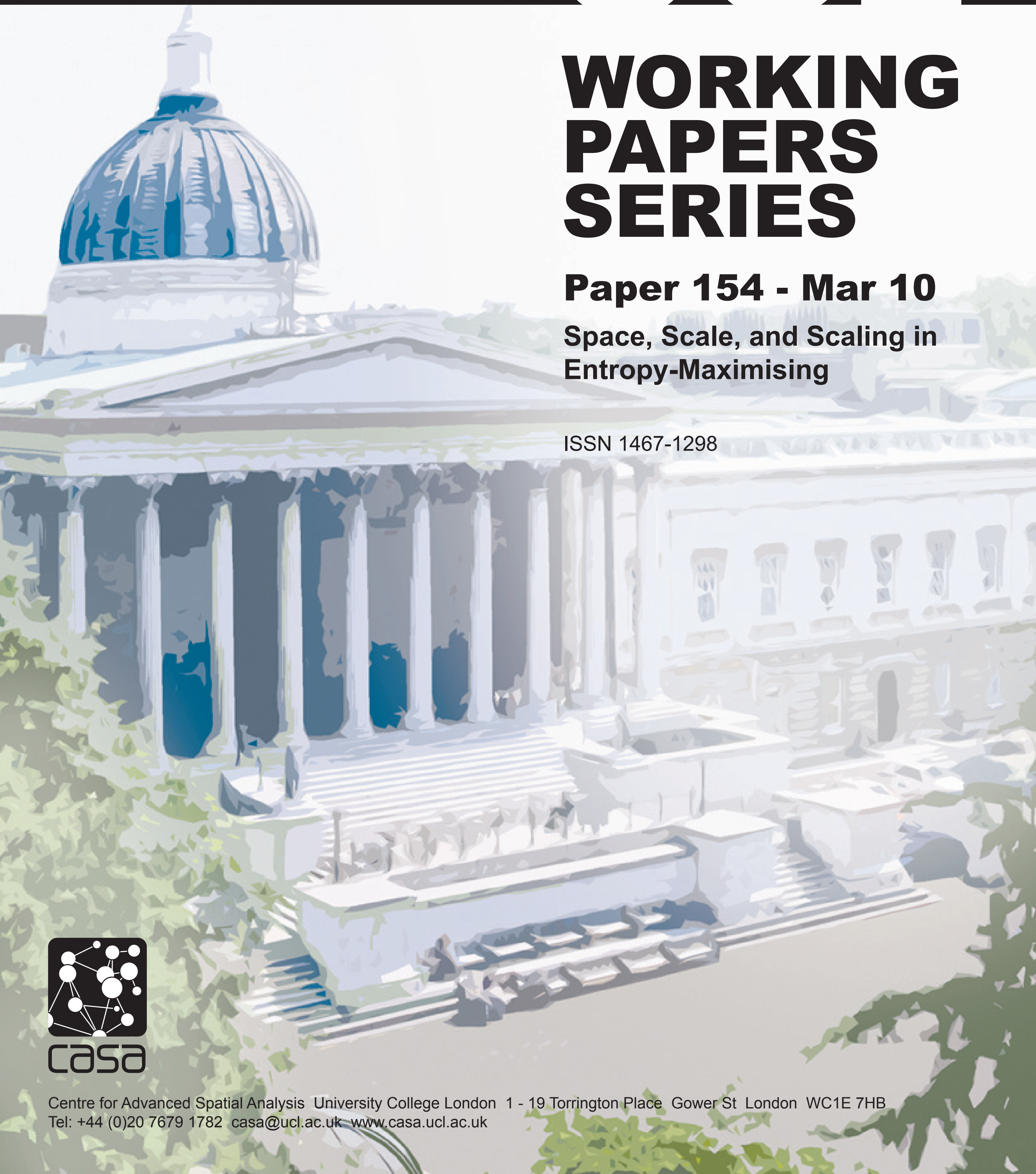
UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

UCL



WORKING PAPERS SERIES

Paper 154 - Mar 10

Space, Scale, and Scaling in Entropy-Maximising

ISSN 1467-1298

casa

Centre for Advanced Spatial Analysis University College London 1 - 19 Torrington Place Gower St London WC1E 7HB
Tel: +44 (0)20 7679 1782 casa@ucl.ac.uk www.casa.ucl.ac.uk

Space, Scale, and Scaling in Entropy-Maximising

Michael Batty

Centre for Advanced Spatial Analysis (CASA)
University College London (UCL)
1-19 Torrington Place, London WC1E 6BT, UK
m.batty@ucl.ac.uk www.casa.ucl.ac.uk

1 March 2010

Abstract

Entropy measures were first introduced into geographical analysis during a period when the concept of human systems as being in some sort of equilibrium was in the ascendancy. In particular, entropy-maximising, in direct analogy to equilibrium statistical mechanics, provided a powerful framework in which to generate location and interaction models. This was introduced and popularised by Wilson (1970) and it led to many different extensions that filled in the framework rather than progressed it to different kinds of models. In particular, we 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, however, have remained 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-maximising, substantive interpretations of entropy measures, and the longer term dynamics of how equilibrium distributions are reached and maintained have not been well-developed. This has many implications for future research and in conclusion, we will sketch the need for new and different entropy measures as well as new forms of dynamics that enable us to see how equilibrium spatial distributions can be generated as the outcomes of dynamic processes that converge to the steady state.

Defining Entropy, Interpreting Entropy

If an event occurs with a probability p, then this will give us a measure of information about the likelihood of that probability being correct. Any event with a very low probability which 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 this as $1/p$. However if we have two independent events with probabilities p_1 and p_2, then if one occurs and then the other occurs, we would expect the information gained to be $1/p_1 p_2$ because the probability of their joint occurrence is $p_1 p_2$. 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 $1/p_1 + 1/p_2 \neq 1/p_1 p_2$ but a function $F(\circ)$ of which the only solution is the logarithm of the inverse of the probability, that is

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

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

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

$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad . \qquad (2)$$

This measure was first defined in this form by Shannon (1948) when considering communication of information over a noisy channel but in fact the formula is central to statistical physics, originating with Clausius in the early 19th century but given specific statistical interpretation by Boltzmann and then Gibbs as the measure for thermodynamic entropy. In particular, the method of entropy-maximising 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 which serves as the baseline for many of the distributions of spatial activity that we will introduce. In fact, when Shannon (1948) introduced this measure, he sought advice 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

[1] (as quoted in **Scientific American**, **225** No. 3, 1971, p. 180).

important, no one really knows what entropy really is, so in a debate you will always have the advantage!"

There are many attractive properties of this function for describing spatial distributions. Here we will initially assume that the probability p_i is some count or density of spatial activity such as population in a zone i which 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, then $p_i = 1$ and $p_k = 0, \forall k \neq i$ and the entropy would be at a minimum where $H_{\min} = 0$. If the population were evenly spread throughout the tracts as $p_i = 1/n, \forall i$, then the entropy would be at a maximum where $H_{\max} = \log n$. Many distributions lie between these extremes and it is possible to construct a variety of related measures that make comparisons with the maximum. For example a measure of information difference can be constructed as

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

The term on the right hand side of the second line of equation (3) is an information difference of the kind widely used in likelihood theory, first popularised by Kullback (1959), and discussed widely in a geographical context by Snickers and Weibull (1977) and Webber (1979). A normalisation 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 comprising the distribution. This is intuitively acceptable for as we have more events, we have more information unless the additional events have zero probability of occurrence. This is easy to show as $H_{\max} = \log n$ 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 normalise it in some way as in equation (3) above and the development of spatial entropy which we will present below is one strategy for doing this. This lack of comparability has meant that methods for deriving spatial probability distributions have been much more to the fore in geographical analysis rather than more substantive interpretations of the entropy measure. This has been unfortunate because there are some important conclusions to be drawn about the structure of different spatial systems with respect to measures of entropy. This is an unfinished quest.

For example, if we consider the hypothetical system in which all the population is piled into one zone, the mile high building example, then such as 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, in systems where the population is spread out evenly, there is maximum entropy, maximum disorder, in that this is the situation which would emerge when there were *no constraints* on the system and every person could live where they wanted. 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 that there are direct connections to thermodynamic entropy where maximum disorder occurs when all particles mix freely which occurs when temperature in the system rises and any differences are ironed out. In fact the order-disorder continuum with respect to H is directly invoked if we consider that as we put more and more constraints on what the form of the distribution is, we successively reduce the entropy. In this sense, there is a direct tie up between the probability distributions which we observe and model and the methods of deriving such distributions using the method of entropy maximisation to which we now turn. We will first present the method for this relates directly to that pioneered by Wilson (1970) for urban and regional systems, although after this, we will describe many new insights that seek to show how such methods can be extended to deal with space, scale, and scaling.

The Entropy-Maximising Framework

To choose a probability distribution that is consistent with information we know the distribution must meet, the best strategy is to maximise its entropy subject to a series of constraints that encode the relevant information. When entropy is maximised, the distribution is the most conservative we can choose and hence the most 'uninformative'. 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. This maximisation is thus equivalent to choosing a distribution that is the most likely or probable within the constraints for it is easy to show as Wilson (1970, 2010) does, that the maximum entropy is an approximation to the probability of a particular macro state occurring amongst all possible arrangements (or microstates) of the events in question.

Unlike Wilson (1970, 2010), we will demonstrate the maximisation for a probability distribution of the location of population p_i in n zones rather than the probability p_{ij} of interactions between zones i and j although all our derivations are immediately generalisable to these more detailed specifications. We must first specify the constraints and we take these to be functions of the probabilities that define totals, averages or more generically 'moments' of the distribution. To demonstrate this, we choose two constraints on the location of population. First there is a normalisation constraint that ensures the probabilities sum to unity

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

and second we choose a constraint on the average cost $\overline{C}$ of locating in any zone which is the sum of the individual locational costs c_i

$$\sum_i p_i c_i = \overline{C} \quad . \tag{5}$$

We first form a Lagrangian L which consists of the entropy H reduced by the information encoded into the constraints in equations (4) and (5) and we then find its maximum with respect to the probability p_i. Then

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

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

$$\frac{\partial L}{\partial p_i} = -\log p_i - \lambda_0 - \lambda_1 c_i = 0 \quad , \qquad (7)$$

rearrangement and exponentiation of which gives the probability model

$$p_i = \exp(-\lambda_0 - \lambda_1 c_i) \quad . \qquad (8)$$

Note that the multiplier which is specified as $\lambda_0 - 1$, enables us to get rid of the free floating 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 it is easy to show that $\exp(-\lambda_0)$ is a partition function defined from

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

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

$$p_i = \frac{\exp(-\lambda c_i)}{\sum_i \exp(-\lambda c_i)} \quad , \quad \sum_i p_i = 1 \quad , \qquad (10)$$

and from this, we can see that if the Lagrangian multiplier on the average cost of location is redundant, that is $\lambda_1 = 0$, 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 from which it is clear that when the entropy for this model is at its maximum,

$$H_{\max} = -\sum_i p_i \log\left[\exp(-\lambda_o - \lambda_1 c_i)\right] = \lambda_0 + \lambda_1 \overline{C} \qquad . \qquad (11)$$

It is quite clear that 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. It is easy to show that for the exponential function, the summations in equations (4)-(6) and (9)-(11) can be generalised 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 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 assume that as $\Delta x_i \to 0$, then $p(x_i) \to p(x)$ and $c(x_i) \to c(x)$. We can thus write and simplify constraint equation (4) as

$$\begin{aligned} \sum_{\substack{i \\ \lim \Delta x_i \to 0}} p(x_i)\Delta x_i &= \int_0^\infty p(x)dx \\ &= \int_0^\infty \exp(-\lambda_o)\exp[-\lambda_1 c(x)]\, dx = \frac{\exp(-\lambda_o)}{\lambda_1} = 1 \end{aligned} \qquad , \qquad (12)$$

which further simplifies to

$$\exp(-\lambda_0) = \lambda_1 \qquad and \qquad \lambda_0 = -\log \lambda_1 \qquad . \qquad (13)$$

The constraint on travel cost in continuous form can now be written as

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

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 $\exp(-\lambda_0) = \lambda_1 = 1/\overline{C}$, the model can be written in its classic form as a density

$$p(x) = \frac{1}{\overline{C}} \exp\left(-\frac{c(x)}{\overline{C}} \right) \qquad , \qquad (15)$$

where in thermodynamics $c(x)$ is the energy at location x and $\overline{C}$ is the average temperature.

The maximum entropy in continuous form is not however the limit of equation (2) with respect to Δx_i as we will show below. Before we do so, let us state this entropy as

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

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

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

It is now clear that for the appropriate measurements of entropy S (and H), these vary with the log of the average cost or temperature and it is also clear that 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, then entropy itself can be see as such a measure. Related insights have been explored by Batty (1983), Erlander and Stewart (1990), and Roy and Thill (2004).

Our last foray into the derivation of this model – which we 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 the movement of costs of location between different places i. Let us assume that the system starts with each place i having the average cost of location as $\overline{C}$, that is, $c_i = \overline{C}, \forall i$. 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. We set up 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 so the total cost (and average cost) of location remains the same. At each time, $c_i(t+1) = c_i(t) + \Delta c$ and $c_j(t+1) = c_j(t) - \Delta c$ such that $\Sigma c_k(t+1) = \Sigma c_k(t)$. We also assume that a location cannot receive a negative cost, that there is a lower bound to $c_i(t) \geq 0, \forall i,t$ where this boundary condition is absolutely essential for the generation of the stable state which ultimately emerges. If this process continues for many time steps, what will emerge is a distribution of costs (in locations) which follows the Boltzmann-Gibbs distribution in equations (10) or (15) which appears when the costs are binned and the relative probability distribution examined. In short, what actually happens is that through a process of random swapping akin to energy collisions in a thermodynamic system, the system self-organises 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 leads to the universal form of a negative exponential which is due to the boundary condition and the conservation of costs. Strictly, the process is best considered as one where each location is an individual engaging in the process with the resulting probability distribution formed by collecting these individuals into 'locations'. Dragulescu and Yakovenko (2000) show many variants of the model which 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 generalise the model by relaxing the boundary constraints and embed it in a wider context where wealth which is not conserved is considered, making the point that these variants also admit the generation of other distributions such as the lognormal 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 leads to entropy-maximising. The dynamics that has been explored, is one in which the entropy-maximising solution is embedded in a wider non-linear dynamics (Wilson, 2010). What this discussion introduces is the possibility of disaggregating the entropy-maximising model to the point where individuals or agents are the basic objects comprising the system, thus opening the framework to much more general types and styles of simulation such as agent-based modelling.

Spatial Entropy: The Continuous Formulation

So far, apart from our brief digression in the last section into continuous entropy, we have not made any formal distinction between density and distribution. We have assumed 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 was the same size as for example in a spatial system arranged on a regular grid. In fact, many spatial models ignore the size of the interval completely and operational models which build on entropy-maximising rarely factor internal size into their simulations which inevitably leads to biased applications. Yet we can easily show how interval size must enter the analysis explicitly. As before, we first define each element of the probability distribution p_i which is the product of an approximation to the density $p(x_i)$ and the interval size Δx_i

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

from which density is defined as

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

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

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

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

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

where the first term on the RHS of equation (21) is the continuous Shannon entropy defined above as S in equation (17). It is clear from equation (21) that $H \to \infty$, as $\lim \Delta x_i \to 0$, which is another way of saying what we have already said in the previous section, that is, 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-maximising 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

$$H_S = -\sum_i p_i \log \left[\frac{p_i}{\Delta x_i} \right] \qquad , \qquad (22)$$

which we define as spatial entropy (Batty, 1974; Goldman, 1968). Using equation (22) instead of equation (2) in the entropy-maximising scheme which involves minimising 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)

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

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 $\Delta x_i \to 0$.

There is however another interpretation of this augmented model. If we write the entropy H_S in the expanded form of equation (22) as

$$\begin{aligned} H_S &= -\sum_i p_i \log p_i + \sum_i p_i \log \Delta x_i \\ &= H + \sum_i p_i \log \Delta x_i \end{aligned} \qquad , \qquad (24)$$

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. In equation (24), this is a very specific constraint 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

$$\sum_i p_i \log \Delta x_i = \log \overline{X} \qquad , \qquad (25)$$

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

$$L = -\sum_i p_i \log p_i - (\lambda_0 - 1)\left(\sum_i p_i - 1 \right) - \lambda_1 \left(\sum_i p_i c_i - \overline{C} \right) - \lambda_2 \left(\sum p_i \log \Delta x_i - \log \overline{X} \right) \quad . \qquad (26)$$

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

$$p_i = \exp(-\lambda_0 - \lambda_1 c_i - \lambda_2 \log \Delta x_i) \quad , \qquad (27)$$

which in more familiar form can be written as

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

The interval or zone size thus 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, 1971) in his family of spatial interaction models. It is also clear comparing equations (23) and (28) that if the multiplier λ_2 is forced to be unity, then the constraint on interval size enters the model in exactly the same way as if it were incorporated into the entropy in the first place, that is as a maximisation of spatial rather than discrete entropy. Note also that in entropy-maximised equations like (28) the sign of the multipliers is undetermined until they are fitted to meet the constraint equations. There is one further point on this augmented maximisation. If constraint equations in the Lagrangian or augmentations to the entropy are of logarithmic form, then the relevant variables enter the model as power laws, that is, they are scaling, and any continuous version of the derivation has to be modified to ensure that these constraints lie within defined limits. We will return to this point later when we deal more formally with scaling.

The standard example that Wilson (1970) uses to demonstrate the logic of entropy maximisation 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 which is subject to an equivalent cost and normalisation constraint is derived by maximising

$$H = -\sum_i \sum_j p_{ij} \log p_{ij} \quad , \qquad (29)$$

subject to the following constraints

$$\sum_i \sum_j p_{ij} = 1 \quad and \quad \sum_i \sum_j p_{ij} c_{ij} = \overline{C} \quad . \qquad (30)$$

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

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

The density equivalent is based on normalising the probability with respect to the size of the zones at each origin and destination Δx_i and Δx_j. 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

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

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 we move to deal with questions of scale and aggregation, we will reintroduce our 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 $\{q_i\}$ is often referred to as the prior and $\{p_i\}$ the posterior. Kullback (1959) and in a geographical context, Snickers and Weibull (1977) and Webber (1980), amongst others, define information I as

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

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

$$q_i = \frac{\Delta x_i}{\sum_o \Delta x_i} = \frac{\Delta x_i}{X} \quad , \tag{34}$$

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

$$I = \sum p_i \log\left[\frac{p_i}{\Delta x_i / X}\right] = \log X + \sum p_i \log \frac{p_i}{\Delta x_i} \qquad . \qquad (35)$$
$$= \log X - H_S$$

When $\Delta x_i = \Delta x, \forall i$, it is clear that equation (35) collapses to equation (3) which we repeat here as

$$I = H_{\max} - H = \log n + \sum_i p_i \log p_i \qquad . \qquad (36)$$

There are many such manipulations of entropy and information which all give oblique insights into the measure and the shape of the relevant distributions, some of which will recur in the subsequent discussion.

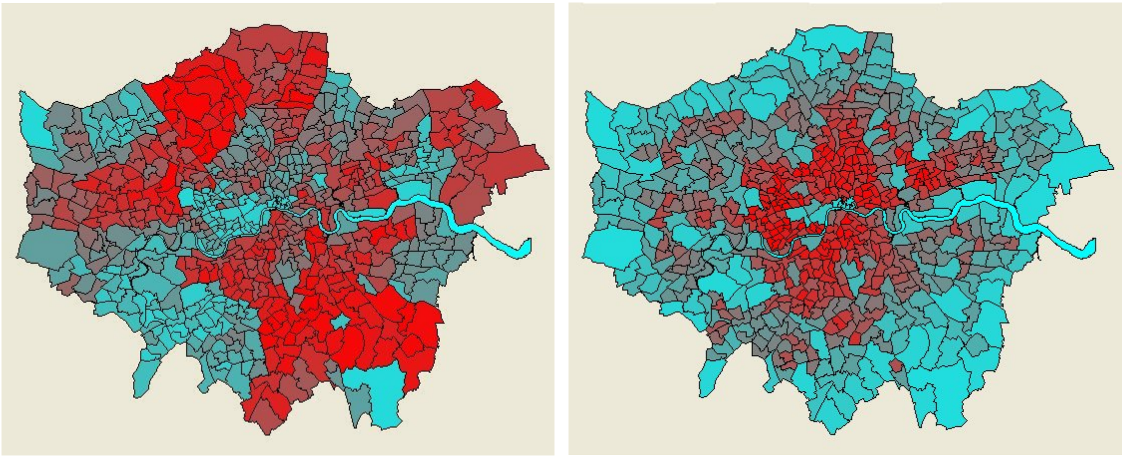
To conclude this section, it is worth noting how straightforward it is to get some idea of the magnitude of these different entropy measures and the way the value of the multipliers invoked through their maximisation can be approximated from data. We have taken population $\{P_i\}$, normalised with respect to total population P as $p_i = P_i / P$ and land area Δx_i of all the wards in Greater London $(n = 633)$ to measure the entropies H and H_S. We have defined the costs $\{c_i\}$ of location as a weighted measure of the transport costs $\{c_{ij}\}$ from any ward i to all others $j \neq i$, defined as $c_i = \sum_j p_j c_{ij}$ and $\overline{C} = \sum_i p_i c_i$. We have mapped the population $\{p_i\}$, the population density $\{p_i / \Delta x_i\}$, and these location costs $\{c_i\}$ in Figures 1(a) to 1(c). The various measures are shown in the table which is part of Figure 1 where we have also computed the two multipliers λ_0 and λ_1 first by solving the constraint equations (4) and (5), and then for the continuous multiplier λ_1, solving for the cost equation in equation (14) and the continuous entropy in equation (17). The differences between these values gives some measure of the extent to which the model fits the data and the extent to which the discrete and continuous models give equivalent results.

Scale and Entropy: Aggregation and Constraints

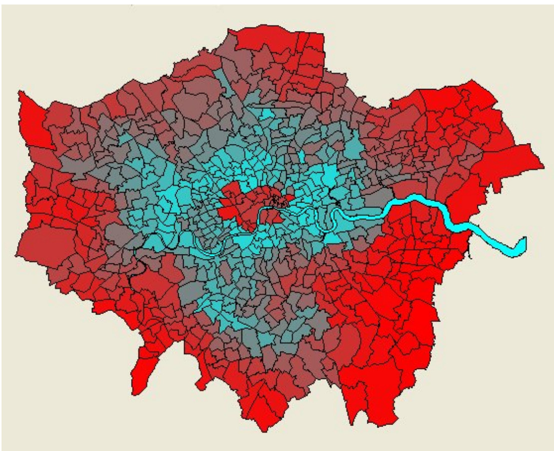
Shannon's entropy in equation (2) has an exceptionally easy-to-manipulate log-linear structure and additive form that allows it to be aggregated with respect to groups of objects that might pertain to some higher level of organisation in the system of interest. Theil (1972) refers to this process of aggregation as the entropy decomposition theorem and to illustrate it, we 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 exhaustive and exclusive in that

$$Z = \bigcup_{k=1}^{K} Z_k \qquad \textit{and} \qquad \Omega = \bigcap_{k=1}^{K} Z_k \qquad , \qquad (37)$$

a *b*



c



Statistic/Parameter	*Value*
Discrete Entropy H	2.7907
Spatial Entropy H_S	3.0927
λ_0 from equations (4)-(5)	2.3442
λ_1 from equations (4)-(5)	0.0078
λ_0 from $\overline{C}$	1.9966
λ_1 from $\overline{C}$	0.0081
λ_0 from H_S	2.0927
λ_1 from H_S	0.0101

Figure 1: Population (a), Population Density (b) and Location Costs (c) Used to Compute Entropies and Multipliers for the Boltzmann-Gibbs Model

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

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

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

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

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. In fact it is easy to show that 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 – that the within-set entropies decrease in sum and the between-set entropy H_B rises in value until all there is one aggregated set for each object, that is $H_B \to H$. Moving the other way, when all the objects are aggregated into one set then, $H_B \to 0$, and $\Sigma P_k H_k \to H$. Proofs of these assertions are given in Theil (1972) and Webber (1979).

It is a simple matter to state the equivalent decomposition formula for spatial entropy as we have defined it in equation (22). Then noting that

$$X_k = \sum_{i \in Z_k} \Delta x_i \qquad , \qquad (40)$$

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

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

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. There is an information difference structure buried in equation (41) as we spelt out earlier for spatial entropy between equations (33) and (36) and similar interpretations apply. In fact, in developing decompositions of entropy and spatial entropy in this fashion, the focus has been 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 which are smaller than Z_k but larger that 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 the effort to design spatial systems that meet some criteria of optimality that pertain to scale and size (see Batty, 1974, 1976; Batty and Sammons, 1978). In this paper, we will not deal with the effect of shape on entropy but there are extensions to deal with idealised spatial systems that also incorporate constraints on shape such as the regularity of boundaries although developments in this area have been limited (Batty, 1972, 1974).

These decomposed entropy measure can be used in entropy maximisation enabling 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 maximised so that the aggregate probabilities sum to those which are fixed by the level of decomposition or aggregation chosen, as fixed in equation (38) above. We set up the Lagrangian to maximise equation (39) with respect to equations (38) and (5) as follows

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

and then minimise this as

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

to derive the model that we can state as

$$p_i = \exp(-\lambda_0^k - \lambda_1 c_i), \quad i \in Z_k \quad . \quad (44)$$

We can compute the partition function directly by substituting for p_i in equation (38) and this leads to

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

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

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

Note that the constraint equation on cost is for the entire system and this effectively couples together 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 for if there are no system-wide constraints, then the models are separable; in fact the entropy-maximising 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

$$\sum_{i \in Z_k} \frac{p_i}{P_k} c_i = \overline{C}_k \ , \quad \forall k \quad . \quad (47)$$

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

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

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

$$p_i = P_k \frac{\exp(-\lambda_1^k c_i)}{\sum_{i\in Z_k} \exp(-\lambda_1^k c_i)} \quad , \quad i \in Z_k \quad . \tag{49}$$

This model is not only separable for each subset Z_k but each model is also calibrated separately with respect to the cost constraint and determination of the multipliers $\{\lambda_1^k\}$. Using spatial entropy-maximising 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

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

where if the system-wide cost constraint in equation (5) applies, 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 non-spatial systems, connecting these ideas to a much wider literature in the measurement of inequality.

Generating Spatial Probability Distributions

So far we have defined both entropy and its method of maximisation with respect to probabilities that pertain to spatial locations. In terms of the typical problem, then we are assuming that the probability of location is some function, in the classic Boltzmann-Gibbs case a negative exponential, 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 it would make sense to assume a higher probability of location measured by a higher population would be associated with a lower cost (or higher benefit) of locating in the place in question. However, there is another interpretation which is less specific with respect to the kinds of probability distribution that emerge from entropy-maximising and this depends on how one sets up the problem. In this section and the rest of this paper, we will assume that it is some measure of size, not cost, that the probability distribution must conserve and probabilities will 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 locating is with respect to some size or cost, we will now develop the model as one in which the probability of location is dependent on the actual population size which is observed in the locations in question. As we will see, this is the obvious way in which to develop entropy-maximising for city-size distributions, an area 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 up the arguments of this paper to size distributions in general and power laws in particular.

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

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

This will be subject to the usual normalisation and constraints associated with the moments of the distribution which we define as

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

This discussion and notation follows Tribus (1969) although there are several other presentations of this process which have more formal roots in probability theory. As befits a paper such as this, the presentation is informal.

The Boltzmann-Gibbs negative exponential model is still the baseline in entropy maximisation since it introduces a constraint on the distribution which is the first moment, the average and no others except for the normalisation of the probabilities. Then following the same logic as we used earlier in equations (4) to (10) and noting that we can assume the intervals over which the discrete frequency is measured to be equal, that is $\Delta x_i = \Delta x, \forall i$ (to avoid any confusion with spatial entropy at this stage), we maximise equation (51) subject to the first two constraints shown in (52). Using the relevant Lagrangian with appropriate multipliers leads to

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

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

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

which gives the familiar exponential form

$$f(V_i) = \frac{\exp(-\lambda_1 V_i)}{\sum_i \exp(-\lambda_1 V_i)} \quad . \qquad (55)$$

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

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

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 for it is much more likely that larger populations live in places where the costs of location are lower, all other things being equal. This is the confusion that has never really been resolved in generating size distributions from entropy-maximising. The motivation for the earlier models such as those developed by Wilson (1970) was always to maximise entropy with respect to a cost constraint whereas the models of this section are to maximise entropy with respect to a size constraint. In this context, it is perfectly reasonable to assume that if an individual were to locate across a space, that there are many more places to locate where populations are smaller than places where populations are large. It is in this sense, that frequency in this section is different from probability in the previous sections, although formally the algebraic expressions are identical.

We can now show very simply how the negative exponential can become a power function if the constraint on average size is replaced by its geometric equivalent, that is

$$\sum_i f(V_i) \log V_i = \log \overline{V} \quad , \qquad (57)$$

where $\log \overline{V}$ is the expected value of the sum of the logarithms of the sizes. We are assuming that that this is defined for a discrete system for there are difficulties which we note below when we examine the rank-size rule and its consistency with entropy maximisation where the continuous version of the model must be invoked for purposes of simplification and demonstration. However if we maximise entropy subject to this and the normalisation constraint, the model becomes

$$\log f(V_i) = -\lambda_0 - \lambda_1 \log V_i \quad , \qquad (58)$$

which in exponential form is

$$f(V_i) = \exp(-\lambda_0)\, V_i^{-\lambda_1} \quad . \qquad (59)$$

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

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

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

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

In this context, V_i also varies inversely with the power of frequency. From equation (61), we can generate the more familiar rank-size rule which has been known for well over a century, first exploited for income sizes by Pareto (1906) and then for city sizes by Zipf (1949). We will explore these functions in the next section.

If we now maximise entropy with respect to the three constraints stated in equations (52), noting that the third constraint can be also simplified to

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

then this leads to

$$\log f(V_i) = -\lambda_0 - \lambda_1 V_i - \lambda_2 V_i^2 \qquad . \qquad (63)$$

In the first exponential form, this is

$$f(V_i) = \exp(-\lambda_0 - \lambda_1 V_i - \lambda_1 V_i^2) \qquad , \qquad (64)$$

which in the more familiar terms is

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

As Tribus (1969) shows, equation (65) is a form of the normal distribution. What is interesting about the entropy-maximising derivation is that 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. It can be shown that 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 of course is always preserved no matter what the value of these multipliers. Moreover if $\lambda_1 << \lambda_2$ then this is tantamount to the variance of the distribution getting smaller and smaller with the normality more and more peaked. We can now complete this set of distributions by assuming that the size distribution is lognormal, that is, that instead of V_i, we now define size as its logarithm $\log V_i$. We will formally restate the constraint equations for the lognormal as

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

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

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

From equation (67), it is now very clear that if $\lambda_1 << \lambda_2$, then the lognormal form collapses to the inverse power law form but only for a range of the largest values of V_i. This in fact is one of the simplest demonstrations that power laws tend to dominate in the upper or heavy tail of the lognormal distribution. Again the same caveats as to the existence of the moments for the discrete case apply, which for the sorts of spatial system to which these models apply, this will always be the case, that is where $1 \le V_i < \infty$. Tribus (1969) has a relatively straightforward demonstration of the properties of the normal distribution with respect to the values of the parameters which 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 that have been generated using entropy-maximising in the previous section represent discrete density functions relating frequency to size. In fact 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 we now call rank r_1 to the smallest which we call rank r_n. In fact, the rank is the counter-cumulative 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 $f(V) \sim \exp(-\lambda_1 V)$. Then the integration defining the counter-cumulative $F(V)$ is

$$F(V) = \int_v^\infty f(V)\, dV \sim \int_v^\infty \exp(-\lambda_1 V)\, dV = \frac{1}{\lambda_1}\left[\exp(-\lambda_1 V)\right]_v^\infty , \qquad (68)$$

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

$$r_k \sim \exp(-\lambda_1 V_k) \qquad , \qquad (69)$$

from which

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

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.

In fact, the classic rank-size relationship is normally developed for 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 counter-cumulative $F(V)$ as

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

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

$$r_k \sim V_k^{-\lambda_1 + 1} \qquad , \qquad (72)$$

from which

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

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

It is worth noting that 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 seen as a regular distortion based on human perception. This is known as the Weber-Fechner-Stevens law which pertains to the fact that increases in how we perceive brightness and sound, even the way our cognitive senses respond to size, are proportional to the logarithm, not the actual value, of the relevant measure of intensity (see Stevens, 1957). In spatial interaction modelling, Wilson (1970, 1971) made use of this to show how the original gravitational hypothesis is consistent with models produced by entropy-maximising, 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. Of course the same arguments are used to incorporate additional constraints which might be thought of as benefits rather than costs, reflecting the fact that agglomeration economies are sometimes perceived logarithmically.

We can also generate rank-size distributions for the normal and lognormal models that we derived in equations (65) and (67) respectively. In fact, there is little point in pursuing this for the normal but the lognormal is a special case largely because there are many arguments that suggest that city, firm and income size distributions are not consistent with power laws but in fact are lognormal where the power law only applies 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, this becomes $f(V) \sim V^{\alpha} V^{-2\beta}$ from which we form the counter-cumulative as

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

From equation (74), it is clear that the shape of the lognormal is completely dependent on the value of the parameters α and β but 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 in analogy to equation (73) can be written as

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

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

This is a somewhat informal way of demonstrating the relationship between inverse power and lognormal functions and readers are referred to more considered sources which elaborate this relationship more strictly. Perline (2005) has an excellent discussion of when one is able to approximate the heavy tail of a lognormal with a power law which builds on earlier expositions that are part of the literature on skewed probability functions as summarised by Montroll and Schlesinger (1982, 1983). It is not our purpose here to develop a treatise on the lognormal or indeed on Zipf and Pareto power laws for we have shown that both can be derived from entropy-maximising but it is important to note that power laws can emerge from two sources: first directly if the constraint on the entropy is a geometric mean and second when the constraints on the entropy are those that define the lognormal 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. If we examine the data we used earlier for Greater London, the range of values for the distribution of population densities is much too narrow to see any evidence of an inverse power function in the upper tail. In such an intra-urban context, the variation between the largest and smallest values it not enough as populations in the centre are very small due to competition from other activities. Readers are referred to the mainstream literature on city-size distributions where these issues are discussed in more detail. The recent paper by Eeckhout (2004) is representative.

There is one last substantive issue that we need to discuss to complete our presentation on how scaling distributions are associated with entropy-maximising. The traditional explanation of how power laws come to dominate spatial and social distributions is essentially based on a generic model that leads to agglomeration economies in which it is becomes increasingly unlikely that any object chosen at random grows to a very large scale, realising 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 comprising such competitive systems is based on Gibrat's (1931) law of proportionate effect in which any object grows or declines by a random amount whose value is proportionate to the size of the object already reached. This process, if operated continually for many time periods, leads to a distribution of objects which is lognormal. 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), then several authors have shown that the resultant distribution is no longer lognormal but 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 (Gabaix, 1999; Saichev, Malevergne, and Sornette, 2010), in earth sciences (Sornette, 2006) as well as in several other areas of social inquiry (Blank and Solomon, 2000; Newman, 2005).

In fact, this dynamics which is referred to by Solomon (2000) as the 'Generalised 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 variable of interest, be it population, income, wealth, cost and so on. The steady state results generated by such models are also consistent with Boltzmann distributions as Richmond and Solomon (2001) show, while Foley (1994) and then Milakovic (2003) show that entropy-maximising can be employed directly with the dynamics being embedded as constraint equations which the process of wealth creation must meet. There is now an enormous literature dealing with stochastic GLV types of model which build on proportionate effect leading to lognormal and power laws. Several oblique interpretations of the steady states associated with such processes as

Boltzmann-Gibbs distributions exist, as Richmond and Solomon (2001) say as " … Boltzmann laws in disguise". The earlier dynamic models developed by Dragulescu and Yakovenko (2000) are also 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 has yet to find its way into spatial or geographical systems for the concern with city-size distributions has been remarkably aspatial in contrast to entropy-maximising in geographical analysis but there are signs of a convergence. Wilson's (2008) recent work, for example, is seeking to generalise entropy-maximising in a dynamical framework that he refers to as Boltzmann-Lotka-Volterra (BLV) models and these have clear links to GLV models. At present approaches to dynamics can be seen as either constructing a Lotka-Volterra dynamics which leads to Boltzmann-Gibbs and related distributions or as Boltzmann-Gibbs distributions which are nested within Lotka-Volterra dynamics. There is much synthesis to be done and there appear many fruitful insights to be gained by these extensions. After a period of reflection and consolidation, it is now entirely possible that there will be a rebirth of interest in measures of entropy and entropy-maximising in geographical analysis.

Future Research: Alternative Entropies, More Explicit Dynamics

Earlier in this paper, we argued that one of the things 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 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 fact in geographical systems, events can be highly auto-correlated in space as well as time and thus the methods used to generate probability distributions in equilibrium or the in the steady state can be badly compromised if more appropriate measures are not chosen.

Amongst these, the measure proposed by Reyni (1961) introduces a parameter α which gives greater weight to larger probabilities if this parameter is greater than 1, lesser weight if less than 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. There are few if any applications in this field (but see March and Batty, 1975) and thus this measure is worth exploring further. A more radical form of measure broaches the question of the independence of events directly 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 non-extensive, 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 Reyni entropy) is that in its use in maximisation, the resulting model is an inverse power law, not the negative exponential. All 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-maximising framework that he introduced is in the generation of specific and applied models while at the same time demonstrating 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 which continue to be useful, indeed essential in consistently specifying and coupling different models together. The development of entropy-maximising in generating economic models came much later and has yet to adopt the systematic procedures demonstrated for spatial systems by Wilson (2010). Yet despite the power of entropy-maximising, it is compromised a little in that it is easy to show that space itself should be directly incorporated into the framework so that dimensional consistency can be ensured through use of spatial entropy rather then it discrete equivalent, the Shannon measure. Existing practice of defining operational location and interaction models has not really followed these procedures, nor has it systemically examined the sets of constraints that are needed to define particular problems with respect to what is known and not known about the systems of interest. There is still much to do with respect to using entropy-maximising to establish formalised 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 that it could generate models in equilibrium. Dynamics was assumed to be benign, even to the point where simple models such as that used to move money around in an economic system developed only a decade or so ago by researchers such as Drăgulescu and Yakovenko (2000), have never been explored in spatial analysis. Seeing geographical systems in equilibrium was enough. When researchers such as Wilson (2008) 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 we argued in the last section, there is now a new momentum 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. But there is now 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.

Literature Cited

Adamic, L. A. (2002). Zipf, Power-laws, and Pareto – A Ranking Tutorial. Information Dynamics Lab, HP Labs, Palo Alto, CA, accessed 04/01/2010, available at http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html.

Batty, M. (1972). "Entropy and Spatial Geometry." **Area, 4,** 230-236.

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**, 256-267.

Batty, M. and Sammons, R. (1978). "On Searching for the Most Informative Spatial Pattern." **Environment and Planning A, 10,** 747-779.

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

Blank, A., and Solomon, S. (2000). "Power Laws in Cities Population, Financial Markets and Internet Sites: Scaling and Systems with a Variable Number of Components." **Physica A**, **287**, 279–288.

Curry, L. (1964). "The Random Spatial Economy: An Exploration in Settlement Theory." **Annals of the Association of American Geographers**, **54**, 138-146.

Drăgulescu, A. A., and Yakovenko, V. M. (2000). "Statistical Mechanics of Money." **The European Physical Journal B**, **17**, 723–729.

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

Erlander, S., and Stewart, N. F. (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), 321–345.

Gabaix, X. (1999). "Zipf's Law for Cities: An Explanation." **Quarterly Journal of Economics**, **114**, 739–767.

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

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

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

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

March, L., and Batty, M. (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**, Pisa, Italy: Laboratory of Economics and Management (LEM), Sant’Anna School of Advanced Studies.

Montroll, E. W. and Schlesinger, M. F. (1982) “On 1/f Noise and Other Distributions with Long Tails.” **Proceedings of the National Academy of Sciences USA**, **79**, 3380-3383.

Montroll, E. W. and Schlesinger, M. F. (1983). Maximum Entropy Formalism, Fractals, Scaling and l/f Noise: A Tale of Tails.” **Journal of Statistical Physics**, **32**, 209-230.

Newman, M. E. J. (2005). “Power Laws, Pareto Distributions and Zipf's Law.” **Contemporary Physics**, **46**, 323-351.

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 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960**, 547–561. at http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf

Richmond, R. and Solomon, S. (2001). “Power Laws are Disguised Boltzmann Laws.” **International Journal of Modern Physics C**, **12** (3), 333-343.

Roy, J. R. and Thill, J. C. (2004). “Spatial Interaction Modeling.” **Papers in Regional Science**, **83**, 339 – 361.

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

Shannon, C. E. (1948). “A Mathematical Theory of Communication.” **Bell System Technical Journal**, **27**, 379–423, and 623–656.

Snickers, F. and Weibull, J. W. (1977). “A Minimum Information Principle: Theory and Practice.” **Regional Science and Urban Economics**, **7**, 137-168.

Solomon, S. (2000). “Generalized Lotka Volterra (GLV) Models of Stock Markets.” in G. Ballot and G. Weisbuch (Editors). **Applications of Simulation to Social Sciences**, Oxford, UK: Hermes Science Publications, 301-322.

Sornette, D. (2006). **Critical Phenomena in Natural Sciences**. Second Edition, Heidelberg, DE: Springer.

Stevens, S. S. (1957). "On the Psychophysical Law." **Psychological Review**, **64**(3), 153–181.

Theil, H. (1972). **Statistical Decomposition Analysis**. Amsterdam, Netherlands: 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 Gell-Mann, M., and Tsallis, C. (Editors) **Nonextensive Entropy: Interdisciplinary Applications**. Oxford, UK: Oxford University Press, 1-53.

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. (1971). "A Family of Spatial Interaction Models, and Associated Developments." **Environment and Planning**, **3**(1), 1–32.

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**, 865-871.

Wilson, A. G. (2010). "Entropy in Urban and Regional Modelling: Retrospect and Prospect." **Geographical Analysis**, forthcoming.

Yakovenko, M. and Rosser Jr., J. B. (2009). "Colloquium: Statistical Mechanics of Money, Wealth, and Income." **Reviews of Modern Physics**, **81**, 1703-1725.

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