SPATIAL STATISTICS OF RANDOM NETWORKS AND A PROBLEM IN RIVER BASIN HYDROLOGY

by

Vijay K. Gupta
Department of Geological Science
& Cooperative Institute for Research in Environmental Science
The University of Colorado
Boulder, CO 80309

and

Ed Waymire
Department of Mathematics
Oregon State University
Corvallis, OR 97331

ABSTRACT

Mathematical problems which arise in river basin hydrology involve the asymptotic analysis, under large source number, of random tree graphs and branching patterns. Some orientation to these problems will be illustrated by a brief survey of some rigorously known results in the case of a well-known empirical bifurcation law due to Robert Horton. Covered are the law of large numbers and central limit theorems which arise in connection with Horton ratios.

1. Introduction

Hydrologists seek to understand and predict the movement of water over land and beneath the surface. This makes the *river basin* a focal point of hydrologic science. River basin anatomy may be viewed at various scales, extending from the microscopic scales of porous media, to the hydraulic scales of flow through a pipe or channel, to the much larger scales of river networks. Our focus here is on network (basin) scale properties as opposed to the single channel hydraulics.

Hydrologic investigations of river basins are carried out with the aid of a number of resources. One way, for example, is through the space/time data base of hydrologic fluxes, e.g., streamflows, rainfall intensity, sediment flows, etc., made available by gauges of various sorts. Although desirable, gauged networks are limited in the US and around the world by the obvious expenses associated with their installation and maintenance. Another important data resource is mapping, e.g., planimetric, topographic, remotely sensed, etc. We shall address here some network properties known empirically through the study of maps.

The hydrologic view of the pattern of streams found in river basins has led to a number of interesting statistical observations which are understood to varying degrees of empiricism and mathematical rigor. In the next couple of sections we shall describe some results in the context of a simple model which were originally obtained from maps of natural river systems. Some comparisons with the results known for real river networks will also be indicated. Apart from scale considerations, there is still a degree of arbitrariness with regard to inclusions or omissions of certain streams in the mapping of networks. This makes asymptotics and robustness problems quite important from the point of view of applications. In place of basin detail, one looks for asymptotic stabilities and laws of averages which exploit the largeness of networks in various ways.

It is noteworthy that certain of the river network codes and statistics described here have also been computed for other naturally occurring branching networks, including lightning patterns, vessels of lung tissue, leaf patterns, root systems, computer storage designs and so on. A few references in these directions are, for example, Berry and Bradley (1976), Borchert and Slade (1981), Flajolet and Prodinger (1986), and Horsfield (1980). As a matter of orientation to neural networks, Hopfield and Tank (1987) liken neural processing to the "motion of a raindrop which lands on a terrain of hills and valleys." Needless to say, the familiar uses of hydraulics to "clarify" the flow of current in an electric network, or that of electric currents to "clarify" the principles governing the flow of water in a pipe or channel network, at least signify an important role for a certain amount of mathematical abstraction and rigor in the study of what are otherwise largely physical problems. The approach illustrated here is to study simple idealizations in an attempt to understand with paper and pencil what may be expected as "typical"behavior of certain network statistics observed in practice. In any case, the problems and results are of a broader interest than hydrology alone.

2. The random model

As we wish to consider probability distributions over a space of rooted trees, it will be convenient to record some basic terminology pertaining to such structures. Broader familiarity with general definitions from graph theory such as *graph, vertex, edge, adjacency, connectivity*, etc., is assumed; see Chartrand and Lesniak (1986), for example.

A *tree graph* is a connected graph without loops. A *binary tree* is a tree graph whose vertices each have degree (valence) one or three. A *rooted* (or *planted*) *tree* refers to a tree graph in which a degree one vertex is selected to be uniquely designated as the *root*. All other degree one vertices are referred to as ***sources***. The degree three vertices are called ***junctions***. The single edge which has the root as an endpoint is called the ***stem***. Edges between two junctions are called ***internal links***, and those between a source and a junction are ***external links***. The stem of a tree graph is regarded as an internal link, except when the graph consists of only a stem joining a single vertex to the root. In the latter case it is defined to be an external link. Finally, the rooted tree may be endowed with a natural edge orientation (***diagraph structure***) consistent with the direction of streamflows when the root represents the network outlet. This makes each vertex, other than the root, ***incident to*** one and only one link (the "downstream" link). There are places where it is convenient to use this identification of vertices and links in the combinatorics. The above terminology is summarized in Figure 2.1.

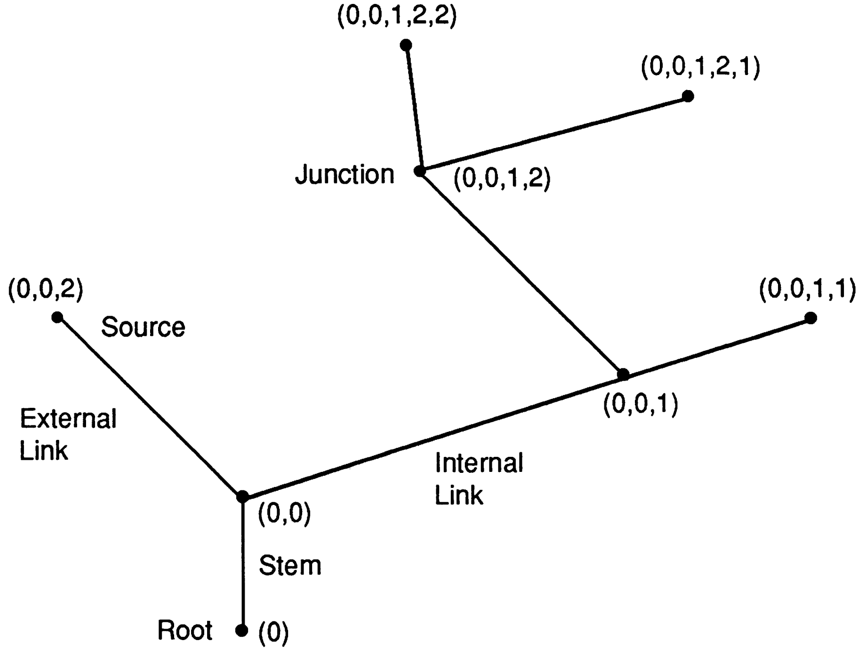

Figure 2.1. Rooted Binary Tree Graph

From here on, the term "tree" will be used to refer to a rooted binary tree diagraph. The number of sources in a tree is a basic parameter called the ***magnitude***. A tree with magnitude $n \geq 1$ has $n - 1$ junctions and a root. The total number of links, including the stem, is $2n - 1$; this also being the total number of vertices (excluding the root).

An equivalent (coordinatized) representation of rooted binary trees may be obtained by a standard coding of the vertices. Namely, consider the *space of words* (sequences) given by

$$W = \bigcup_{k=0}^{\infty} W_k \tag{2.1}$$

where $W_0 = \{(0)\}$, $W_1 = \{(0,0)\}$, $W_k = \{(\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k) : \varepsilon_0 = \varepsilon_1 = 0,\ \varepsilon_j \in \{1,2\},$ $j \geq 2\}$, in general. Then W may be regarded as the set of vertices of the *full infinite rooted binary tree graph*, with root (0), when endowed with the obvious edge connections between vertices $(\alpha) \in W$. A (finite) *rooted binary tree graph of magnitude* n can be defined within this framework as a connected subgraph of W containing the root (0) and having $2n-1$ edges.

Let Ω_n denote the sample space of trees of magnitude n. Then Ω_n consists of (*Cayley's formula*)

$$|\Omega_n| = \frac{1}{2n-1}\binom{2n-1}{n} \tag{2.2}$$

distinct trees. The *finite random model* is defined by the equiprobable (uniform) distribution p_n on Ω_n according to which each tree graph τ in Ω_n occurs with equal probability $p_n(\{\tau\}) = |\Omega_n|^{-1}$. For $0 < p \leq 1/2$, $q = 1-p$, an *infinite* (*grand canonical*) *random model* can be defined as the probability distribution concentrated on the denumerably infinite sample space $\Omega^{(f)}$ of finite trees given by

$$\Omega^{(f)} = \bigcup_{n=1}^{\infty} \Omega_n, \tag{2.3}$$

as

$$Q_p(\{\tau\}) = p^{n-1}q^n \text{ for } \tau \in \Omega_n \text{ for some } n \geq 1. \tag{2.4}$$

The root-stem is certain, and may be terminated with probability q or twice replicated with probability p. For this model the *magnitude* is a random variable M_0 with probability distribution given by

$$Q_p(\{M=n\}) = \frac{1}{2n-1}\binom{2n-1}{n} p^{n-1}q^n, \quad n = 1,2,\ldots. \tag{2.5}$$

In particular, M_0 is Q_p-almost surely finite for $p \leq 1/2$ and

$$Q_p(\{\tau\}|M=n) = \begin{cases} \dfrac{2n-1}{\binom{2n-1}{n}} & \text{if } \tau \in \Omega_n \\ 0 \quad , & \text{otherwise} \end{cases} \tag{2.6}$$

This latter property may be viewed as *sufficiency* of the statistic M_0 (for estimating p) based on the tree sample.

The successive family trees $\tau^{(0)}, \tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(n)}, \ldots$ of the *Bienaymé-Galton-Watson binary branching process* (starting from a single root-stem) may be regarded as randomly generated rooted trees. The offspring distribution has mean $\mu = 2p \leq 1$, making eventual extinction and,therefore, a finite limiting family tree certain to occur. The distribution of the entire (limiting) family tree $\tau^{(\infty)} \in \Omega^{(f)}$ so generated is given by (2.4).

Channel lengths are certainly among the most important quantities one observes in a river network. Investigations of certain basic notions pertaining to river length may be found in the classic papers of Steinhaus (1954) and Einstein (1934) which continue to inspire modern research. In the present framework one may consider the (*graphical or topological*) *distance* (or *height*) from a vertex $v = (\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k) \in W$ to the root $(0) \in W$ is defined by $|v| = k$. While this distance is sufficient for the problems described here, more general notions of distance between vertices and lengths of links are easily introduced with the aid of *weighted graphs*; i.e., positive numbers are (randomly) assigned to the edges according to various possible schemes to represent link lengths. One may then study various statistics such as total channel length, main channel length, etc.; the latter being an extreme value statistic. For networks one seeks to obtain laws of averages for river lengths by exploiting the largeness of the link population. The literature on problems pertaining to main channel length is relatively large. Readers interested in some recent theoretical results on main channel length may consult Gupta, Mesa, and Waymire (1989) and references therein.

3. Horton laws

The *Horton-Strahler ordering* of a channel network is a coding scheme which weights the network bifurcation pattern. Horton's ideas led to the rather striking estimate on the total length of channels in the US to be on the order of 3 million miles [Leopold, 1962]. The ordering of a network is constructed recursively as follows. The external links, equivalently source vertices, are defined to have order one. The vertex and its associated edge *incident from* a pair of first order links has order two. The rule from here on says that a vertex and its associated edge incident from a pair of links of orders m and n, respectively, each has order $\max(m,n)$ if $m \neq n$. The order of the stem is referred to as the *order of the network* (see Figure 3.1). In particular, the network order is the maximal stream order in the network.

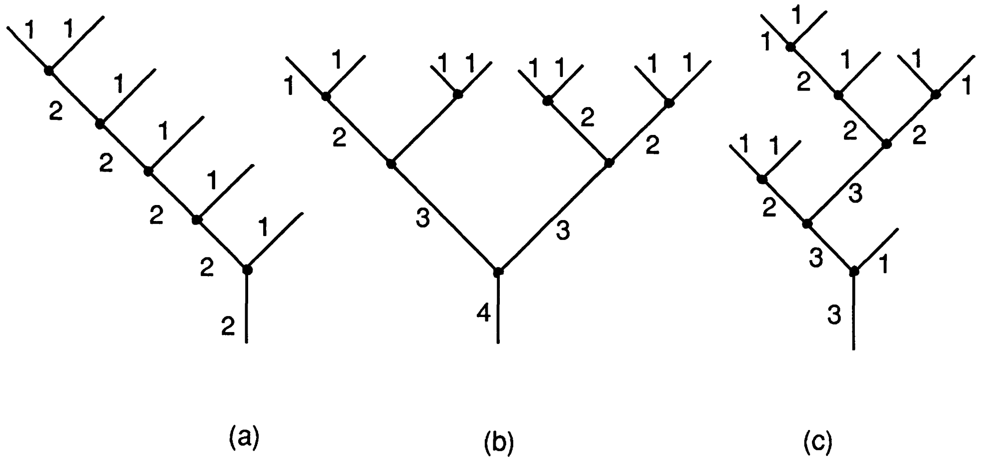

Figure 3.1. Horton Order

Note that if the magnitude $M_0 = n \geq 2$ then the network order H satisfies

$$2 \leq H \leq \log_2 M_0 + 1. \tag{3.1}$$

If $M_0 = n = 1$ then $H = 1$.

A *stream* of order h is a maximal connected directed path of links of equal order h. The stream is said to originate at the vertex (α) associated with the link of the path at the farthest graphical distance from the root. Let $S^{(h)}$ denote the number of streams of order h and let $L^{(h)}$ denote the number of links of order h in a network. For the networks of Figure 3.1 one has $L^{(1)} = S^{(1)} = 5, L^{(2)} = 4, S^{(2)} = 1$ in (a); $L^{(1)} = S^{(1)} = 8$, $L^{(2)} = S^{(2)} = 4$, $L^{(3)} = S^{(3)} = 2$, $L^{(4)} = S^{(4)} = 1$ in (b); $L^{(1)} = S^{(1)} = 8$, $L^{(2)} = 4$, $S^{(2)} = 3$, $L^{(3)} = 3$, $S^{(3)} = 1$ in (c). In any case, $L^{(1)} = S^{(1)} = M_0$ (magnitude) and $S^{(H)} = 1$. The ratios

$$R_L^{(h)} = \frac{L^{(n+1)}}{L^{(h)}} \quad \text{and} \quad R_S^{(h)} = \frac{S^{(h+1)}}{S^{(h)}}$$

are referred to as (Horton) *link and stream number bifurcation ratios*, respectively. Robert Horton (1945) observed early on that such ratios were quite stable for natural basins, with values of $R_S^{(h)}$, for example, near 1/4.

The first rigorous result in this connection was a calculation of the following *ratios of probabilities* by R. Shreve (1967). Let $H_S(\alpha)$ denote the order of a stream which originates at the vertex (α), and let $H_L(\alpha)$ denote the order of (α) (or the link associated with (α)). Note that the random field $H_S(\alpha)$ is *not* defined for all vertices (α); one may arbitrarily define $H_S(\alpha) = 0$ if no stream originates at (α).

Theorem 3.1. *Under the random model distribution (2.4),*

i. $$\frac{Q_{1/2}(H_S(\alpha) = h+1)}{Q_{1/2}(H_S(\alpha) = h)} = \frac{1}{4}$$

ii. $$\frac{Q_{1/2}(H_L(\alpha) = h+1)}{Q_{1/2}(H_L(\alpha) = h)} = \frac{1}{2}.$$

The ratios are obtained by considering the tree recursions

$$\begin{aligned} Q_p(H_L(\alpha) = h) = pQ_p(H_L(\alpha,1) = h)\sum_{j=1}^{h-1} Q_p(H_L(\alpha,2) = j) & \\ + pQ_p(H_L(\alpha,1) = h)\sum_{j=1}^{h-1} Q_p(H_L(\alpha,2) = j) & \\ + pQ_p(H_L(\alpha,1) = h-1)Q_p(H_L(\alpha,2) = h-1), h \geq 2, & \quad (3.2a) \\ Q_p(H_L(\alpha) = 1) = q & \quad (3.2b) \end{aligned}$$

and

$$\begin{aligned} Q_p(H_S(\alpha) = h) = pQ_p(H_L(\alpha,1) = h-1)Q_p(H_L(\alpha,2) = h-1), h \geq 2, & \quad (3.3a) \\ Q_p(H_S(\alpha) = 1) = q, & \quad (3.3b) \end{aligned}$$

under the *invariance property*, $Q_p(H_S(\alpha) = h) = Q_p(H_S(\beta) = h)$ and $Q_p(H_L(\alpha) = h) = Q_p(H_L(\beta) = h)$, $h \geq 1$, for all (α), (β). In particular, one obtains $Q_{1/2}(H_L(\alpha) = h) =$

2^{-h}, $h \geq 1$, and $Q_p(H_S(\alpha) = h) = 2/4^h$, $h \geq 1$. The distribution of $H_S(\alpha)$ is defective since a stream that originates at (α) need not occur; the conditional distribution of $H_S(\alpha)$ given that a stream originates at (α) is of the form $3/4^h$, $h \geq 1$.

For a given value of n, the largest network order possible is $1 + log_2 n$. Precise asymptotics on the expected network order were obtained by Meir,Moon, and Pounder (1980) which show that for fixed n, the expected network order is $\frac{1}{2} log_2 n + O(1)$. This and the above result of Shreve describe the structure of network statistics in terms of "phase averages". From the point of view of network data analysis, one also seeks the behavior of sample values for large networks. In this connection, Gupta and Waymire (1983) provide a law of large numbers for rigorous interpretation of Shreve's probability ratios (ie.,Thm 3.1 above) in the form of a statistical law of stream numbers as observed by Horton. The precise result is as follows.

Theorem 3.2. *(Law of Large Numbers). Let $S_n^{(2)}, S_n^{(1)} = n$, $L_n^{(2)}, L_n^{(1)} = n$ denote the numbers of second and first order streams and links, respectively, in the random model P_n of magnitude n. Then*

i. $\dfrac{S_n^{(2)}}{S_n^{(1)}} \longrightarrow \dfrac{1}{4}$ *in probability as* $n \longrightarrow \infty$

ii. $\dfrac{L_n^{(2)}}{L_n^{(1)}} \longrightarrow \dfrac{1}{2}$ *in probability as* $n \longrightarrow \infty$.

The proofs are simple applications of Chebyshev's inequality based on first and second moment calculations. In the case of stream numbers Shreve (1966) used simple combinatorics which provide the exact form of the distribution of $S_n^{(2)}$ according to $(n \geq 2)$,

$$P_n(S_n^{(2)} = j) = Z_n^{-1} \binom{n-2}{n-2j} \frac{2^{n-2j}}{2j-1} \binom{2j-1}{j}, \quad j = 1, 2, \ldots, [n/2], \tag{3.4}$$

where

$$Z_n = \sum_{j=1}^{[n/2]} \binom{n-2}{n-2j} \frac{2^{n-2j}}{2j-1} \binom{2j-1}{j}$$

is the normalization constant. The first two moments are, as first computed by Werner (1972),

$$ES_n^{(2)} = \frac{n}{4}\left(1 + \frac{1}{2n-3}\right) \sim \frac{n}{4} \text{ as } n \longrightarrow \infty \tag{3.5a}$$

$$\text{Var } S_n^{(2)} = \frac{n(n-1)(n-2)(n-3)}{2(2n-3)^2(2n-5)} \sim \frac{n}{16} \text{ as } n \longrightarrow \infty. \tag{3.5b}$$

So part (i) follows directly from this and Chebyshev's inequality. The companion calculations for part (ii) were made by Mesa (1986). Self-contained derivations of these results can also be found in Wang and Waymire (1989).

The fluctuation laws for the bifurcation ratios were recently worked out by Wang and Waymire (1989) in the case of stream order ratios. The precise result is as follows.

Theorem 3.3. *(Central limit theorem) Let $S_n^{(2)}, S_n^{(1)}, L_n^{(2)}, L_n^{(1)}$ be as in Theorem 3.2. Then, denoting convergence in distribution by $\Longrightarrow$,*

i. $$\sqrt{n}\left(\frac{S_n^{(2)}}{S_n^{(1)}} - \frac{1}{4}\right) \Longrightarrow N(0,4) \text{ as } n \to \infty$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

The proof is based on methods from large deviation theory. In fact, we obtain that

$$\lim_{n\to\infty} n^{-1} E e^{\xi S_{2,n}} = -\frac{\xi}{2} + log(\frac{e^{-\frac{\xi}{2}}+1}{2}), \tag{3.6}$$

in a suitably small neighborhood of zero.

In the case of link number ratios companion calculations suggest the following conjecture:

$$\sqrt{n}\left(\frac{L_n^{(2)}}{L_n^{(1)}} - \frac{1}{2}\right) \Longrightarrow N(0,2) \text{ as } n \to \infty. \tag{3.7}$$

However we have not been able to completely resolve this problem. Another somewhat related class of problems is treated by Flajolet and Odlyzko (1984) with methods which may be relevant here, or vice-versa.

In the course of trying to identify the slowly varying part of a Tauberian limit for a moment sequence in the proof of Theorem 3.3, the following curious identity was discovered

$$\frac{1}{4\sqrt{\pi}} \sum_{j=1}^{n} \binom{n+1}{j} \Gamma\left(\frac{2j-1}{2}\right) \Gamma\left(\frac{2n+1-2j}{2}\right) = \Gamma\left(\frac{2n-1}{2}\right), n = 1, 2, \ldots, \tag{3.8}$$

where $\Gamma(x)$ is the gamma function. Among its many innocent-looking equivalents is the identity

$$\sum_{j=1}^{n} \frac{\binom{n+1}{j}\binom{n-1}{j-1}}{\binom{2n-2}{2j-2}} = 4n-2, n = 1, 2 \ldots,. \tag{3.9}$$

It may be worthwhile remarking that this is precisely the sort of identity for which the symbolic software package *Macsyma* is very quick to furnish a hi-tech "proof". Although we could not come up with a "ball and urn"explanation, we did eventually find a more conventional induction proof and we have since learned a bit about the combinatorial content of (3.9) from Otto G. Ruehr. In fact he observed that it may also be obtained as a special case of Gauss's theorem for ${}_2F_1$ hypergeometric functions; see Wang and Waymire (1989) for a sketch.

Acknowledgements

This research was partially supported by grant 27772-GS from the Army Research Office and NSF grant DMS-8801466.

References

Berry, M.,& Bradley, P. M. (1976). The application of network analysis to the study of branching patterns of large dendritic fields. *Brain Research* **109** 111-132.

Borchert, R. & Slade, N. A. (1981). Bifurcation ratios and adaptive geometry of trees. *Bot. Gaz.* **142** 394-401.

Chartrand, G., & Lesniak, L. (1986). *Graphs and DiGraphs.* Wadsworth, 2nd ed., Monterey, CA.

Einstein, A. (1934). The cause of the formation of meanders in the courses of rivers and of the so-called Bear's law, in *Essays in Science.* Wisdom Library, NY. (First published in *Die Naturwissenschaften* **14** 1926).

Flajolet, P. & Odlyzko, A. M. (1984). Limit distributions for coefficients of iterates of polynomials with applications to combinatorial enumerations. *Math. Proc. Camb. Phil. Soc.* **96** 237-284.

Flajolet, P. & Prodinger, H. (1986). Register allocation for unary-binary trees. *SIAM J. Comput.* **96** 629-640.

Gupta, V. K., Mesa, O., & Waymire, E. (1989). Tree dependent extreme values: The exponential case. *Jour.Appld.Prob.* (in press)

Gupta, V. K., & Waymire, E. (1983). On the formulation of an analytical approach to hydrologic response and similarity at the basin scale. *J. Hydrol.* **65** 95-124.

Hopfield, J., & Tank, D. (1987). Collective computation in neuronlike cells. *Scientific American* (December) 104–114.

Horsfield, K. (1980). Are diameter, length, and branching ratios meaningful in the lung? *Jour. Theor. Biol.* **87** 773-784.

Horton, R. (1945). Erosional development of streams and their drainage basins: Hydrophysical approach to quantitative morphology. *Bul. Geol. Soc. Amer.* **56** 275–370.

Leopold, L. B. (1962). Rivers. *American Scientist* **50** (4) 511-537.

Meir, J. W., Moon, J. W.,& Pounder, J. R. (1980). On the order of random channel networks. *SIAM J. Alg. Disc. Meth.* **1** 25-33.

Mesa, O. (1986). Analysis of Channel Networks Parametrized by Elevation. Ph.D. Dissertation, Dept. of Civil Engineering, University of Mississippi.

Shreve, R. L. (1967). Infinite Topologically random channel networks. *J. Geol.* **75** 179–186.

Steinhaus, H. (1954). Length, Shape, and Area. *Comun. of Colloquium Mathematicum* **3** 1–13.

Wang, Xi, & Waymire, E. (1989). Central Limit Theorems for Horton Ratios. *preprint.*

Werner, C. (1972). Two models for Horton's law of stream numbers. *Can. Geogr.* **16** (1) 50–68.