Lecture 7: 31st October 2011

Entropy, Complexity, and Information:

Michael Batty

m.batty@ucl.ac.uk

twitter @jmichaelbatty

http://www.complexcity.info/
http://www.spatialcomplexity.info/

Outline of Lecture 7

Defining Entropy: Probability (Population) Densities

Interpreting Entropy

Entropy Maximising: Deriving the Density Model

Information and Entropy

Spatial Entropy – Size and Shape and Distribution

Spatial Entropy as Complexity

Symmetry again – I will leave you to read this section from the post of the pdf

Defining Entropy: Probability (Population) Densities

Last time we introduced spatial interaction for two dimensional systems – for networks defined as flows between origins and destinations i and j.

We need to simplify this so we can introduce our definitions and interpretations of 'entropy' as clearly as possible.

We first define the probability as the proportion of the population in i but we could take any attribute – we use population because it is an easy to understand attribute of a geographical system. We thus define the probability p_i as

$$p_i = \frac{P_i}{P}$$

The population P_i sums to P as

$$P = \sum_i P_i$$

And this means that the probabilities will sum to 1

$$\sum_i p_i = 1$$

Now let us define raw information in terms of p_i

$$\frac{1}{p_i}$$

Note that when the probability is small the information is large and vice versa. I.E. high info occurs when the event is unlikely and we get a lot of info if and when it occurs

But if an event occurs and another event occurs which is independent, then the joint info should be

$$\frac{1}{p_i p_j} = \frac{1}{p_i} . \frac{1}{p_j}$$

Now information gained should in fact be additive, we should be able to add the first info + the second info to get this but

$$\frac{1}{p_i p_j} \neq \frac{1}{p_i} + \frac{1}{p_j}$$

The only function which will allow this is the log of

$$\log \frac{1}{p_i}$$

And we thus write the information as follows

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

And if we take the average or expected value of all these probabilities in the set, we multiply the info by the probability of each and sum

to get for n events

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

This is the entropy. The minimum value of this function is clearly 0 which occurs when

$$p_i = 1, \quad \textit{and the rest are } p_j = 0, \forall j \neq i$$

And we can easily find out that the entropy is at a maximum when the probabilities are all equal and H=log n when

$$p_i = \frac{1}{n}$$

Interpreting Entropy

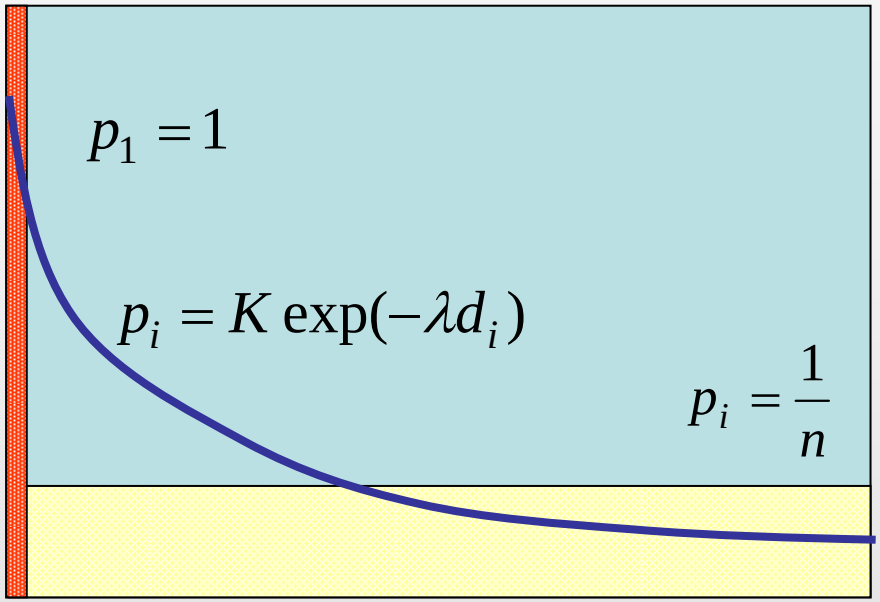
Entropy has two aspects that are relevant to complexity. These are based on the distribution and on the number of the events – i.e. the size of the system in terms of the number of objects or populations n

This immediately means there is a tradeoff between the shape of the distribution and the number of events. In general, as the number of events goes up – i.e. n gets bigger – then the entropy H gets larger.

But the shape of the distribution also makes a difference.

Let us imagine that we are looking at the population density profile from the centre for the edge of a city. This is a one-dimensional distribution.

We can graph this as follows:



We can calculate H as minimum 0 where p_1=1, p_i=0, i>1

H maximum = log n for the uniform distribution

For the negative exponential $H = \log K + \lambda \sum_i p_i d_i$

Basically what all this implies is that when we have an extreme distribution, the entropy or information is zero. This means that if the probability is 1, and the event occurs, then the information we get is zero.

In the case where the probability is the same everywhere, if an event occurs then the information is at a maximum.

Now also as the number of events goes up, we get more information.

There is thus a tradeoff. We can have any system with entropy from zero to log n. But as n goes up, then we can have a system with very few n and greater entropy than a system with many n but with an extreme distribution. This entropy measures distribution as well as number; distribution is a little like shape as the previous graphs show.

Entropy Maximising: Deriving the Density Model

Essentially in E-M, we choose a probability distribution so that we let there be as much uncertainty as possible subject to what information we know which is certain

This is not the easiest point to grasp – why would we want to maximise this kind of uncertainty – well because if we didn't we would be assuming more than we knew – if we know there is some more info, then we put it in as constraints. If we know p=1, we say so in the constraints. Let us review the process,

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

Subject to $\sum_i p_i = 1$ and $\sum_i p_i c_i = \overline{C}$

We can think of this as a one dimensional probablity density model where this might be population density

And we then get the classic negative exponential density function which can be written as

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

Now we don't know that this is a negative function, it might be positive – it depends on how we set up the problem but in working out probabilities wrt to costs, it implies the higher the cost, the lower the probability of location.

We can now show how we get a power law simply by using a log constraint on travel cost instead of the linear constraint.

We thus maximise entropy subject to a normalisation constraint on probabilities and now a logarithmic cost constraint of the form

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

Subject to $\sum_i p_i = 1$ and $\sum_i p_i \log c_i = \overline{C}$

Note the meaning of the log cost constraint. This is viewed as the fact that travellers perceive costs logarithmically according the Weber-Fechner law and in some circumstances, this is as it should be.

If we do all this we get the following model where we could simply put $\log c_i$ into the negative exponential getting

$$p_i = \frac{\exp(-\lambda \log c_i)}{\sum_i \exp(-\lambda \log c_i)} \quad \Rightarrow \quad p_i = \frac{c_i^{-\lambda}}{\sum_i c_i^{-\lambda}}$$

A power law. But this is not the rank size relation as in the sort of scaling we looked at last week. We will see if we can get such a relation below but first let me give one reference at this point to my GA 2010 paper

Space, Scale, and Scaling in Entropy Maximizing, *Geographical Analysis 42* (2010) 395–421 which is at http://www.complexcity.info/files/2011/06/batty-ga-2010.pdf

Before I look at the rank size derivation, let me show you a simple model of how we can generate an entropy maximising distribution which is negative exponential. We assume that we start with a random distribution of probabilities which in fact we can assume are resources – i.e. money

Now assume each zone has $c_i(t)$ units and two of these chosen at random engage in swapping a small unit of their resource – say one unit of money in each time period. In short at each time period, two zones i and j are chosen randomly and then one of them gives one unit of resource to another, again determined randomly; then $c_i(t+1)=c_i(t)-1$ and $c_j(t+1)=c_j(t)+1$.

In this way, the total resources are conserved i.e. $\sum_i c_i(t) = C$

Now this is like a process of random collisions. In much the same way that we showed how networks generate large hubs through preferential attachment, and the way cities get bigger or smaller through random growth through proportionate effect, then population units gain or lose in the same kind of competitive fashion

This leads to a negative exponential distribution

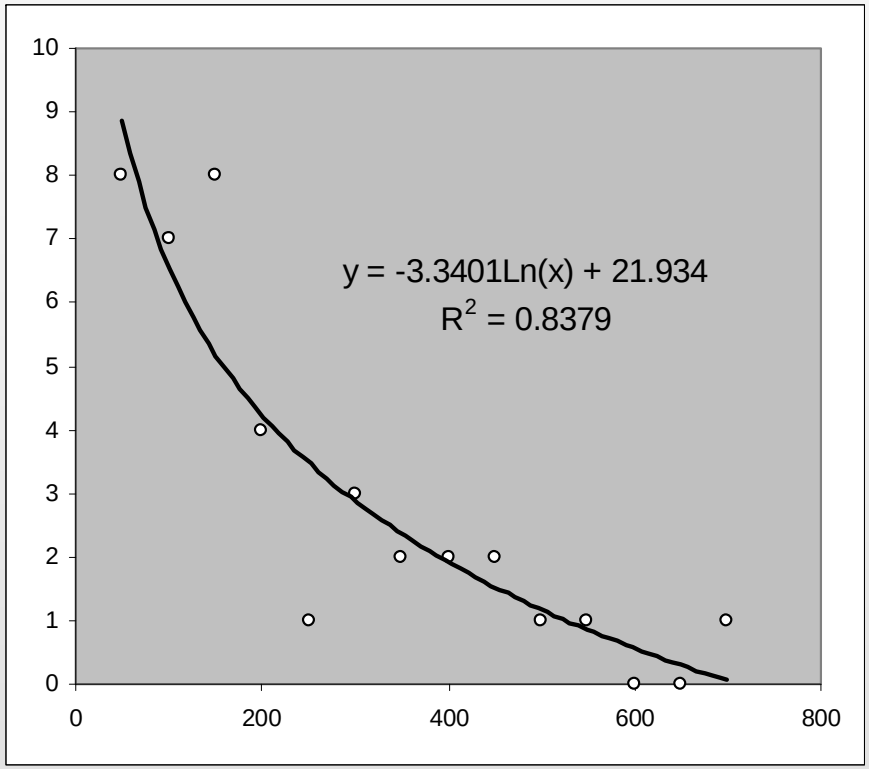
It is kind of obvious but we need to demo it and the following program shows how this occurs:

The Random Collisions Model

Note this model starts with something different from a uniform distribution – an extreme distribution with H=0 and then entropy increases as the collisions move the money around

If we start with an extreme distribution with H=log n, then the entropy reduces to that of the negative exponential

```
Private Sub Command1_Click()
Dim People(100) As Single
Money = 100
SwapMoney = 1
n = 40
For i = 1 To n
People(i) = Money
'Print i, People(i)
Next i
t = 1
For i = t To 1000000
ii = Int((Rnd(1) * n) + 1)
jj = Int((Rnd(1) * n) + 1)
If ii = jj Then GoTo 777
If People(ii) = 0 Then People(ii) = 1: GoTo 777
If People(jj) = 0 Then People(jj) = 1: GoTo 777
d = Rnd(1)
If d > 0.5 Then
fid = SwapMoney
fjd = -fid
End If
People(ii) = People(ii) + fid
People(jj) = People(jj) + fjd
Total = 0
For iz = 1 To n
Total = Total + People(iz)
Next iz
'Print ii, jj, fid, fjd, People(ii), People(jj), Total
777 Next i
For i = 1 To n
Print i, People(i)
Next i
NewFile = "Money.txt"
Open NewFile For Output As #2
For i = 1 To n
Print #2, i, People(i)
Next i
Close #2
End Sub
```



This is my own program which gradually converges on an exponential as the graph shows after 1 million runs

Our last foray into generating power laws using EM involves showing how we can get a rank size distribution.

There is a key difference between entropy-maximising location models which tend to look at location probabilities as functions of cost and benefit of the locations, and scaling models of city size or firm size or income size which tend to look at probabilities of sizes which have nothing to do with costs

Thus the problems of generating a location model or a size model are quite different.

Thus we must maximise entropy with respect to average city size not average locational cost and then we get the probabilities of locating in small cities much higher than in large cities as city size is like cost.

It is entirely possible of course for probabilities of locating in big cities to be higher than in small cities but as there are so many more small cities than big cities, small ones dominate.

So we to look at the city size problem, we must substitute cost with size and we thus set up the problem as

$$\max H = -\sum_i p_i \log p_i \quad st \quad \sum_i p_i = 1 \quad and \quad \sum_i p_i \log P_i = \overline{P}$$

$$p_i = \frac{\exp(-\lambda \log P_i)}{\sum_i \exp(-\lambda \log P_i)} \quad \Rightarrow \quad p_i = \frac{P_i^{-\lambda}}{\sum_i P_i^{-\lambda}}$$

And then we take the frequency as p_i and then the size as P_i, form the counter cumulative which is the rank and then twist the equation round to get the rank size rule – and hey presto we can connect up with our argument of a previous lecture

Information and Entropy

There is another measure of information which is important in spatial analysis and that is the information difference. Imagine we have prior and posterior probability distributions

$$q_i \quad where \quad \sum_i q_i = 1$$

$$p_i \quad where \quad \sum_i p_i = 1$$

We could form the entropy for each and make comparisons but there is an integrated formula based on the entropy of each with respect to the posterior probabilities only, that is

$$H(p:q) = -\sum_i p_i \log q_i \qquad H(p) = -\sum_i p_i \log p_i$$

$$I(p:q) = H(p:q) - H(p) = \sum_i p_i \log \frac{p_i}{q_i}$$

This is the Kullback information difference formula and it is always positive from the way we have formed it

In fact what we might do is not maximise this information difference but minimise it and we can set up the problem as one where we

$$\min I = \sum_i p_i \log \frac{p_i}{q_i} \quad st \quad \sum_i p_i = 1 \quad and \quad \sum_i p_i c_i = \overline{C}$$

This then leads to a model in which the prior probability appears in the model as one which is moderated by the additional information on cost, that is

$$p_i = \frac{q_i \exp(-\lambda c_i)}{\sum_i q_i \exp(-\lambda c_i)}$$

In fact if we then set $q_i=1/n$, that is, the uniform distribution, then this prior probability has no effect and the model simplifies to the usual EM model

As a parting shot on this, consider what happens when the prior probability is equal to the space available for population, that is $q_i \sim \Delta x_i$

Then our model becomes

$$p_i = \frac{\Delta x_i \exp(-\lambda c_i)}{\sum_i \Delta x_i \exp(-\lambda c_i)} \quad and\ thus \quad \rho_i = \frac{p_i}{\Delta x_i}$$

Note that this density can in fact be derived rather differently by developing a spatial version of entropy S and this we will now do. It is in fact equivalent formally to I

Spatial Entropy – Size and Shape and Distribution

Imagine that we now want to find the entropy of the probability density which is

$$\rho_i = \frac{p_i}{\Delta x_i}$$

We can simply take the expected value of the log of the inverse of this, that is the expected value of

$$\log \frac{1}{\rho_i} = -\log \rho_i$$

So the spatial entropy formula becomes

$$S = -\sum_i p_i \log \rho_i = -\sum_i p_i \log \frac{p_i}{\Delta x_i}$$

If we follow through the logic of EM then we get the same model as the one we have just shown but this time by maximising S, not minimising I

Now what we are doing here is using a rather different equation – spatial entropy is really entropy with an additional component

Let us expand it as

$$S = -\sum_i p_i \log \rho_i = -\sum_i p_i \log \frac{p_i}{\Delta x_i}$$

$$= -\sum_i p_i \log p_i + \sum_i p_i \log \Delta x_i$$

This is the distribution and the number size effect in terms of n in entropy

This is the area size effect

In fact this spatial entropy is really only the distribution effect for the number size effect is cancelled out – i.e. the second term cancels the number effect but in a convoluted way

In fact the spatial entropy is really just the discrete approximation to the continuous entropy which deals only with the distribution/density not simply the size effect

We can write spatial entropy thus or entropy as

$$S = H + \sum_i p_i \log \Delta x_i \qquad H = S - \sum_i p_i \log \Delta x_i$$

Now here we have an excellent definition of spatial complexity because we have in entropy both a size and distribution effect.

Note that the continuous equivalent of S is

$$S = -\int_x \rho(x) \log \rho(x)$$

By introducing spatial entropy, we get at both distribution and number-area size effects and are able to disaggregate this.

Spatial Entropy as Complexity

What we can now do is examine how entropy as complexity changes under different assumptions of the distribution and the size.

First let us note what happens when the probability is uniform, that is

$$p_i = \frac{1}{n}$$

$$S = \log n + \sum_i \frac{\log \Delta x_i}{n}$$

Then if we also have a uniform distribution of land

$$\Delta x_i = \frac{X}{n} = \sum_i \Delta x_i / n$$

Then we get S as

$$S = \log n + \sum_i \frac{1}{n} \log X + \sum_i \frac{1}{n} \log \frac{1}{n}$$

$$= \log X$$

We could of course maximise S and then we can easily see this.

We thus have different ways of computing the components of size and distribution and making comparisons of the shape of the distribution – what entropy comes from this – and the size of the distribution – what entropy comes from that

Moreover we can also employ extensive spatial disaggregation of these log linear measures. And I refer you back to the entropy paper in GA last year

As a conclusion, let us return to our spatial interaction model and look at its entropy – this is now

$$H = -\sum_{ij} p_{ij} \log p_{ij}$$

And there are various versions of spatial entropy

$$S = -\sum_{ij} p_{ij} \log \frac{p_{ij}}{\Delta x_{ij}}$$

$$S = -\sum p_{ij} \log \frac{p_{ij}}{\Delta x_i \Delta x_j}$$

These can be expanded but they are quite different: the first assumes that the space is an ij term whereas the second assumes space is at i and j separately – that interaction is not a space but that space is a location. This is not just a play on words – the ij space could be the space of the network

Symmetry again, and to conclude

What I want to do here is take a basic pattern of spatial interaction and then following a paper by Tobler decompose it into symmetric and skew symmetric components. The derivation follows: we are using distance not cost again now- a good model when we add the two asymmetries would be twice the symmetric gravity model

$$T_{ij} + T_{ji} = 2GP_iP_jd_{ij}^{-\alpha}$$

As the model is symmetric we write this as

$$\hat{T}_{ij} = \hat{T}_{ji} = [T_{ij} + T_{ji}]/2$$

$$\hat{T}_{ij} = \frac{T_{ij} + T_{ji}}{2} = GP_iP_jd_{ij}^{-\alpha}$$

and produce an average of the asymmetries

To model this asymmetry, we introduce a global term r and a directional term c_{ij} to the distance and our model becomes

$$T_{ij} = K \frac{P_i P_j}{d_{ij}^{\alpha}} (r + c_{ij})$$

$$T_{ji} = K \frac{P_i P_j}{d_{ij}^{\alpha}} (r - c_{ij}) = K \frac{P_j P_i}{d_{ji}^{\alpha}} (r + c_{ji})$$

$$\frac{T_{ij}}{T_{ji}} = \frac{r + c_{ij}}{r - c_{ij}}$$

From which we compute the term c_{ij} as

$$c_{ij} = r \frac{T_{ij} - T_{ji}}{T_{ij} + T_{ji}}$$

Clearly the model is symmetric so we can do this

The rest of the derivation is as follows and then we can see the model is divided into symmetric and asymmetric components

$$T_{ij} = K \frac{P_i P_j}{d_{ij}^{\alpha}} \left[r + r \frac{(T_{ij} - T_{ji})}{(T_{ij} + T_{ji})} \right]$$

$$\frac{T_{ij} + T_{ji}}{2} = K \frac{P_i P_j}{d_{ij}^{\alpha}}$$

$$T_{ij} = \left[\frac{T_{ij} + T_{ji}}{2} \right] + \left[\frac{T_{ij} - T_{ji}}{2} \right]$$

Tobler, W. R. (1976) Spatial Interaction Patterns, Journal of Environmental Systems, 6 (4), 271-301.

Tobler, W. R. (1983) An Alternative Formulation for Spatial-Interaction Modeling, Environment and Planning A, 15, 693-703.

The blog will have more and more references as the course continues

Questions

www.complexity.info

www.spatialcomplexity.info