geographical analysis

Geographical Analysis ISSN 0016-7363

Space, Scale, and Scaling in Entropy Maximizing

Michael Batty

Centre for Advanced Spatial Analysis (CASA), University College London (UCL), London, U.K.

Entropy measures were first introduced into geographical analysis during a period when the concept of human systems in equilibrium was in its ascendancy. In particular, entropy maximizing, in direct analogy with equilibrium statistical mechanics, provides a powerful framework in which to generate location and interaction models. This was introduced and popularized by Wilson, and it led to many different extensions that elaborate the framework rather than extend it to different kinds of models. I review two such extensions here: how space can be introduced into the formulation through defining a ''spatial entropy'' and how entropy can be decomposed and nested to capture spatial variation at different scales. Two obvious directions to this research remain implicit. First, the more substantive interpretations of the concept of entropy for different shapes and sizes of geographical systems have hardly been developed. Second, an explicit dynamics associated with generating probability distributions has not been attempted until quite recently with respect to the search for how power laws emerge as signatures of universality in complex systems. In short, the connections between entropy maximizing, substantive interpretations of entropy measures, and the longer-term dynamics of how equilibrium distributions are reached and maintained have not been well developed. This literature gap has many implications for future research, and, in conclusion, I sketch the need for new and different entropy measures that enable us to see how equilibrium spatial distributions can be generated as the outcomes of dynamic processes that converge to a steady state.

Defining and interpreting entropy

An event occurring with probability p gives us a measure of information about the likelihood of that probability being correct. Any event with a very low probability that occurs gives us a great deal of information, whereas when an event with a high probability occurs, this is less of a surprise and gives us correspondingly less information. Information thus varies inversely with probability, and we can define

Correspondence: Michael Batty, Centre for Advanced Spatial Analysis (CASA), University College London (UCL), 1-19 Torrington Place, London WC1E 6BT, U.K.
e-mail: m.batty@ucl.ac.uk

Submitted: March 1, 2010. Revised version accepted: May 4, 2010.

this as $1/p$. However, if we have two independent events with probabilities p_1 and p_2, if one occurs and then the other occurs, we would expect the information gained to be $1/(p_1p_2)$ because the probability of their joint occurrence is p_1p_2. Yet when an event occurs, it is reasonable to suppose that the information gained should be additional to any information already gained, and, thus, one might expect the information for both events to be the sum of each. Clearly, this is not $1/p_1 + 1/p_2 \neq 1/(p_1p_2)$ but a function $F(\circ)$, of which the only solution is the qlogarithm of the inverse of the probability, that is,

$$\left.\begin{aligned} F\left(\frac{1}{p_1p_2}\right) &= F\left(\frac{1}{p_1}\right) + F\left(\frac{1}{p_2}\right) \\ -\log(p_1p_2) &= -\log(p_1) - \log(p_2) \end{aligned}\right\} \tag{1}$$

In short, the information gained by the occurrence of any event is $\log(1/p) = -\log(p)$, which also can be thought of as a measure of the uncertainty of the event occurring or as a measure of surprise (Tribus 1969).

For a series of n events, with probabilities p_i, $i = 1, 2, \ldots, n$, the average information is the expected value of this series, which can be written as

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

This measure was first defined in this form by Shannon (1948) when considering the communication of information over a noisy channel. But the formula is central to statistical physics, originating with Clausius in the early 19th century, and given specific statistical interpretation by Boltzmann and then by Gibbs as the measure for thermodynamic entropy. In particular, the method of entropy maximizing, which is a major theme here, was first associated with finding the distribution of particles in a physical context, giving rise to the Boltzmann–Gibbs distribution that serves as the baseline for many of the distributions of spatial activity introduced here (Ben-Naim 2008). When Shannon (1948) introduced this measure, he sought advice as to what to call it from John von Neumann, who had worked with a version of the measure in quantum physics. Although apocryphal, von Neumann[1] reportedly said, ''You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage!''

This function has many attractive properties for describing spatial distributions. Here, we initially assume that the probability p_i is proportional to some count or density of spatial activity, such as population in a zone i that might be a census tract. If all the population were located in a ''mile-high building'' such as the one proposed for a town of 100,000 people in 1956 by Frank Lloyd Wright (Rybczynski 2010), then $p_i = 1$ and $p_k = 0, \forall k \neq i$, and the entropy would be at a minimum, with $H_{\min} = 0$. If the population were evenly spread throughout the tracts as

$p_i = 1/n$, $\forall i$, then the entropy would be at a maximum, with $H_{\max} = \log n$. Many distributions lie between these extremes, and the construction of a variety of related measures that make comparisons with the maximum is possible. For example, a measure of information difference can be constructed as

$$\begin{aligned} I =& H_{\max} - H = \log n + \sum_i p_i \log p_i \\ =& \sum_i p_i \log\left[\frac{p_i}{1/n}\right] = \sum_i p_i \log\left[\frac{p_i}{q_i}\right] \end{aligned} \tag{3}$$

The term on the right-hand side (RHS) of the second line of equation (3) is an information difference of the kind widely used in likelihood theory, first popularized by Kullback (1959). Snickars and Weibull (1977) and Webber (1979) discussed it in a geographical context where $\{q_i\}$ can be interpreted as a prior and $\{p_i\}$ as a posterior probability distribution. The normalization of I as $R = I/H_{\max}$ is called relative redundancy, which is a measure varying between 0 and 1.

The entropy measure in equation (2) increases with the number of events or objects making up a distribution. This is intuitively acceptable because as we have more events, we have more information, unless the additional events have zero probability of occurrence. This feature is easy to show because $H_{\max} = \log n$; but it also constitutes a problem for spatial analysis because it means that we cannot compare systems with unequal numbers of objects, or, in our case, different numbers of spatial subdivisions or zones. We have to normalize the quantity in some way, such as in equation (3), and the development of spatial entropy that I present subsequently is one strategy for doing this. This lack of comparability means that methods for deriving spatial probability distributions have been much more at the fore in geographical analysis than more substantive interpretations of the entropy measure. This focus is unfortunate because some important conclusions need to be drawn about the structure of different spatial systems with respect to measures of entropy. This is an unfinished quest.

If we consider a hypothetical system in which all the population is piled into one zone—the mile-high building example—then such a system is completely ordered; it has minimum entropy, there is no uncertainty about its structure, and it has no variety. To make this kind of system possible, we would need enormous constraints on its manufacture to the point where everything would have to be controlled. In contrast, systems in which the population is spread out evenly have maximum entropy and maximum disorder and constitute the situation that would emerge when the system has *no constraints* on the system and all persons can live where they want. Given enough time, people would spread out evenly in the absence of any reason for locating in any particular place. What is significant about this interpretation is its direct connections to thermodynamic entropy, where maximum disorder occurs when all particles mix freely, which occurs when temperature in a system rises and any differences are ironed out. This order–disorder continuum with respect to H is directly invoked if we consider that as we put more and more constraints on the form of a distribution we successively reduce the entropy. In this sense, a direct link exists between the probability distributions that we observe and the model and the methods of deriving such distributions using the method of entropy maximization, to which we now turn. I first present the method, which relates directly to that pioneered by Wilson (1970) for urban and regional systems, although after this presentation, I describe many new insights that seek to show how such methods can be extended to deal with space, scale, and scaling.

The entropy-maximizing framework

The best strategy to choose a probability distribution consistent with information we know the distribution must meet is to maximize its entropy subject to a series of constraints that encode the relevant information. When entropy is maximized, the distribution is the most conservative and hence the most ''uninformative'' we can choose. Were we to choose a distribution with lower entropy, we would be assuming information that we did not have, while a distribution with higher entropy would violate the known constraints. Thus, this maximization is equivalent to choosing a distribution that is the most likely or probable within the constraints, because it is easy to show, as Wilson (1970 and in this issue) does, that the maximum entropy is an approximation to the probability of a particular macrostate occurring among all possible arrangements (or microstates) of the events in question.

Unlike Wilson (1970 and in this issue), I demonstrate the maximization for a probability distribution of the location i of population p_i in n zones, rather than the probability p_{ij} of interactions between zones i and j, although all my derivations are immediately generalizable to these more detailed specifications. We must first specify the constraints, which we take to be functions of the probabilities that define totals, averages, or more generically ''moments'' of a distribution. To demonstrate this derivation, I choose two constraints for the location of population. First, a normalization constraint ensures the probabilities sum to unity:

$$\sum_i p_i = 1 \tag{4}$$

Second, I choose a constraint for the average cost, $\bar{C}$, of locating in any zone, which is the sum of the individual locational costs c_i weighted by their probabilities of occurrence:

$$\sum_i p_i c_i = \bar{C} \tag{5}$$

Next I form a Lagrangian L that consists of the entropy H reduced by the information encoded into the constraints in equations (4) and (5), and then find its maximum with respect to the probability p_i. In other words,

$$L = -\sum_i p_i \log p_i - (\lambda_0 - 1)\left(\sum_i p_i - 1\right) - \lambda_1\left(\sum_i p_i c_i - \bar{C}\right) \tag{6}$$

where the parameters $\lambda_0 - 1$and λ_1 are Lagrangian multipliers that ensure the maximization meets these constraints. Differentiating equation (6) with respect to each probability p_i and setting the result equal to zero yields

$$\frac{\partial L}{\partial p_i} = -\log p_i - \lambda_0 - \lambda_1 c_i = 0 \tag{7}$$

Rearrangement and exponentiation of equation (7) gives the probability model

$$p_i = \exp(-\lambda_0 - \lambda_1 c_i) \tag{8}$$

Note that the multiplier specified as $(\lambda_0 - 1)$ enables us to get rid of the free-floating negative number –1 resulting from the differentiation in equations (6) and (7), thus clarifying the ensuing algebra.

The model in equation (8) has some intriguing and appealing properties. The values of the parameters λ_0 and λ_1 can be determined by solving the model according to the constraint equations (4) and (5). If we substitute equation (8) into (4), then $\exp(-\lambda_0)$ becomes a partition function defined from

$$\left.\begin{aligned} \exp(\lambda_0) &= \sum_i \exp(-\lambda_1 c_i), \text{ or} \\ \lambda_0 &= \log\left[\sum_i \exp(-\lambda_1 c_i)\right] \end{aligned}\right\} \tag{9}$$

The exponential model in equation (8) can then be more clearly written as

$$p_i = \frac{\exp(-\lambda_1 c_i)}{\sum_i \exp(-\lambda_1 c_i)}, \sum_i p_i = 1 \tag{10}$$

and from this we see that if the Lagrangian multiplier for the average cost of location is redundant—that is, $\lambda_1 = 0$—then the exponential model collapses to a uniform distribution where $p_i = 1/n$.The last step of the derivation is to substitute the model into the entropy equation H; the entropy for this model is at its maximum when

$$H_{\max} = -\sum_i p_i \log[\exp(-\lambda_o - \lambda_1 c_i)] = \lambda_0 + \lambda_1 \bar{C} \tag{11}$$

This maximum is a function of each multiplier and its constraint, with the implication that entropy is a function of the spread of the distribution, which is determined by the cost constraint. In this sense, entropy can be seen as a system-wide accessibility function in that the partition and cost relate to the spread of probabilities across the system.

The exact form of the relationship in equation (11) requires a little more insight into the form of its exponential function. To this end, we need to anticipate the next section in moving from a discrete to a continuous form of model. For the exponential function, the summations in equations (4)–(6) and (9)–(11) can be generalized to continuous form by assuming that $p_i = p(x_i)\Delta x_i$ and $c_i = c(x_i)$, where $p(x_i)$ is an approximation of the size of the population x at the point location i to the probability density over the interval or area defined by Δx_i, and $c(x_i)$ is an equivalent approximation to the cost density in zone i. We can assume that, as $\Delta x_i \rightarrow 0$, $p(x_i) \rightarrow p(x)$ and $c(x_i) \rightarrow c(x)$. Thus, we can write and simplify constraint equation (4) as

$$\sum_{\substack{i \\ \lim \Delta x_i \rightarrow 0}} p(x_i)\Delta x_i = \int_0^\infty p(x)dx = \int_0^\infty \exp(-\lambda_o)\exp[-\lambda_1 c(x)]\,dx$$
$$= \frac{\exp(-\lambda_o)}{\lambda_1} = 1 \tag{12}$$

which further simplifies to

$$\exp(-\lambda_0) = \lambda_1 \text{ and } \lambda_0 = -\log\lambda_1 \tag{13}$$

Now the constraint on travel cost in continuous form can be written as

$$\begin{aligned}\sum_{\substack{i \\ \lim \Delta x_i \rightarrow 0}} p(x_i)c(x_i)\Delta x_i &= \int_0^\infty p(x)c(x)dx \\ &= \int_0^\infty \lambda_1 \exp[-\lambda_1 c(x)]c(x)dx = \frac{1}{\lambda_1} = \bar{C}\end{aligned} \tag{14}$$

From the derivations in equations (13) and (14), the exponential model can be stated in a much simpler form, equivalent to the Boltzmann–Gibbs distribution in statistical mechanics. Noting now that $\exp(-\lambda_0) = \lambda_1 = 1/\bar{C}$, the model can be written in its classic form as a density:

$$p(x) = \frac{1}{\bar{C}}\exp\left(-\frac{c(x)}{\bar{C}}\right) \tag{15}$$

where in thermodynamics $c(x)$ is the energy at location x and $\bar{C}$ is related to the average temperature T because $\bar{C} = kT$, where k is Boltzmann's constant. Note that, as I am using Shannon's rather than Boltzmann's entropy, the expression for average cost is dimensionless when $k = 1$, but this does not make any significant difference to the interpretation (Ben-Naim 2008).

The maximum entropy in continuous form is not the limit of equation (2) with respect to Δx_i as I show here. Before I do this demonstration, let me state this entropy as

$$S = -\int_0^\infty p(x)\log p(x)dx \tag{16}$$

Then, substituting equation (15) into (16), the continuous entropy at its maximum has the same form as equation (11), which simplifies to

$$\begin{aligned} S &= -\int_0^\infty p(x)\log p(x)dx = \lambda_0 + \lambda_1\bar{C} \\ &= -\int_0^\infty p(x)\log\left[\frac{1}{\bar{C}}\exp\left(\frac{c(x)}{\bar{C}}\right)\right]dx = \log\bar{C} + 1 = -\log\lambda_1 + 1 \end{aligned} \tag{17}$$

Therefore, the appropriate measurements of entropy S (and H) vary with the log of the average cost or temperature, and the parameters λ_0 and λ_1 can be approximated from this average cost. In the sense that average cost in the system might be interpreted as a kind of accessibility, entropy itself can be seen as such a measure. Batty (1983), Erlander and Stewart (1990), and Roy and Thill (2004) explore related insights.

Our last foray into the derivation of this model—which I regard as a baseline for geographical systems that must meet some conservation constraint such as average cost—involves sketching how such exponential distributions can emerge from a simple dynamics that involves changes to the costs of location between different places i. Let us assume that a system starts with each place i having the average cost of location as $\bar{C}$, that is, $c_i = \bar{C}$, $\forall i$. Also assume that each place has some sort of collective consciousness or "agent" that is willing to increase or decrease the cost of location if instructed to do so. I design a simulation where, at each time, two places i and j are chosen at random and a small fixed fraction of the cost of location, Δc, is transferred such that the total (and average) cost of location remains the same. Each time, $c_i(t+1) = c_i(t) + \Delta c$ and $c_j(t+1) = c_j(t) - \Delta c$ such that $\Sigma c_k(t+1) = \Sigma c_k(t)$. Let us also assume that a location cannot receive a negative cost, that a lower bound exists for $c_i(t) \geq 0, \forall i, t$, where this boundary condition is absolutely essential for the generation of the stable state that ultimately emerges. If this process continues for many time steps, a distribution of costs (in locations) emerges that follows the Boltzmann–Gibbs distribution in equations (10) or (15) that appears when the costs are binned and the relative probability distribution examined.

In short, through a process of random swapping akin to energy collisions in a thermodynamic system, the system self-organizes to the exponential distribution from any starting point, which in our case is the uniform distribution. This process is robust in that many variations of the swapping mechanism involving randomness lead to the universal form of a negative exponential that is due to the boundary condition and the conservation of costs. Strictly speaking, this process is best considered as one where each location is an individual engaging in the process with the resulting probability distribution formed by collecting each of these individuals into "locations." Drãgulescu and Yakovenko (2000) show many variants of the model that lead to the same ultimate form with respect to a simple economic system where individuals engage in swaps involving a conserved quantity such as money. They also generalize the model by relaxing the boundary constraints and embed it in a wider context where wealth that is not conserved is considered, making the point that these variants also admit the generation of other distributions such as the log normal and the power law. This kind of model has not been explored in geographical analysis hitherto for there has been no consideration of the dynamics that lead to entropy maximizing. The dynamics that have been explored is one in which the entropy-maximizing solution is embedded in a wider nonlinear dynamics (Wilson in this issue). This discussion introduces the possibility of disaggregating the entropy-maximizing model to the point where individuals or agents are the basic objects constituting a system, thus opening the framework to much more general types and styles of simulation such as agent-based modeling.

Spatial entropy: the continuous formulation

So far, apart from my brief digression in the preceding section into continuous entropy, I make no formal distinction between density and distribution. I assume implicitly that distribution and density covary, which would be the case where each interval $\Delta x_i = \Delta x$, $\forall i$, that is, each interval is the same size as, for example, in a spatial system arranged on a regular grid. Many spatial models ignore the size of the interval completely, and operational models that build on entropy maximizing rarely factor internal size into their simulations, which inevitably leads to biased applications. Yet I can easily show how interval size must enter an analysis explicitly. As before, I first define each element of the probability distribution p_i that is the product of an approximation to the density $p(x_i)$ of population size x at location i and the interval size Δx_i;

$$p_i = p(x_i)\Delta x_i \tag{18}$$

from which density is defined as

$$p(x_i) = \frac{p_i}{\Delta x_i} \tag{19}$$

Using equation (18) in the entropy H, equation (2) can be rewritten as

$$\begin{aligned} H &= -\sum_i p(x_i)\Delta x_i \log[p(x_i)\Delta x_i] \\ &= -\sum_i p(x_i)\log[p(x_i)]\Delta x_i - \sum_i p(x_i)[\log \Delta x_i]\Delta x_i \end{aligned} \tag{20}$$

When we pass to the limit, $\lim \Delta x_i \to 0$, equation (20) can be written as

$$\lim\{\Delta x_i \to 0\}\, H = -\int_0^\infty p(x)\log p(x)dx - \int_0^\infty p(x)\log dx \tag{21}$$

where the first term on the RHS of equation (21) is the continuous Shannon entropy defined as S in equation (17). Equation (21) implies that $H \to \infty$, as $\lim \Delta x_i \to 0$, which is another way of saying what I have already said in the previous section, namely, if $\Delta x_i = X/n$, $\forall i$, then $H \sim \log n$, and this goes to infinity in an equivalent way.

The key to augmenting the entropy-maximizing method is to use a discrete approximation to the continuous entropy S. Using equation (19) in the approximation to S, which is the first term on the RHS of the second line of equation (20), gives

$$H_S = -\sum_i p_i \log\left[\frac{p_i}{\Delta x_i}\right] \tag{22}$$

which I define as spatial entropy (Batty 1974; Goldman 1968). Using equation (22) instead of equation (2) in the entropy-maximizing scheme, which involves minimizing the Lagrangian in equation (6) with $p_i/\Delta x_i$ for p_i in equation (7), leads to the augmented Boltzmann–Gibbs exponential model, the equivalent of equation (10):

$$p_i = \frac{\Delta x_i \exp(-\lambda c_i)}{\sum_i \Delta x_i \exp(-\lambda c_i)} \tag{23}$$

Equation (23) can be interpreted as a model in which the interval size has been introduced as a weight on the probability and is consistent with the continuous version of the Boltzmann–Gibbs model when passing to the limit $\Delta x_i \rightarrow 0$.

However, another interpretation exists for this augmented model. If we write the entropy H_S in the expanded form of equation (22) as

$$\begin{aligned} H_S &= -\sum_i p_i \log p_i + \sum_i p_i \log \Delta x_i \\ &= H + \sum_i p_i \log \Delta x_i \end{aligned} \tag{24}$$

then we can consider the second term on the RHS of equation (24)—the expected value of the logarithm of the interval sizes—as a constraint on the discrete entropy H. This is a very specific constraint in equation (24) in that it is simply a direct augmentation to the discrete entropy. Instead, we set this as a freely varying constraint on the discrete entropy in the form

$$\sum_i p_i \log \Delta x_i = \log \bar{X} \tag{25}$$

and introduce this into the Lagrangian in equation (6), which we now write as

$$\begin{aligned} L = &-\sum_i p_i \log p_i - (\lambda_0 - 1)\left(\sum_i p_i - 1\right) - \lambda_1\left(\sum_i p_i c_i - \bar{C}\right) \\ &- \lambda_2\left(\sum p_i \log \Delta x_i - \log \bar{X}\right) \end{aligned} \tag{26}$$

The model that we derive from this minimization can be written as

$$p_i = \exp(-\lambda_0 - \lambda_1 c_i - \lambda_2 \log \Delta x_i) \tag{27}$$

which in a more familiar form can be written as

$$p_i = \frac{(\Delta x_i)^{-\lambda_2} \exp(-\lambda c_i)}{\sum_i (\Delta x_i)^{-\lambda_2} \exp(-\lambda c_i)} \tag{28}$$

Thus, the interval or zone size enters the model as a scaling factor, a kind of benefit rather than cost, in the same way such factors are introduced by Wilson (1970) in his family of spatial interaction models. By comparing equations (23) and (28), if the multiplier λ_2 is forced to be unity, then the constraint on interval size enters the model in exactly the same way it would if it were incorporated into the entropy in the first place, that is, as a maximization of spatial rather than discrete entropy. Note also that, in entropy-maximized equations like (28), the sign of the multipliers is undetermined until they are fitted to meet the constraint equations. One further point about this augmented maximization is that if constraint equations in the Lagrangian or augmentations to the entropy are of logarithmic form the relevant variables enter a model as power laws: they are scaling, and any continuous version of the derivation has to be modified to ensure that these constraints lie within defined limits. I return to this point subsequently when dealing more formally with scaling.

The standard example that Wilson (1970) uses to demonstrate the logic of entropy maximization is for trip distribution or spatial interaction where the entropy is based on the probability p_{ij} that a person makes a trip T_{ij} from an origin zone i such as a workplace to a destination zone j such as a residence. An example of the unconstrained model that is subject to an equivalent cost and normalization constraint is derived by maximizing

$$H = -\sum_i \sum_j p_{ij} \log p_{ij} \tag{29}$$

subject to the following constraints:

$$\sum_i \sum_j p_{ij} = 1 \text{ and } \sum_i \sum_j p_{ij} c_{ij} = \bar{C} \tag{30}$$

where c_{ij} is the cost of interaction between zones i and j, and the model is derived as

$$p_{ij} = \frac{\exp(-\lambda_1 c_{ij})}{\sum_i \sum_j \exp(-\lambda_1 c_{ij})} \tag{31}$$

The density equivalent is based on normalizing the probability with respect to the size of the zones at each origin and destination Δx_i and Δx_j, respectively. Following through the same logic used to derive equation (23) for the one-dimensional case and using the appropriate spatial entropy with respect to $p_{ij}/(\Delta x_i \Delta x_j)$, we generate the equivalent interaction model as

$$p_{ij} = \frac{\Delta x_i \Delta x_j \exp(-\lambda_1 c_{ij})}{\sum_i \sum_j \Delta x_i \Delta x_j \exp(-\lambda_1 c_{ij})} \tag{32}$$

Note that all the same conclusions about the measure of entropy and the way the model can be simplified, as developed for the location model, follow for the interaction model in equation (32). If $\Delta x_i \Delta x_j = \Delta x \Delta x$, the model collapses to the distributional form in equation (31), while if $\lambda_1 = 0$, the model collapses to the uniform distribution, weighted according to the interval size for the distributional form. The way in which attractors or benefits can be introduced either as augmented measures to the entropy or as constraints also follows, and in this sense; equations (23) and (32) are generic forms.

Before moving on to questions of scale and aggregation, I reiterate my earlier definition of information differences with respect to entropy. Statistical information is defined as the difference between two distributions $\{p_i\}$ and $\{q_i\}$, where $\{q_i\}$ often is referred to as the prior and $\{p_i\}$, the posterior. Kullback (1959) and, in a geographical context, Snickars and Weibull (1977) and Webber (1979), among others, define information I as

$$I = \sum p_i \log\left[\frac{p_i}{q_i}\right], \sum_i p_i = \sum_i q_i = 1 \tag{33}$$

I varies between zero and infinity, zero being the measure when $p_i = q_i, \forall i$, that is, no difference exists between prior and posterior distributions; in short, no information is gained by moving from the prior to the posterior. If we assume that the prior probability distribution is proportional to the interval size—that is,

$$q_i = \frac{\Delta x_i}{\sum_i \Delta x_i} = \frac{\Delta x_i}{X} \tag{34}$$

where X is the area of the entire system—then the information in equation (33) becomes

$$\begin{aligned} I &= \sum p_i \log\left[\frac{p_i}{\Delta x_i / X}\right] = \log X + \sum p_i \log\left[\frac{p_i}{\Delta x_i}\right] \\ &= \log X - H_S \end{aligned} \tag{35}$$

When $\Delta x_i = \Delta x, \forall i$, equation (35) collapses to equation (3), which is repeated here as

$$I = H_{\max} - H = \log n + \sum_i p_i \log p_i \tag{36}$$

Many such manipulations of entropy and information exist that all give oblique insights into the measure and the shape of the relevant distributions, some of which recur in the subsequent discussion.

To conclude this section, one noteworthy concern is how we might proceed to develop substantive interpretations of the various entropy measures as derived so far in this article, which does not broach any empirical applications. Nevertheless, although these measures have rarely been used other than for derivation of model structures using entropy maximizing, obvious and straightforward applications exist in which their actual values lead to interesting and informative insights into the structure of spatial systems. The thermodynamic relations in which entropy is the difference between free-energy and fixed-energy use, where free energy also can be thought of as the difference between fixed energy and entropy, can generate many substantive interpretations of the extent to which spatial structures are constrained by known energy use. In terms of models that are derived using entropy maximization, their parameters and constraints can be interpreted as a function of their energies (Morphet 2010). Although entropy can be interpreted as a measure of spread or dispersion in a spatial system that ties it quite strongly to its thermodynamic interpretation, its real value is in illustrating differences between spatial systems, particularly where the energy constraint in a given system changes over time. These energy and entropy differences are what are important, because they are tied quite strongly to measures of difference between accessibility and utility, and to consumer surplus in transport evaluation (Batty 2010). It is not possible to develop these ideas further here, but suffice it to say an entirely new research agenda can be formulated with respect to the substantive meaning of entropy and related energy measures that tie these quantities back to their more fundamental thermodynamic origins.

Consequently, far from being of mainly historical interest in spatial analysis, entropy maximizing still has enormous potential for generating new insights into the structure and functioning of spatial systems, which I illustrate by deriving models that pertain to scaling that are also central to new developments in complexity theory (Batty 2009).

Scale and entropy: aggregation and constraints

Shannon's entropy in equation (2) has an exceptionally easy-to-manipulate log-linear structure and additive form that allows it to be aggregated with respect to groups of objects that might pertain to some higher level of organization in the system of interest. Theil (1972) refers to this process of aggregation as the entropy-decomposition theorem and, to illustrate it, I first divide the set Z of n objects, in this case the spatial zones of the geographical system, into K sets, Z_k, $k = 1, 2, ..., K$, each with n_k objects such that $\Sigma n_k = n$. The sets are mutually exclusive and collectively exhaustive in that

$$Z = \bigcup_{k=1}^{K} Z_k \text{ and } \phi = \bigcap_{k=1}^{K} Z_k \tag{37}$$

where ϕ is the empty set. Note now that each probability $p_i \in Z_k$ is defined so that

$$P_k = \sum_{i \in Z_k} p_i \text{ and } \sum_k P_k = \sum_k \sum_{i \in Z_k} p_i = 1 \tag{38}$$

Substituting these definitions into equations (37) and (38), we can write the discrete entropy in equation (2) as

$$\begin{aligned} H &= -\sum_i P_k \log P_k - \sum_k P_k \sum_{i \in Z_k} \frac{p_i}{P_k} \log \frac{p_i}{P_k} \\ &= H_B + \sum_k P_k H_k \end{aligned} \tag{39}$$

where H_B is the between-set entropy at the higher system level, and the second term on the RHS of the second line of equation (39) is the sum of the within-set entropies H_k weighted by their probability of occurrence P_k at the higher level. As the sets Z_k get fewer and progressively larger from the original set Z—which is tantamount to disaggregation of the entire set into smaller and smaller sets—the within-set entropies decrease in sum and the between-set entropy H_E rises in value until all that remains is one aggregated set for each object, that is, $H_B \rightarrow H$. Moving the other way, when all the objects are aggregated into one set, then $H_B \rightarrow 0$, and $\Sigma P_k H_k \rightarrow H$. Proofs of these assertions are given in Theil (1972) and Webber (1979).

The equivalent decomposition formula for spatial entropy as we have defined it in equation (22) can be stated. Then, noting that

$$X_k = \sum_{i \in Z_k} \Delta x_i \tag{40}$$

where X_k is the sum of the intervals (areas) in each aggregated set Z_k, spatial entropy can be decomposed as

$$\begin{aligned} H_S &= -\sum_k P_k \log\left[\frac{P_k}{X_k}\right] - \sum_k P_k \sum_{i \in Z_k} \frac{p_i}{P_k} \log\left[\frac{p_i}{P_k} \Big/ \frac{\Delta x_i}{X_k}\right] \\ &= H_{SB} + \sum_k P_k H_{Sk} \end{aligned} \tag{41}$$

where H_{SB} is the between-set spatial entropy, and $\Sigma P_k H_{Sk}$ is the sum of the weighted within-set spatial entropies. An information difference structure is buried in equation (41), as spelled out earlier for spatial entropy between equations (33) and (36), and similar interpretations apply. In developing decompositions of entropy and spatial entropy in this fashion, the focus is on explaining the variation in entropy at different spatial scales, noting that entropies can be nested into a hierarchy of levels, that is, the between-set entropies can be further subdivided into sets that are smaller than Z_k but larger than the basic sets for each object or zone Z_i. These ideas have been used to redistrict zones to ensure equal populations in the case of the discrete entropy and equal population densities in the case of spatial entropy in an effort to design spatial systems that meet some criteria of optimality that pertain to scale and size (see Batty 1974, 1976). In this article, I do not deal with the effect of shape on entropy, but extensions exist to deal with idealized spatial systems that also incorporate constraints on shape, such as the regularity of boundaries, although developments in this area have been limited (Batty 1974).

These decomposed entropy measures can be used in entropy maximization to enable models to be derived that are constrained in different ways at different system levels. Let us assume that the cost constraint on probabilities pertains to the entire system, as in equation (5), but that entropy needs to be maximized so that the aggregate probabilities sum to those that are fixed by the level of decomposition or aggregation chosen, as in equation (38). I set up the Lagrangian to maximize equation (39) with respect to equations (38) and (5) as follows:

$$L = -\sum_k P_k \log P_k - \sum_k P_k \sum_{i \in Z_k} \frac{p_i}{P_k} \log \frac{p_i}{P_k} - \sum_k (\lambda_0^k - 1)\left(\sum_{i \in Z_k} p_i - P_k\right) \\ -\lambda_1\left(\sum p_i c_i - \bar{C}\right) \tag{42}$$

and then minimize the expression

$$\frac{\partial L}{\partial p_i} = -\log p_i - \lambda_0^k - \lambda_1 c_i = 0,\ i \in Z_k \tag{43}$$

to derive the model that we can state as

$$p_i = \exp(-\lambda_0^k - \lambda_1 c_i),\ i \in Z_k \tag{44}$$

We can compute the partition function directly by substituting for p_i in equation (38), yielding

$$\left.\begin{aligned} \exp(-\lambda_0^k) &= \frac{P_k}{\sum_{i \in Z_k} \exp(-\lambda_1 c_i)} \text{ or} \\ \lambda_0^k &= \log \frac{\sum_{i \in Z_k} \exp(-\lambda_1 c_i)}{P_k} \end{aligned}\right\} \tag{45}$$

from which the relevant exponential model in equation (44) can be more clearly written as

$$p_i = P_k \frac{\exp(-\lambda_1 c_i)}{\sum_{i \in Z_k} \exp(-\lambda_1 c_i)},\ i \in Z_k \text{ and } \sum_{i \in Z_k} p_i = P_k \tag{46}$$

Note that the constraint equation on cost is for the entire system and, as such, effectively couples the various models for each subset in terms of their calibration but not in terms of their operation.

We need to be careful about the way these models are coupled because if no system-wide constraints exist, then the models are separable; the entropy maximizing is separable into K subproblems. For example, assume that the cost constraint in equation (5) is replaced with cost constraints that pertain to the subsets written as

$$\sum_{i \in Z_k} \frac{p_i}{P_k} c_i = \bar{C}_k,\ \forall k \tag{47}$$

Then, from equation (47), that the system-wide constraint is also met as

$$\sum_k P_k \sum_{i \in Z_k} \frac{p_i}{P_k} c_i = \sum_k P_k \bar{C}_k = \sum_k \sum_{i \in Z_k} p_i c_i = \bar{C},\ \forall k \tag{48}$$

If we substitute equation (48) in (42), noting that now we have K multipliers λ_1^k, then the derived model has the same structure as equation (44) but now can be written, following equation (46), as

$$p_i = P_k \frac{\exp(-\lambda_1^k c_i)}{\sum\limits_{i \in Z_k} \exp(-\lambda_1^k c_i)}, \; i \in Z_k \tag{49}$$

This model is not only separable for each subset Z_k, but each model also is calibrated separately with respect to the cost constraint and determination of the set of multipliers $\{\lambda_1^k\}$. Using spatial entropy maximizing adds little to this logic other than ensuring that the interval or area for each zone appears in the exponential equation. If we follow the same process, the equivalent model to that in equation (49) can be written as

$$p_i = P_k \frac{\Delta x_i \exp(-\lambda_1^k c_i)}{\sum\limits_{i \in Z_k} \Delta x_i \exp(-\lambda_1^k c_i)}, \; i \in Z_k \tag{50}$$

where if the system-wide cost constraint in equation (5) applies, then the only difference is that there is one multiplier, λ_1, not K. To provide some sense of closure to this argument, readers are referred to Theil (1972), who provides many applications of these kinds of decomposition to the measurement of variance and difference at different levels of disaggregation for both spatial and nonspatial systems, connecting these ideas to a much wider literature about the measurement of inequality.

Generating spatial probability distributions

So far I have defined both entropy and its method of maximization with respect to probabilities that pertain to spatial locations. In terms of the typical problem, there is the assumption that the probability of location is some function—a negative exponential in the classic Boltzmann–Gibbs case—of some size variable such as cost. Implicitly, in this case, the probability of location might be proportional to the observed population in any zone, and a sensible assumption is that a higher probability of location measured by a higher population is associated with a lower cost (or higher benefit) of locating in the place in question. However, another interpretation exists that is less specific about the kinds of probability distributions that emerge from entropy maximizing and depends on how one sets up a problem. In this section and in the rest of this article, we can assume that some measure of size, not cost, is what a probability distribution must conserve, and that probabilities vary with respect to this size variable. In short, rather than thinking of the spatial location problem as one in which the probability of population location is related to some size or cost, we now develop a model in which the probability of location is dependent on the actual population size that is observed in the locations in question. This is the obvious way to develop entropy maximizing for city–size distributions, a topic that has remained quite confused since Berry (1964) and Curry

(1964) first speculated about these questions over 40 years ago. This is also the route by which we can connect the arguments of this article to size distributions in general and to power laws in particular.

To extend entropy maximizing in this way, I replace the probability p_i of each event with its frequency $f(\circ)$. I define a function of the size of the event V_i, which in many of these cases literally is the population size, although it could be defined as any related measure. Then I derive the appropriate discrete probability frequency for $f(V_i)$ by maximizing its entropy H defined in analogy to equation (2) as

$$H = -\sum_i f(V_i) \log f(V_i) \tag{51}$$

This expression is subject to the usual normalization and constraints associated with the moments of the distribution that are defined as

$$\sum_i f(V_i) = 1, \sum_i f(V_i) V_i = \bar{V}, \sum_i f(V_i)\left[V_i^2 - \bar{V}\right] = \sigma^2, \text{ and so on} \tag{52}$$

where $\sigma^2$is the variance of the distribution. This discussion and notation follows Tribus (1969), although several other presentations of this process have more formal roots in probability theory. A good contemporary start for these more formal presentations between entropy, scale, and scaling can be found in the books by Sornette (2006) and Saichev, Malevergne, and Sornette (2010).

The Boltzmann–Gibbs negative exponential model is still the baseline in entropy maximization because it introduces a constraint on the distribution that is the first moment, the average, and no others apart from the normalization of the probabilities. Following the same logic used earlier in equations (4)–(10) and assuming the intervals over which the discrete frequency is measured are equal, that is, $\Delta x_i = \Delta x, \forall i$ (to avoid any confusion with spatial entropy at this stage), we maximize equation (51) subject to the first two constraints shown in (52). Using the relevant Lagrangian with appropriate multipliers yields

$$\log f(V_i) = -\lambda_0 - \lambda_1 V_i \tag{53}$$

This has the classic log-linear form that generates the Boltzmann–Gibbs probability frequency

$$f(V_i) = \exp(-\lambda_0 - \lambda_1 V_i) \tag{54}$$

which gives the familiar exponential form

$$f(V_i) = \frac{\exp(-\lambda_1 V_i)}{\sum_i \exp(-\lambda_1 V_i)} \tag{55}$$

Equation (55) implies that the larger the size, the lower the probability, which is the same as the previous interpretation with size equivalent to locational cost. This relationship is made more graphic if we rearrange equation (53), where size is now a function of frequency, as

$$V_i = -\frac{\lambda_0}{\lambda_1} - \frac{1}{\lambda_1}\log f(V_i) \tag{56}$$

However, if the size V_i is population as measured in terms of the number of individuals living in zone *i*, then we cannot equate cost with size in any way because larger populations are much more likely to live in places where the costs of location are lower, all other things being equal. This feature is the confusion that has never really been resolved in generating size distributions with entropy-maximizing techniques. The motivation for the earlier models, such as those developed by Wilson (1970), was always to maximize entropy with respect to a cost constraint, whereas for the models in this section, the motivation is to maximize entropy with respect to a size constraint. In this context, a perfectly reasonable assumption is that an individual locating across a space has many more places to locate where populations are small than places where populations are large. It is in this sense that frequency in this section differs from probability in the previous sections, although formally the algebraic expressions are identical.

Now we can show how the negative exponential can become a power function if the constraint on average size is replaced by its geometric equivalent, that is,

$$\sum_i f(V_i)\log V_i = \log \bar{V} \tag{57}$$

where $\log \bar{V}$ is the expected value of the sum of the logarithms of the sizes. The logic is that agglomeration economies of size or diseconomies of cost or energy are perceived logarithmically rather than absolutely, as enshrined in the Weber–Fechner law (Stevens 1957). We assume this perception is defined for a discrete system because difficulties noted below arise when we examine the rank–size rule and its consistency with entropy maximization. The continuous version of the model must be invoked for purposes of simplification and demonstration. However, if we maximize entropy subject to equation (57) and the normalization constraint, the model becomes

$$\log f(V_i) = -\lambda_0 - \lambda_1 \log V_i \tag{58}$$

which in exponential form is

$$f(V_i) = \exp(-\lambda_0)V_i^{-\lambda_1} \tag{59}$$

Equation (59) is a power function that, in more familiar terms, can be written as

$$f(V_i) = \frac{V_i^{-\lambda_1}}{\sum_i V_i^{-\lambda_1}} \tag{60}$$

where, from equation (60), we can write the model in inverse form in analogy to equation (56) as

$$V_i = \exp\left(-\frac{\lambda_0}{\lambda_1}\right) f(V_i)^{-\frac{1}{\lambda_1}} \tag{61}$$

In this context, V_i also varies inversely with the power of frequency. From equation (61), which, in turn, is derived from the assumption about logarithmic costs made in equation (57), we can generate the more familiar rank–size rule that has been known for well over a century, first exploited for income sizes by Pareto (1906) and then for city sizes by Zipf (1949). I explore these functions in the next section.

In maximizing entropy with respect to the three constraints stated in equations (52), one notes that the third constraint also can be simplified to

$$\sum_i f(V_i)(V_i - \bar{V})^2 = \sum_i f(V_i)V_i^2 - \bar{V} = \sigma^2 \tag{62}$$

yielding

$$\log f(V_i) = -\lambda_0 - \lambda_1 V_i - \lambda_2 V_i^2 \tag{63}$$

In the first exponential form, this is

$$f(V_i) = \exp(-\lambda_0 - \lambda_1 V_i - \lambda_1 V_i^2) \tag{64}$$

which in more familiar terms is

$$f(V_i) = \frac{\exp(-\lambda_1 V_i - \lambda_1 V_i^2)}{\sum_i \exp(-\lambda_1 V -_i \lambda_1 V_i^2)} \tag{65}$$

As Tribus (1969) shows, equation (65) is a form of the normal distribution. The entropy-maximizing derivation is interesting because it makes explicit the polynomial form of the normal with the contribution of the mean and the variance directly associated with the multipliers λ_1 and λ_2. The parameter λ_1 is negative, making this exponential positive, and λ_2 is positive, meaning the variance term acts as a negative exponential. The normality of the distribution is always preserved no matter what the value of these multipliers. Moreover, if $\lambda_1 \ll \lambda_2$, the variance of the distribution becomes increasingly smaller, while the skewness become increasingly peaked. We can complete this set of distributions by assuming that the size distribution is log-normal, that is, instead of V_i, we now define size as its logarithm, log V_i. We can formally restate the constraint equations for the log-normal as

$$\left.\begin{aligned} &\sum_i f(V_i) = 1 \\ &\sum_i f(V_i)\log V_i = \log \bar{V} \\ &\sum_i f(V_i)\Big[(\log\ V_i)^2 - \log\ \bar{V})\Big] = \sigma^2 \end{aligned}\right\} \tag{66}$$

Maximizing equation (51) subject to equations (66) gives the model in final form as

$$\begin{aligned} f(V_i) &= \frac{\exp(-\lambda_1 \log V_i - \lambda_1 (\log V_i)^2)}{\sum_i \exp(-\lambda_1 \log V_i - \lambda_1 \log (V_i)^2)} \\ &= \frac{V_i^{-\lambda_1} (V_i^2)^{-\lambda_2}}{\sum_i V_i^{-\lambda_1} (V_i^2)^{-\lambda_2}} \end{aligned} \tag{67}$$

Equation (67) implies that, if $\lambda_1 \ll \lambda_2$, the log-normal form collapses to the inverse power law form but only for a range of the largest values of V_i. This is one of the simplest demonstrations that power laws tend to dominate in the upper or heavy tail of the log-normal distribution. Again, the same caveats apply as for the existence of the moments for the discrete case, which will always be true for the sorts of spatial systems to which these models apply, that is, where $1 \leq V_i < \infty$. Tribus (1969) has a relatively straightforward demonstration of the properties of the normal distribution with respect to the values of the parameters that can be determined from an approximation to the continuous probability density function.

Approximating scaling: the rank–size rule and Zipf's law

The negative exponential and power law models generated in the previous section using entropy maximizing represent discrete density functions relating frequency to size. These distributions already define the form of the population or city–size distributions (where spatial locations i define the locations of distinct cities). However, a more popular form, particularly for city sizes, firm sizes, incomes, and related social phenomena involves ranking these sizes from the largest value of V_i, which I now call rank r_1, to the smallest, r_n. The rank is the countercumulative of the frequency (Adamic 2002). If we accumulate the frequencies from, let us say, some value of $i = m < n$ to the largest value of $i = n$, then this accumulation would define the rank r_{n-m}. We can only express this formally if we consider the continuous approximation to $f(V_i)$ as $f(V)$, which is defined when $\Delta x_i \to 0$. Let us first take the exponential model defined in equation (55) in its continuous limit as $f(V) \sim \exp(-\lambda_1 V)$. The integration defining the countercumulative $F(V)$ is

$$F(V) = \int_V^\infty f(V) dV \sim \int_V^\infty \exp(-\lambda_1 V) dV = \frac{1}{\lambda_1} [\exp(-\lambda_1 V)]_V^\infty \tag{68}$$

where $F(V) \sim r_{n-m} = r_k$, $i = m$, and $k = n - i$. Thus,

$$r_k \sim \exp(-\lambda_1 V_k) \tag{69}$$

from which

$$\left.\begin{aligned} \log r_k &\sim -\lambda_1 V_k \\ V_k &\sim \frac{1}{\lambda_1}\log\left(\frac{1}{r_k}\right) \end{aligned}\right\} \tag{70}$$

Equations (70) define rank as a function of population and population as a function of rank, which exposes the clear log-linear structure of the exponential rank–size relationship.

The classic rank–size relationship commonly is developed with the relationship between size and frequency expressed as a power law. The continuous limit based on equation (60) can be written as $f(V) \sim V^{-\lambda_1}$, from which we define the countercumulative $F(V)$ as

$$F(V) = \int_v^\infty f(V)dV \sim \int_v^\infty V^{-\lambda_1} dV = \frac{1}{\lambda_1 + 1}\left[V^{-\lambda_1+1})\right]_v^\infty \tag{71}$$

where $F(V)$ is the rank r_k as defined for the integration of the exponential following equation (68). This rank can be written as

$$r_k \sim V_k^{-\lambda_1+1} \tag{72}$$

from which

$$\left.\begin{aligned} \log r_k &\sim (1-\lambda_1)\log V_k \\ V_k &\sim r_k^{-\frac{1}{1-\lambda_1}} \end{aligned}\right\} \tag{73}$$

Equations (73) define rank as a function of population, and population as a function of rank. Equations (72)–(73) imply that these power laws are scaling, that is, if we scale size by α as αV_k, then the rank does not change, which can be demonstrated by substituting αV_k for V_k in any of the preceding equations. A power law is the only function that has this property; hence, its claim as a signature of universality.

Using the logarithmic mean of the size as the major constraint in generating distributions in the inverse power or Zipf–Pareto form is consistent with assuming that size (or cost) can be viewed as a regular distortion based on human perception. We noted this feature previously as the Weber–Fechner law, which pertains to how we perceive brightness and sound. Even the way our cognitive senses respond to size is proportional to the logarithm, not to the actual value, of the relevant measure of intensity (see Stevens 1957). In spatial interaction modeling, Wilson (1970) made use of this property to show how the original gravitational hypothesis is consistent with models produced by entropy maximizing, particularly in the context of very long distance flows, such as those measured as commodities in trade systems, where the perception of travel cost is more likely to be logarithmic than absolute. The same arguments are used to incorporate additional constraints that might be thought of as benefits rather than costs, reflecting the fact that agglomeration economies are sometimes perceived logarithmically.

We also can generate rank–size distributions for the normal and log-normal models that we derived in equations (65) and (67), respectively. Although pursuing this development is not very illuminating, the log-normal is a noteworthy special case largely because many arguments exist suggesting that city, firm, and income size distributions are not consistent with power laws, but rather are log-normal, with the power law only applying as an approximation to these distributions in their upper tail. Writing equation (67), noting the signs of the multipliers as determined by Tribus (1969), expressing the first multiplier as α and the second as β, and then passing to the limit renders $f(V) \sim V^{\alpha}V^{-2\beta}$, from which we form the countercumulative as

$$F(V) = \int_{V}^{\infty} f(V)dV \sim \int_{V}^{\infty} V^{\alpha}V^{-2\beta}dV = \frac{1}{\alpha - 2\beta + 1}\left[V^{\alpha-2\beta+1})\right]_{V}^{\infty} \qquad (74)$$

Equation (74) indicates that the shape of the log-normal is completely dependent on the value of the parameters α and β. Nevertheless, we can speculate on the shape of the function for various ranges of size from these values and the size $\{V_i\}$. The rank and size relationships, analogous to equation (73), can be written as

$$\left.\begin{array}{l} \log r_k \sim (\alpha + 1)\log V_k - 2\beta \log V_k \\ V_k \sim r_k^{\frac{1}{(\alpha+1-2\beta)}} \end{array}\right\} \qquad (75)$$

If $\alpha+1 \ll 2\beta$, then for the largest values of V_k the second term in the first line of equation (75) dominates, implying that the rank–size relation is more like a power law in its upper or heavy tail.

The preceding development is a somewhat informal way of demonstrating the relationship between inverse power and log-normal functions, and readers are referred to more considered sources that elaborate this relationship. Perline (2005) formalizes an excellent discussion about when one is able to approximate the heavy tail of a log-normal with a power law that builds on earlier expositions that are part of the literature on skewed probability functions, as summarized by Montroll and Schlesinger (1982). The purpose here is not to develop a treatise about the log-normal or, indeed, about the Zipf and Pareto power laws, for we see that both can be derived from entropy maximizing. Rather, power laws can emerge from two sources: (1) directly if the constraint on the entropy is a geometric mean and (2) when the constraints on the entropy are those that define the log-normal but for very large values of the size distribution where the variance of the distribution is also very large, effectively meaning that the heavy tail occurs over several orders of magnitude. For empirical applications to city–size distributions, readers are referred to the mainstream literature where these issues are discussed in great detail. The recent article by Eeckhout (2004) is representative.

One last substantive issue requires us to complete this presentation about how scaling distributions are associated with entropy maximizing. The traditional explanation of how power laws come to dominate spatial and social distributions essentially is based on a generic model that leads to agglomeration economies, in which any object chosen at random increasingly is unlikely to grow to a very large scale, realizing agglomeration economies that are associated with large cities, people with large incomes, the domination of large firms, and so on. In essence, the growth or decline in size of any object making up such competitive systems is based on Gibrat's (1931) law of proportionate effect, in which any object of size V_{it} grows or declines to V_{it+1} by a random amount ε_{it}, whose value is proportionate to the size of the object already reached, that is, $V_{it+1} = (1 + \varepsilon_{it})V_{it}$. This process, if operated continually for many time periods, leads to a distribution of objects that is log-normal. If the process is constrained so that objects do not decline in size below a certain threshold (which is tantamount to not letting size become negative), several authors show that the resultant distribution is no longer log-normal but rather is scaling in the form of an inverse power function. These conclusions have emerged from several sources in physics (Levy and Solomon 1996), in economics (Saichev, Malevergne, and Sornette 2010), in earth sciences (Sornette 2006), and in several other areas of social inquiry (Newman 2005).

This dynamic, referred to by Solomon (2000) as the "generalized Lotka–Volterra (GLV) model," essentially illustrates that in the steady state, power laws emerge from processes in which there is random proportionate growth against a background of transitions between individuals or places in terms of the interest variable, be it population, income, wealth, cost, or some other size measure. The steady-state results generated by such models also are consistent with Boltzmann distributions, as Richmond and Solomon (2001) show, while Foley (1994) and then Milakovic (2003) demonstrate that entropy maximizing can be employed directly with the dynamics being embedded as constraint equations that the process of wealth creation must meet. An enormous literature now exists that deals with stochastic GLV types of models, which build on proportionate effect, leading to log-normal and power laws. Several oblique interpretations of the steady states associated with such processes as Boltzmann–Gibbs distributions exist, which Richmond and Solomon (2001) say are ". . . Boltzmann laws in disguise." The earlier dynamic models developed by Drãgulescu and Yakovenko (2000) also are being extended to deal with systems where the constraints on distributions of money, wealth, and income all vary with consequent differences in their distributions, in turn, providing a rich source of interpretations for the way inequalities emerge in economic systems (Yakovenko and Rosser 2009).

Little of this discussion has yet to find its way into spatial or geographical systems because the concern with city–size distributions has been remarkably aspatial, in contrast to entropy maximizing in geographical analysis; but signs of a convergence are beginning to appear. Wilson's (2008) recent work, for example, seeks to generalize entropy maximizing in a dynamical framework that he refers to as

Boltzmann–Lotka–Volterra models, which have clear links to GLV models. At present, approaches to dynamics can be seen as either constructing a Lotka–Volterra dynamic that leads to Boltzmann–Gibbs and related distributions, or to Boltzmann–Gibbs distributions that are nested within a Lotka–Volterra dynamic. Much synthesis needs to be done, and many fruitful insights can be gained by these extensions. After a period of reflection and consolidation, a rebirth of interest in measures of entropy and entropy maximizing in geographical analysis is now entirely possible through developments in modern systems theory that now fall under the guise of the complexity sciences (Batty 2009).

Future research: alternative entropies, more explicit dynamics

In this article, I argued that one of the issues that has never been systematically tackled with respect to the application of entropy measures and methods in geographical analysis involves a thorough interpretation of what the various measures actually mean in terms of spatial distributions with respect to their size, scale, and shape. The Shannon entropy measure in equation (2) is only one of many such measures, albeit perhaps the most natural in that it satisfies the multiplicity requirement for independent events in terms of the additivity of information as defined in equation (1). But if events are not independent and if the entropy phase space is structured in ways that do not allow probabilistic events to occur in all parts of the space, then the Shannon measure is not necessarily the most appropriate. In geographical systems, events can be highly autocorrelated in space as well as time, and thus the methods used to generate probability distributions in equilibrium or in the steady state can be badly compromised if more appropriate measures are not chosen.

Among these, the measure proposed by Rényi (1961) introduces a parameter α that gives greater weight to larger probabilities if the parameter is >1, lesser weight if <1, and is the same as the Shannon entropy for $\alpha = 1$. It has many similar properties to the Shannon measure in terms of its maximum and minimum but could be more useful for spatial systems where larger probabilities imply greater importance. Few, if any, applications in this field exist (but see March and Batty 1975), and thus this measure is worth exploring further. A more radical form of measure broaches directly the question of the independence of events and breaks with the assumptions in equation (1) defining a measure of joint information for any addition of information due to a sequence of probability events. This is called Tsallis entropy, which Tsallis (2004) argues represents an entropy where events are nonextensive; that is, events that apply to a more structured phase space than that assumed for the original Shannon measure. The attraction of the Tsallis measure (which formally is not unlike the Rényi entropy) is that in its use in maximization, the resulting model is an inverse power law, not the negative exponential. All of these measures can be decomposed for different scales, continuous equivalents can be approximated, and they can be reconciled with methods and models that generate their form as either equilibrium distributions or as the outcome of stochastic proportional growth processes. An agenda for testing their applicability to geographical systems would not be hard to fashion.

Wilson's (1970) contribution, however, is that he introduced a framework for generating consistent models, rather than a set of methods, for enabling measurement of actual entropies. Actual measures do fall out along the way, but the real power of the entropy-maximizing framework that he introduced is in the generation of specific and applied models and the demonstration that entire families of models could be pictured across a spectrum of possible types. In this sense, his methods provide a lasting framework for the derivation of operational models that continue to be useful, indeed essential, in consistently specifying and coupling different models together. The development of entropy maximizing in generating economic models came much later and has yet to adopt the systematic procedures demonstrated for spatial systems by Wilson (in this issue). Yet, despite its power, entropy maximizing is compromised somewhat because space itself should be directly incorporated into the framework so that dimensional consistency can be ensured through the use of spatial entropy rather than its discrete equivalent, the Shannon measure. The existing practice of defining operational location and interaction models has not really followed these procedures; nor has it systemically examined the sets of constraints necessary to define particular problems with respect to what is known and not known about the systems of interest. Much remains to be done in using entropy maximizing to establish formalized methods for aiding the spatial model-building process.

Last, but not least, dynamics has slowly entered the picture. The great attraction of the framework when it was first proposed 40 years ago was its ability to generate models in equilibrium. Dynamics was assumed to be benign, even to the point where simple models (such as those used to move money around in an economic system), developed only a decade or so ago by researchers such as Drãgulescu and Yakovenko (2000), have never been explored in spatial analysis. Seeing geographical systems in equilibrium was enough. When Wilson (2008), among others, following the tradition established by Harris and Wilson (1978), began to explore how such models could be made dynamic, they decoupled the dynamics from the statics, assuming that Boltzmann–Gibbs models represented a shorter, faster equilibrium that could be nested in the longer-term dynamics associated with the models originally proposed by Lotka and Volterra. As argued in the preceding section, now a new momentum is emerging. These different but related approaches are generating a new synergy about how geographical systems develop, consistent with emergence and far-from-equilibrium structures, as well as new concepts about how to model such systems from the bottom up. During the last 20 years, entropy in geographical systems has no longer been at the cutting edge. Now, however, there is every sign that these ideas will be resurrected as part of the burgeoning interest in complexity science, which is forcing upon us the notion that equilibrium is a convenient fiction that we must move beyond.

Note

[1] As quoted in *Scientific American* 225(3) (1971), 180.

References

Adamic, L. A. (2002). ''Zipf, Power-laws, and Pareto—A Ranking Tutorial.'' Information Dynamics Lab, HP Labs, Palo Alto, CA. Available at http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html (accessed April 1, 2010).

Batty, M. (1974). ''Spatial Entropy.'' *Geographical Analysis* 6, 1–31.

Batty, M. (1976). ''Entropy in Spatial Aggregation.'' *Geographical Analysis* 8, 1–21.

Batty, M. (1983). ''Cost, Accessibility and Weighted Entropy.'' *Geographical Analysis* 15, 256–67.

Batty, M. (2009). ''Cities as Complex Systems: Scaling, Interactions, Networks, Dynamics and Urban Morphologies.'' In *Encyclopaedia of Complexity and Systems Science*, Vol. 1: 1041–71, edited by R. Meyers. Berlin: Springer.

Batty, M. (2010). ''Visually-Driven Urban Simulation: Exploring Fast and Slow Change in Residential Location.'' Working Paper 162, Centre for Advanced Spatial Analysis, University College London, U.K. Available at http://www.casa.ucl.ac.uk/working_papers/paper156.pdf (accessed September 1, 2010).

Ben-Naim, A. (2008). *A Farewell to Entropy: Statistical Thermodynamics Based on Information Theory*. Singapore: World Scientific Publishing Corporation.

Berry, B. J. L. (1964). ''Cities as Systems within Systems of Cities.'' *Papers and Proceedings, Regional Science Association* 13, 147–63.

Curry, L. (1964). ''The Random Spatial Economy: An Exploration in Settlement Theory.'' *Annals of the Association of American Geographers* 54, 138–46.

Drãgulescu, A. A., and V. M. Yakovenko. (2000). ''Statistical Mechanics of Money.'' *The European Physical Journal B* 17, 723–29.

Eeckhout, J. (2004). ''Gibrat's Law for (All) Cities.'' *American Economic Review* 94, 1429–51.

Erlander, S., and N. F. Stewart. (1990). *The Gravity Model in Transportation Analysis: Theory and Extensions*. Utrecht, Netherlands: VSP.

Foley, D. K. (1994). ''A Statistical Equilibrium Theory of Markets.'' *Journal of Economic Theory* 62(2), 321–45.

Gibrat, R. (1931). *Les Inegalités Economiques*. Paris: Librarie du Recueil, Sirey.

Goldman, S. (1968). *Information Theory*. New York: Dover Publications.

Harris, B., and A. G. Wilson. (1978). ''Equilibrium Values and Dynamics of Attractiveness Terms in Production-Constrained Spatial-Interaction Models.'' *Environment and Planning A* 10, 371–88.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Levy, M., and S. Solomon. (1996). ''Power Laws are Logarithmic Boltzmann Laws.'' *International Journal of Modern Physics C* 7(4), 595–601.

March, L., and M. Batty. (1975). ''Generalised Measures of Information, Bayes' Likelihood Ratio and Jaynes' Formalism.'' *Environment and Planning B* 2, 99–105.

Milakovic, M. (2003). ''Maximum Entropy Power Laws: An Application to the Tail of Wealth Distributions.'' LEM Papers Series 2003/01. Laboratory of Economics and Management, Sant'Anna School of Advanced Studies, Pisa, Italy.

Montroll, E. W., and M. F. Schlesinger. (1982). ''On 1/f Noise and Other Distributions with Long Tails.'' *Proceedings of the National Academy of Sciences USA* 79, 3380–83.

Morphet, R. (2010). ''Thermodynamic Potentials and Phase Change for Transport Systems.'' Working Paper 155, Centre for Advanced Spatial Analysis, University College, London, UK. Available at http://www.casa.ucl.ac.uk/working_papers/paper155.pdf (accessed July 1, 2010).

Newman, M. E. J. (2005). ''Power Laws, Pareto Distributions and Zipf's Law.'' *Contemporary Physics* 46, 323–51.

Pareto, V. (1906). *Manual of Political Economy*. (English translation, Ann S. Schwier, 1971). New York: Augustus M. Kelley Publishers.

Perline, R. (2005). ''Strong, Weak and False Inverse Power Laws.'' *Statistical Science* 20(1), 68–88.

Rényi, A. (1961). ''On Measures of Information and Entropy.'' *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability* 1960, 547–561. Available at http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf (accessed July 26, 2010).

Richmond, R., and S. Solomon. (2001). ''Power Laws Are Disguised Boltzmann Laws.'' *International Journal of Modern Physics C* 12(3), 333–43.

Roy, J. R., and J. C. Thill. (2004). ''Spatial Interaction Modeling.'' *Papers in Regional Science* 83, 339–61.

Rybczynski, W. (2010). ''Dubai Debt: What the Burj Kahlifa—the Tallest Building in the World—Owes to Frank Lloyd Wright.'' *Slate*, January 13. Available at http://www.slate.com/id/2241275/ (accessed July 27, 2010).

Saichev, A., Y. Malevergne, and D. Sornette. (2010). *Theory of Zipf's Law and Beyond. Lecture Notes in Economics and Mathematical Systems 632*. Heidelberg, Germany: Springer.

Shannon, C. E. (1948). ''A Mathematical Theory of Communication.'' *Bell System Technical Journal* 27, 379–423, 623–56.

Snickars, F., and J. W. Weibull. (1977). ''A Minimum Information Principle: Theory and Practice.'' *Regional Science and Urban Economics* 7, 137–68.

Solomon, S. (2000). ''Generalized Lotka Volterra (GLV) Models of Stock Markets.'' In *Applications of Simulation to Social Sciences*, 301–22, edited by G. Ballot and G. Weisbuch. Oxford, UK: Hermes Science Publications.

Sornette, D. (2006). *Critical Phenomena in Natural Sciences*, 2nd ed. Heidelberg, Germany: Springer.

Stevens, S. S. (1957). ''On the Psychophysical Law.'' *Psychological Review* 64(3), 153–81.

Theil, H. (1972). *Statistical Decomposition Analysis*. Amsterdam: North Holland.

Tribus, M. (1969). *Rational, Descriptions, Decisions and Designs*. New York: Pergamon Press.

Tsallis, C. (2004). ''Nonextensive Statistical Mechanics: Construction and Physical Interpretation.'' In *Nonextensive Entropy: Interdisciplinary Applications*, 1–53, edited by M. Gell-Mann and C. Tsallis. Oxford, UK: Oxford University Press.

Webber, M. J. (1979). *Information Theory and Urban Spatial Structure*. London: Croom Helm.

Wilson, A. G. (1970). *Entropy in Urban and Regional Modelling*. London: Pion Press.

Wilson, A. G. (2008). ''Boltzmann, Lotka and Volterra and Spatial Structural Evolution: An Integrated Methodology for Some Dynamical Systems.'' *Journal of the Royal Society, Interface* 5, 865–71.

Yakovenko, M., and J. B. Rosser Jr.. (2009). ''Colloquium: Statistical Mechanics of Money, Wealth, and Income.'' *Reviews of Modern Physics* 81, 1703–25.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.