Jul 1st, 12:00 AM

Ecosystem as a Text: Semantic Analysis of the Global Vegetation Pattern

Yuri Svirezhev

Ecosystem as a Text: Semantic Analysis of the Global Vegetation Pattern

Yuri M. Svirezhev

Potsdam Institute for Climate Impact Research, Potsdam
P.O. Box 601203, D-14412 Potsdam, Germany
E-mail: Yuri.Svirezhev@pik-potsdam.de

Abstract: Let us consider a text, which is written by Russian language. At the first level of reception we know only a number of letters. Then the information per one letter I_1=5bits. At the second level of perception we take into account the frequencies of letters then I_2=4.35bits. At the next levels, when we take into account double, triple, etc. correlations, we get I_3=3.5bits, I_4=3bits, etc. Then the redundancy of information at each level, $R_i=1-I_i/I_1$, is equal to R_1=0, R_2=0.13, R_3=0.3, R_4=0.4, etc. By defining the cost of information as a degree of non-redundancy, $C_i=1/(1-R_i)$ we get C_1=1 C_2=1.15, C_3=1.43, C_4=1.67, etc.

Let us consider now a description of the Global Vegetation Pattern (GVP). At the first level of description we have a number of biomes or vegetation types. In accordance with Bazilevich this number is equal to 30 then I_1=4.9bits. At the second level we take into account the relative areas covered by biomes. Then I_2=4.41bits and R_2=0.1, C_2=1.11. At the next level of description we consider the spatial correlations between different pairs of biomes (Bazilevich's biomes map is used). We get the following results: I_3=3.6bits, R_3=0.265, C_3=1.36.

Semantic methods allow comparing two different texts, for instance, the GVP and the spatial distributions of temperature and precipitation.

Let us consider two divisions of one rectangle, which correspond to the indicators A (biomes) and B (different types of climate). The areas of domains induced by these divisions, are S_i^A and S_j^B, $i=\overline{1,n^A}, j=\overline{1,n^B}$. Let the area of intersection S_i^A and S_j^B be S_{ij}, respectively. We define the matrices $\mathbf{P}^{A/B}=\left\| p_{ij}^{A/B}=S_{ij}/S_j^B \right\|$ and $\mathbf{P}^{B/A}=\left\| p_{ij}^{B/A}=S_{ij}/S_i^A \right\|$. The amounts of information, contained in j^{th} row of $\mathbf{P}^{A/B}$ and i^{th} column of $\mathbf{P}^{B/A}$, are equal to $I_j^{A/B}=-\sum_i p_{ij}^{A/B}\ln p_{ij}^{A/B}$ and $I_i^{B/A}=-\sum_i p_{ij}^{B/A}\ln p_{ij}^{B/A}$. The total amounts of information, contained in $\mathbf{P}^{A/B}$ and $\mathbf{P}^{B/A}$ are equal to $I^{A/B}=\sum_j I_j^{A/B}S_j^B$ and $I^{B/A}=\sum_i I_i^{B/A}S_i^A$. The relative measures of information are $h^{A/B}=I^{A/B}/\ln n^A$ and $h^{B/A}=I^{B/A}/\ln n^B, 0\le h^{A/B}, h^{B/A}\le 1$. Then the *information coefficient of correlation* is defined as $\eta = 1-\dfrac{h^{A/B}+h^{B/A}}{2}, 0\le\eta\le 1$. This coefficient for the current climate and the GVP in Siberia is equal to 0.42.

Keywords: Information; Semantics; Global Vegetation Pattern; Climate change.

1. INTRODUCTION

If looking at a standard botanical description of some territory we can see that it contains, firstly, a list of species (types, forms, etc.) of plants represented at the territory, and secondly, the percents of covering, p_i, i.e. the percent of the total territory covered by i^{th} species. This is a typical linguistic construction, in which the alphabet of corresponding language is the names of all species contained in the list. Immediately an idea appears to apply semantic methods to its analysis. To clarify the sentence we show how to do this for some purely linguistic problem.

Let us consider a text, which is written by the Russian language. At the first level of reception we know only a number of letters. Then the information per one letter I_1=5 bits. At the second level of perception we take into account the frequencies of letters then I_2=4.35 bits. At the next levels, when we take into account double, triple, etc. correlations of letters, we get I_3=3.5 bits, I_4=3 bits, etc. At the given level only new non-redundant information has a cost. The repeated information, the transmission of which decreases the probability of its destruction by a noise, is called *redundant*. Then the redundancy of information at each level, $R_i=1-I_i/I_1$, is equal to R_1=0, R_2=0.13, R_3=0.3, R_4=0.4, etc. It means that only 40% of letters are redundant at the fourth level, i.e., 60% of randomly distributed letters are sufficient for the understanding of the text. By defining the cost of information as a degree of non-redundancy, $C_i=1/(1-R_i)$ we get C_1=1, C_2=1.15, C_3=1.43, C_4=1.67, etc. (Volkenstain [1988]).

2. INFORMATION ANALYSIS OF THE GLOBAL VEGETATION PATTERN

Let us consider now a description of the Global Vegetation Pattern (GVP). At the first level of description we have the list of biomes. When the biome name is considered as a letter of alphabet. In accordance with Bazilevich [1969, 1979] this number is equal to 30 (see Tab. 1). The information per one letter is equal to a logarithm of number of biomes or vegetation types, I_1=4.9bits.

At the second level we take into account the relative areas covered by biomes. Let S_i be the area occupied by i^{th} biome, $p_i = S_i / \sum_i S_i$. Then $I_2 = -\sum_i p_i \log_2 p_i = 4.41$ bits. In accordance with Volkenstein the redundancy R_2=0.1 and the cost of information C_2=1.11. At the next level of description we consider the spatial correlations between different pairs of biomes (Bazilevich's biomes map is used). In this case the entropy is a convolution of two-dimensional value. We get the following results: I_3=3.6 bits, R_3=0.265, C_3=1.36. In addition to the area every biome is also characterised by three values: annual productivity, living biomass and dead organic matter. Considering each of them as the second informational level we get

Tab. 1. Different types of vegetation (biomes).

1. Polar desert	16. Dry steppe
2. Tundra	17. Sub-boreal desert
3. Mountainous tundra	18. Sub-boreal saline desert
4. Forested tundra	19. Subtropical semi-desert
5. North taiga	20. Subtropical desert
6. Middle taiga	21. Mountainous desert
7. South taiga	22. Alpine and Sub-alpine meadows
8. Temperate mixed forest	23. Evergreen tropical rain forest
9. Aspen-birch lower taiga	24. Deciduous tropical forest
10. Deciduous forest	25. Tropical xerophytic woodland
11. Subtropical deciduous and coniferous forest	26. Tropical savannah
12. Xerophytic woods and shrubs	27. Tropical desert
13. Forest steppe	28. Mangrove forest
14. Temperate dry steppe (incl. mountainous)	29. Saline land
15. Savannah	30. Subtropical and tropical woodland and tugaj shrubs

By applying the same method we get for

1. Productivity: I_2 = 3.71 bits, R_2 = 0.24, C_2 = 1.32.
2. Biomass: I_2 = 3.26 bits, R_2 = 0.33, C_2 = 1.5.
3. Dead organic matter: I_2 = 4.13 bits, R_2 = 0.16, C_2 = 1.17.

This means that the information about the distributions of biomes productivity, living biomass and dead organics matter is more valuable than the information about the area distribution of biomes. The maximal cost has the information about the distribution of living biomass (C_2=1.5).

3. DETERMINATION GVP BY CLIMATE

A concept about the strong determination of the vegetation locality by such climatic factors as the annual temperature and vegetation (so-called Holdridge's scheme) was always very attractive. Because of this a lot of the Global Vegetation Pattern models were based either on the original Holdridge scheme [1947] or on its different modifications (see, for instance, Monserud, Leemans [1992], Belotelov et al. [1993]). However, the degree of climatic determination depends very strongly, on the one hand, on the used classification of vegetation types, and, on the other hand, on the spatial scale of averaging. For instance, if the standard Walter classification

(Walter [1964, 1968]) of plant types (biomes) was used then the degree of determination dropped drastically (Svirezhev [1997]). The same result gets if we consider a regional vegetation pattern where the effect of global averaging is manifested more weakly. All this forces us to find a new more generalised bioclimatic scheme in order to use it for modelling of the shift of vegetation zones under a climate change.

4. GENERALISATION OF THE BIOCLIMATIC SCHEME

At the beginning we have to introduce a few definitions. Let i (i=1,...,n) be an identifier of some type of vegetation. All the continental area is divided on (n+1) non-overlapping domains V_i (i = 0,1,...,n), so that if the point with co-ordinates (x,y) belongs to domain V_i then i^{th} type of vegetation is at this point. If $i=0$ then this indicates there is not any vegetation at the point. It is obvious that $V_i \cap V_j = \varnothing$. We define the set $\mathbf{V}:\{V_i\}$ as the GVP.

The "climatic space" **C** is defined as a space of local climatic parameters so that **C**: {**c**}. The vector **c**(x, y) is a continuous vector-function of co-ordinates x, y, which describes an ambiguous mapping of the "geographic" space {x, y} onto **C**: $\{x,y\} \overset{\mathbf{c}(x,y)}{\Rightarrow} \mathbf{C}$. It is obvious that these functions map also **V** onto **C** (since V_i= $V_i(x, y)$), the domains V_i in the geographic space are transformed into the domains K_i in the climatic space. In consequence of the ambiguity of the mapping ***c*** *(x, y)* $K_i \cap K_j \neq \varnothing$.

We may define some measure on the product ***V***⊗***C***. It may be a standard entropy measure if the set ***V***⊗***C*** has some more or less regular structure, but if the set has a fractal structure, it will be more correct to introduce Hausdorff's measures. In any case, this value has to reflect a degree of the importance of climatic parameters in the procedure of the separability of different types of vegetation in the climatic space. In other words, what is a measure of determination of the geographic location of biomes by climatic factors?

Let us introduce the concept of the information correlation coefficient h. It is obviously that every *elementary map* (climatic, hydrological, soil and other indicators) may be presented in the form of *division* of some rectangle by n non-overlapping domains. More complex maps including two and more elementary indicators (for instance, climatic + vegetation) can be presented as a superposition of several divisions.

Let us consider two divisions of one rectangle (its area is equal to one), which correspond to the indicators A and B: A and B-divisions. The areas of domains induced by these divisions, are S_i^A and S_j^B, $i = \overline{1,n^A}, j = \overline{1,n^B}$. It is obvious that $\sum_i S_i^A = \sum_j S_j^B = 1$. Let the intersection of S_i^A and S_j^B have the area S_{ij}. Introducing the vectors $\mathbf{S}^A = \{S_i^A,\, i = \overline{1,n^A}\}$ and $\mathbf{S}^B = \{S_j^B, j = \overline{1,n^B}\}$, which can be interpreted as the *abundance* of A *and* B, *respectively*, we define the matrices $\mathbf{P}^{A/B}$ ("A in B")= $\left\| p_{ij}^{A/B} = S_{ij} / S_j^B \right\|$ and $\mathbf{P}^{B/A}$ ("B in A")= $\left\| p_{ij}^{B/A} = S_{ij} / S_i^A \right\|$. It is obvious that $\sum_i p_{ij}^{A/B} = \sum_j p_{ij}^{B/A} = 1$. The value $p_{ij}^{A/B}$ is a fraction of i^{th} domain, generated by A-division, which is contained in j^{th} domain, generated by B-division. The analogous interpretation takes place for $p_{ij}^{B/A}$. The amount of information contained in j^{th} row of $\mathbf{P}^{A/B}$ is equal to $H_j^{A/B} = -\sum_i p_{ij}^{A/B} \ln p_{ij}^{A/B}$; the same one for the i^{th} column of $\mathbf{P}^{B/A}$ is equal to $H_i^{B/A} = -\sum_i p_{ij}^{B/A} \ln p_{ij}^{B/A}$. The total amounts of information contained in $\mathbf{P}^{A/B}$ and $\mathbf{P}^{B/A}$ are equal to $H^{A/B} = \sum_j H_j^{A/B} S_j^B$ and $H^{B/A} = \sum_i H_i^{B/A} S_i^A$, correspondingly. The relative measures of information are $h^{A/B} = H^{A/B} / \ln n^A$ and $h^{B/A} = H^{B/A} / \ln n^B$, $0 \le h^{A/B}, h^{B/A} \le 1$. Then the *information coefficient of correlation*, η, can be defined as

$$\eta = 1 - \frac{h^{A/B} + h^{B/A}}{2}, 0 \le \eta \le 1, \qquad (1)$$

Let us consider the situation when A and B-divisions are identical, then $\eta = 1$.

We introduce into consideration the following criteria:

Criterion 1. If $H^{A/B} > H^{B/A}$, then the hierarchical subordination the map A is beyond than the map B. If $H^{A/B} < H^{B/A}$ then the hierarchical subordination the map B is beyond than the map A.

Criterion 2. The same situation with the substitutions: $h^{A/B} \to H^{A/B}$, $h^{B/A} \to H^{B/A}$.

Criterion 3. If h is close to 1, then two maps (A and B*)* are equivalent; for the complete description one of these maps is sufficient. If h is close to 0 then two maps are independent; for the complete description both the maps are necessary.

5. SELECTIVE PRINCIPLE FOR THE MOST SUITABLE GVP

In order to apply these criteria to our case we shall consider two subdivisions of the plane $\{x,y\}$on the domains S_i and C_i correspondingly. Note that if the first subdivision is not a problem since biomes possess more or less well-defined borders then it is not clear how to quantify such a continuous field as a field of climatic parameters. This is not a problem if keeping in mind that in classic climatology are widely used such concepts as specific isotherms and precipitation isolines. Later on we shall also use the same principle of subdivision for climatic space. Thus, we have two subdivisions S_i ; i=1,...,n and C_j; j=1,...,m of the same „geographic“ plane. Superimposing these subdivisions we can calculate the areas of the intersection of S_i and C_j and corresponding fractions. And finally, we calculate the coefficient h as an information measure of correlation between the vegetation and the pair {annual temperature, annual precipitation} as climatic variables.

Let **S*** be the real Global Vegetation Pattern corresponding to the contemporary climate and **C*** be the subdivision of climate space (we consider the subdivision by curves of the same annual temperatures and precipitation). We assume that there is the ordered set of so-called „virtual GVP“ $\mathbf{S}_k$, k = 1,…,r connected each to other by the relation of local variations, i.e., $\|\mathbf{S}_k - \mathbf{S}_{k+1}\| \sim \varepsilon$. In order words, the difference between two neighbouring patterns is small. A climate scenario is given by the subdivision **C**.

Formally we can calculate the information correlation coefficient h_k for any pair $\{\mathbf{S}_k,\mathbf{C}\}$. If we want to remain in the framework of evolutionary paradigm we have to assume that *the value of h** (information coefficient of correlation between **S*** and **C***) *must be maximal* (in comparison with h_k). This statement is a quantitative form of the hypothesis about the maximal adaptation of the contemporary GVP to the contemporary climatic conditions. For instance, this coefficient for Siberia vegetation is equal to 0.42. This principle gives us the method of selection of the most suitable GVP to the changed climate **C**: *the most suitable GVP,* $\mathbf{S}_s$, *is corresponded to the condition* $h_s : \min_k |h_k - h^*|$.

6. CASE STUDY: THE YENISEI MERIDIAN

We shall illustrate our method applying it to one territory in the Central Siberia (the Yenisei meridian) as a case study area. An idealised structure of biomes is shown in Figure 1.

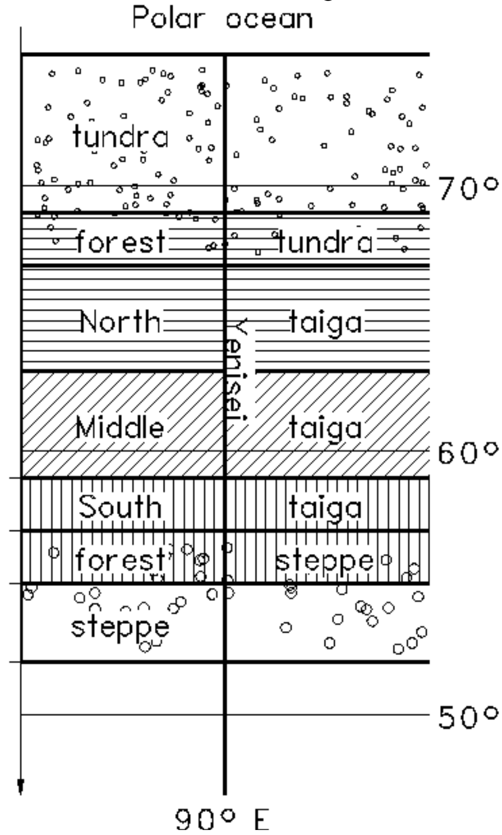

Figure 1. Idealised structure of vegetation along the Yenisei meridian (90° E): the biome of the Southern taiga is located from 59°N to 56° N , the transition zone (forest - steppe) - from 56°N to 54°N and the steppe biome to 51°N.

In accordance with the standard scenario of climate change for CO_2-doubling (Petoukhov et al. [1999]) in this region the annual temperature is expected to rise by 5°C and the precipitation increases by 10% (in average). Almost all the borders between biomes are shifted northwards, except the border between “forest-steppe” and steppe zones, which is shifted by 0.8° southwards. The border between tundra and "forest-tundra" is shifted by 2.6° (1° ~ 111km), between "forest-tundra" and the North taiga - 1.9°; as a result, the transition zone between tundra and taiga increases in 0.7°. 2.0° and 1.5° shifts the north and south borders of the Middle taiga; as a result its width increases in 0.5°. This shift for the South taiga is equal to 1.5° and 1.1°. As a result, the width of transition zone between taiga and steppe ("forest-steppe") increases in 1.9°. The information coefficient of correlation for the

GVP is equal to 0.38, i.e., a degree of climatic determination decreases.

REFERENCES

Bazilevich, N.I., *Biological Productivity of Ecosystems of Northern Eurasia*, Moscow, Nauka, 293 pp., 1993.

Bazilevich, N.I. and L.E. Rodin, Geographical regularities of productivity and turnover of chemical elements for the main types of Earth's vegetation, *General theoretical problems of biological productivity*, Leningrad, Nauka: 24-32, 1969.

Belotelov, N.V., B.G. Bogatyrev, A.P. Kirilenko, S.V.Venevsky, Bioclimatic indices used to estimate vegetation pattern transformation, *STEP International, 3*(3): 13, 1993.

Holdridge, L.R., Determination of world plant formation from simple climate data, *Science, 105*: 367-368, 1947.

Monserud, R.A. and R. Leemans, Comparing global vegetation maps with the Kappa statistic, *Ecological Modelling, 62*: 275-293, 1992.

Petoukhov, V., A. Ganopolsky, V. Brovkin, M. Claussen, A. Eliseev, C. Kubatzki, S. Rahmstorf, CLIMBER–2: a climate system model of intermediate complexity. Part I: model description and performance for present climate, *Climate Dynamics, 15*: 950-966, 1999.

Svirezhev, Yu. (Ed.), *The Brandenburg Biosphere Model: a new version of the Moscow Biosphere Model*, PIK, Potsdam, 835pp., 1997.

Volkenstein, M.V., *Biophysics*, Moscow: Nauka. 592pp., 1988.

Walter, H., *Die Vegetation der Erde in öko-physiologischer Betrachtung. Vol. 1: Die tropischen und subtropischen Zonen*, Jena: Fischer. 538 pp., 1964.

Walter, H., *Die Vegetation der Erde in öko-physiologischer Betrachtung. Vol. 2: Die gemässigten und arktischen Zonen*, Jena: Fischer. 1001 pp., 1968.