the Sepik River (Papua New Guinea) or the Rhine River delta (Holland).

We have identified a clear world divide between two main regions (Europe and the Americas versus Africa and extended Asia) characterized by different states regarding their LD and the exchangeability of the environmental variables that explain the latter. In the light of the historical events in the last few centuries, that divide can be interpreted in a temporal sense, where regions of high diversity would be in the 'past' and evolve towards the LD patterns of low-diversity regions. With this scenario in mind, we have attempted a rough prediction of how the future spatial distribution of languages might look like after some relaxation period whose duration we cannot estimate with current data and the methods used here. Several ingredients may affect this transient. First, as diversity decreases, it is to be expected that surviving languages, likely spoken by larger communities, would enter a regime of slower competitive dynamics with a decreasing rate of language extinction. Second, changes in world organization steadily modify the main players in these dynamics. Examples are the substitution of colonization by globalization and of persistent, water-mediated movements by fast, long-range transportation. Third, demographic pressure and climatic change might be factors accelerating the extinction of a language, in the same way that they are affecting the biosphere [33]: natural habitats are modified, reduced and fragmented, populations are displaced and, eventually, increasing competition and further weakening of small populations supervenes. There is evidence to support that these small populations may suffer from an accelerated decline, in resemblance to the Allee effect in biology [10,34]. On the other hand, new languages are also emerging in the form of new urban hybrid forms or in slow divergence processes that can be already witnessed in English and Spanish. However, this number is relatively small when compared with the rate of language loss [35].

Language extinction is certainly not so strongly driven by colonization as it was some centuries ago. The process that finishes with one language taking over another one is at present a passive process in many cases. Widespread languages usually perform as dominant languages, and the subjective appreciation of their higher status leads to a non-violent but inexorable extinction of minor coexisting languages [14]. Though eventual extinction is the final fate of most world languages, the time they may take to disappear depends strongly on conservation policies and environmental factors. In the light of our results, an eye should be kept on river systems. Nowadays, owing to climatic change and other human activities—as flow regulation through dams–, river systems have been deeply modified, often losing at once their dual ancient role as providers of fresh water and means of transportation [36].

The significant relation between river systems and LD here uncovered may hopefully open new ways of looking at how languages, and thus human cultures, arise, interact and can be preserved. If linguistic recombination is behind the proliferation of languages along fluvial networks, linguistic phylogenies in those regions may appear as complex ensembles of networked languages [37], with a strong component of horizontal transmission that may eventually call for an analysis of the parallelisms between linguistic and biological evolution beyond the analogy. An active preservation of river systems appears as essential to maintain not only the linguistic heritage of the humankind, but also possible unknown remnants of a rich and diverse cultural past.

Acknowledgement. We thank L. Stone, J. Á. Capitán, J. Aguirre, N. H. Axelsen and E. M. Bock for critical reading of the manuscript, and E. Aboh, J. Bascompte and A. Sánchez for insightful suggestions.

Funding statement. This work was supported by projects from Comunidad de Madrid (MODELICO S2009/ESP-1691), Spanish MINECO (FIS2011-27569) and the Carlsberg Foundation (J.B.A.).

References

1. Maffi L. 2005 Linguistic, cultural, and biological diversity. *Annu. Rev. Anthropol.* **29**, 599–617. (doi:10.1146/annurev.anthro.34.081804.120437)
2. Burnside WR, Brown JH, Burger O, Hamilton MJ, Moses M, Bettencourt L. 2012 Human macroecology: linking pattern and process in big-picture human ecology. *Biol. Rev.* **87**, 194–208. (doi:10.1111/j.1469-185X.2011.00192.x)
3. Gomes MAF, Vasconcelos GL, Tsang IJ, Tsang IR. 1999 Scaling relations for diversity of languages. *Phys. A Stat. Mech. Appl.* **271**, 489–495. (doi:10.1016/S0378-4371(99)00249-6)
4. Preston FW. 1962 The canonical distribution of commonness and rarity: part I. *Ecology* **43**, 185–215. (doi:10.2307/1931976)
5. Humboldt AV. 1824 Voyage aux Régions equinoxiales de Nouveau Continent fait en 1799, 1800, 1801, 1802, 1803 et 1804 por Al. v. Homboldt et A. Bonpland, 8. A la librairie Grecque-Latine-Allemande.
6. Collard IF, Foley RA. 2002 Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evol. Ecol. Res.* **4**, 371–383.
7. Mace R, Pagel M. 1995 A latitudinal gradient in the density of human languages in North America. *Proc. R. Soc. Lond. B* **261**, 117–121. (doi:10.1098/rspb.1995.0125)
8. Moore JL, Manne L, Brooks T, Burgess ND, Davies R, Rahbek C, Williams P, Balmford A. 2002 The distribution of cultural and biological diversity in Africa. *Proc. R. Soc. Lond. B* **269**, 1645–1653. (doi:10.1098/rspb.2002.2075)
9. Nettle D. 1998 Explaining global patterns of language diversity. *J. Anthropol. Archaeol.* **17**, 354–374. (doi:10.1006/jaar.1998.0328)
10. Sutherland WJ. 2003 Parallel extinction risk and global distribution of languages and species. *Nature* **423**, 276–279. (doi:10.1038/nature01607)
11. Gavin MC *et al*. 2013 Toward a mechanistic understanding of linguistic diversity. *Bioscience* **63**, 524–535. (doi:10.1525/bio.2013.63.7.6)
12. Crystal D. 1997 *The Cambridge encyclopedia of language*. Cambridge, UK: Cambridge University Press.
13. Gordon RG, Grimes BF and Summer Institute of Linguistics. 2005 *Ethnologue: languages of the world*, vol. 15. Dallas, TX: SIL International Dallas.
14. Kandler A. 2009 Demography and language competition. *Hum. Biol.* **81**, 181–210. (doi:10.3378/027.081.0305)
15. Abrams DM, Strogatz SH. 2003 Modelling the dynamics of language death. *Nature* **424**, 900. (doi:10.1038/424900a)

16. Lewis MP, Simons GF, Fennig CD (eds). 2013 *Ethnologue: Languages of the World*, 17th edn. Dallas, TX: SIL International. See http://www.ethnologue.com.
17. Gobron N, Knorr W, Belward AS, Pinty B. 2010 Fraction of absorbed photosynthetically active radiation (FAPAR). *Bull. Am. Meteorol. Soc.* Special supplement on State of the Climate in 2009, **91**, S50–S51. (doi:10.1175/BAMS-90-8-Stateofthe Climate)
18. Mitchell TD, Carter TR, Jones PD, Hulme M, New M. 2004 A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901–2000) and 16 scenarios (2001–2100). *Tyndall Centre for Climate Change Research Working Paper* 55, 0 25. See http://www.cru.uea.ac.uk/cru/data/hrg/timm/grid/CRU_TS_2_0.html.
19. Wessel P, Smith WHF. 1996 A global, self-consistent, hierarchical, high-resolution shoreline database. *J. Geophys. Res.* (All Series) **101**, 8741–8743. (doi:10.1029/96JB00104)
20. Becker JJ *et al*. 2009 Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Mar. Geodesy* **320**, 355–371. (doi:10.1080/01490410903297766)
21. Hurst HE. 1951 Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* **116**, 770–808.
22. Moran PAP. 1950 Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23. (doi:10.2307/2332142)
23. Baldassarri A, Montuori M, Prieto-Ballesteros O, Manrubia SC. 2008 Fractal properties of isolines at varying altitude revealing different dominant geological processes on earth. *J. Geophys. Res.* **113**, 09002. (doi:10.1029/2007JE003066)
24. Krauss M. 1992 The world's languages in crisis. *Language* **68**, 4–10. (doi:10.1353/lan.1992.0075)
25. Currie TE, Mace R. 2009 Political complexity predicts the spread of ethnolinguistic groups. *Proc. Natl Acad. Sci. USA* **106**, 7339–7344. (doi:10.1073/pnas.0804698106)
26. Ayres JM, Clutton-Brock TH. 1992 River boundaries and species range size in Amazonian primates. *Am. Nat.* **140**, 531–537. (doi:10.1086/285427)
27. Drake NA, Blench RM, Armitage SJ, Bristow CS, White KH. 2011 Ancient watercourses and biogeography of the sahara explain the peopling of the desert. *Proc. Natl Acad. Sci. USA* **108**, 458–462. (doi:10.1073/pnas.1012231108)
28. Campos D, Fort J, Mendez V. 2006 Transport on fractal river networks: application to migration fronts. *Theor. Popul. Biol.* **69**, 88–93. (doi:10.1016/j.tpb.2005.09.001)
29. Pitts FR. 1978/79 The medieval river trade network of Russia revisited. *Soc. Netw.* **1**, 285–292. (doi:10.1016/0378-8733(78)90025-4)
30. Mufwene SS. 2005 Language evolution: the population genetics way. In *Gene, Sprachen und ihre evolution*, vol. 29 (ed. G Hauska), pp. 30–52. Regensburg, Germany: Universitäts Regensburg.
31. Aboh EO. 2009 Competition and selection: that's all! In *Complex processes in new languages of Creole language library,* vol. 35 (eds EO Aboh, N Smith), pp. 317–344. Amsterdam, The Netherlands: Benjamins.
32. Vennemann Th. 2003 Languages in prehistoric Europe north of the Alps. In *Languages in prehistoric Europe* (Indogermanische Bibliothek, Dritte Reihe), (eds A Bammesberger, Th Vennemann), pp. 319–332. Heidelberg, Germany: Carl Winter.
33. Barnosky AD *et al*. 2012 Approaching a state shift in Earth's biosphere. *Nature* **486**, 52–58. (doi:10.1038/nature11018)
34. Stephens PA, Sutherland WJ, Freckleton RP. 1999 What is the Allee effect? *Oikos* **87**, 185–190. (doi:10.2307/3547011)
35. Crystal D. 2000 *Language death*. Cambridge, UK: Cambridge University Press.
36. Nilsson Ch, Reidy CA, Dynesius M, Revenga C. 2005 Fragmentation and flow regulation of the world's large river systems. *Science* **308**, 405–408. (doi:10.1126/science.1107887)
37. Jourdan C. 1991 Pidgins and creoles: the blurring of categories. *Ann. Rev. Anthropol.* **20**, 187–209. (doi:10.1146/annurev.an.20.100191.001155)