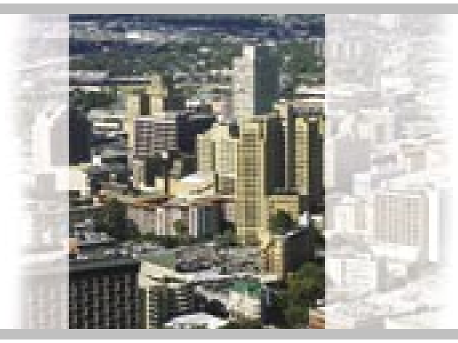
[Spatial modelling]

Chapter 16



Population growth dynamics in cities, countries and communication systems

Michael Batty and Narushige Shiode

This chapter discusses the aggregate dynamics of population systems that maintain a consistent regularity through time in the sizes of their elements. City size distributions represent the classic example, with the size of cities being inversely proportional to their rank. This is enshrined in Zipf's (1949) rank-size rule that provides one of the strongest and most exact relationships throughout the sciences. We first review fundamentals and then suggest that a wide class of random multiplicative processes provide good models for systems that grow to produce the kind of rank-size regularities that emerge for cities, countries and communications systems. We then focus on the internal dynamics that take place between time periods marked by these regularities showing that dramatic changes can take place, even though such systems appear to be rather stable at an aggregate level.

1 Social physics and Zipf's Law

Most of the models and methods presented in this book deal with spatial systems ranging from entire cities to local neighborhoods where analysis takes place in census tracts down to the geometry of buildings and streets. In this chapter, we depart from this focus on the meso and the micro, scaling up to cities and regions in nation states and the global space economy. Although we lose the rich detail that characterises more disaggregate spatial analysis, we gain the advantage of dealing with systems where events are aggregated to the point where trends and discontinuities in temporal and spatial patterns can be clearly detected. It is in this domain that some of the strongest regularities in the way spatial systems are organised become apparent, enabling us to build simple but effective models of their dynamics.

The earliest formal theories of how cities and regions were structured emerged from what came to be called 'social physics' (Stewart 1950). Social physics applied classical mechanics, fashioned around the concepts of force, gravitation and potential, to the way social systems were organised, usually in space. Initial applications tended to see city systems in equilibrium, because the signatures of such systems were usually based on regularities such as distance decay and diffusion, which appeared to follow well-defined power laws. However, in the last decade with the advent of complexity theory, this traditional analysis has been enriched with a concern for how robust and long-lasting regularities observed in such systems come to sustain themselves over many years in the face of quite volatile and effervescent dynamics. Here we begin by describing typical regularities and then show how stability in such patterns is consistent with rapid and sustained growth and decline, illustrating our argument with examples from city systems, population growth at the world level, and the penetration of new communications devices in nation states.

The best known and strongest regularity in the social sciences, some say even the sciences en masse, relates to the size and frequency with which different populations are distributed, in space as cities and through the economy in various forms of wealth. For such systems, it is widely agreed that there are a large number of small events and a small number of large events, with the relationship between being continuous and regular. For cities, this means that there are a large number of small places and a small number of large ones. Casual observation supports this regularity as we illustrate in Figure 1, but it was Zipf (1949) who first popularised and formalized the relationship that he enshrined in the term the

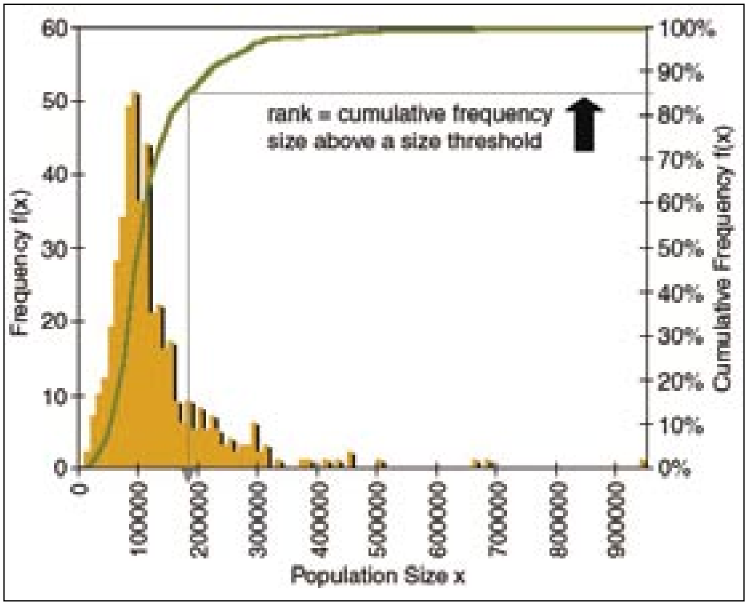
Figure 1 Scaling laws: Few large cities, many small towns: Prague to Dorset Shaftsbury

'rank-size law' or 'rank-size rule'. For the most part, we will try to avoid formal algebra here, but it is still necessary to state the law mathematically since much of our subsequent analysis depends on being clear about this meaning.

If we rank all cities by size from the largest to the smallest with the largest P_1 associated with rank 1, the next largest P_2 with rank 2, P_3 with rank 3, and in general P_r with rank r, then the relationship between city size and rank is that population is inversely proportional to its rank. The formal relation can be written as $P_r = P_1/\mathrm{r}^{\alpha}$ where α is a scaling constant. It was Zipf who first showed that $\alpha \approx 1$ not only for populations but also for word counts from novels and speeches, income distributions, and so on, thus establishing an even more curious characteristic of the law. Thus from the largest population P_1 which is often called the primate city, the second largest population is half the first $P_2 = P_1/2$, the third is a third of the first $P_3 = P_1/3$, and so on down the hierarchy with more and more smaller towns appearing, clustered ever nearer to one other in terms of size.

At first sight, this law seems to defy common sense. How, one might ask, can objects as complicated and complex as cities are be ordered in such a simple manner? Krugman (1996) sums it up extremely well when he says: 'We are unused to seeing regularities this exact in economics—it is so exact that I find it spooky. The picture gets even spookier when you find out that the relationship is not something new—indeed the rank-size rule seems to have applied to U.S. cities at least since 1890!' (page 40). If this is a universal law, then we need to explain how such regularities persist at the macro level when we know that at the micro level all manner of changes are taking place. Moreover, cities have changed a lot during this period although in our analysis, we have kept the aerial definitions the same. The urban United States in 1890 was a very different place from 1990 with the focus changing from east to west and north to south, the rust belt declining economically in the face of a growing sunbelt, and the drift from farms to the cities reversing itself during this period. In this chapter, we first review some novel explanations, but from our own data, we suggest that we are but at a beginning in terms of seeking good explanations. In this sense, we illustrate how important it is to explain space in terms of time, statics in terms of dynamics.

To make progress, we must trace a little history and reformulate the law. Although it is impossible to say who first observed these types of distribution, it was Alfredo Pareto in the late nineteenth century who first formalised a similar relationship concerning the distribution of income. Pareto said that if you count the number of individuals who have an income greater than a certain level, then this number would decline inversely with the level of income. In other words, as



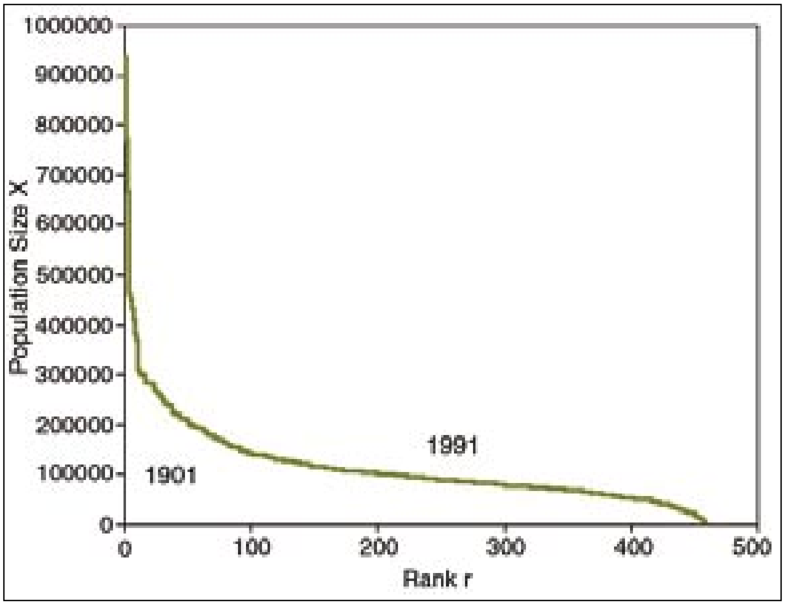


Figure 2 Frequency, cumulative frequency and rank size based on the British population in 1991

income gets larger, the number of persons who have that level of income gets smaller. In fact, this is none other than the rank-size distribution with the number of persons with an income greater than a certain amount the rank and the income level itself the size. It is worth noting that frequency and rank are consistently related. For every rank-size relation $P_r = P_1/r^{-\alpha}$, we can transform this into its frequency defined in terms of population as $f(P_i)$ and cumulative frequency $F(P_i)$ where i is now the unranked observation. The details are presented in the useful tutorial by Adamic (1999, Web access 2/9/2002).

We show these frequencies for the 1991 population of British urban areas in Figure 2, where the link from frequency to cumulative frequency to rank-size is implied graphically. It is possible of course to fit power laws to any of the three relations, but the rank-size is more convenient. Frequencies do not need to be counted and binned. We will come back to Figure 2 because the processes that we introduce will not, in general, yield distributions that are true Zipf laws. This might be anticipated in Figure 2A; it shows that the frequency is a highly skewed normal distribution—a lognormal—rather than an inverse power law although in its right hand tail, it might be thought of as a power law. This has been a source of considerable controversy, which we will return to later.

We conclude this brief introduction with a little more history. There is a massive research literature on Zipf's law. Wentian Li's reference archive (Li 1999, Web access 2/9/2002) only sketches the main contributions and he cites 220. In the physics literature of the last 10 years there are probably at least five hundred articles on power laws that reference Zipf, and for over 50 years there has been a steady stream of articles in regional science and urban economics. The earliest formal proposal to apply the model to cities was by Auerback in 1913, but Lotka in 1924, Goodrich in 1925, and Singer in 1936 speculated on city size distributions, all preceding Zipf's (1949) seminal text. Rashevsky and Lewis Fry Richardson, both writing in the 1940s, also made prescient contributions to the debate. Harry Richardson (1973) reviews the early work with an emphasis on how rank-size is generated using hierarchical models consistent with central place theory, while Carroll (1982) provides the most recent comprehensive review.

2 Explanations: multiplicative random processes generating order and growth

There are several models built around spatial economic theory that generate city size distributions consistent with Zipf's law, but most of these generate cities in terms of static spatial structure. Insofar as they are dynamic, they presume a spatial equilibrium. We have already noted models that generate size distributions consistent with notions about the hierarchy of central places. To these we must add various macro-economic analogues, often incorporating trade and growth theory of which one of the most recent is Krugman's (1996) and adaptations thereof (Brakman et al 1999). We do not propose to review these explanations further as we prefer a simpler and more parsimonious class of model. This model is based on random growth which, as we will show, is one generating considerable excitement at present, and the one which in terms of the systems we are interested in here seems the most applicable.

The most basic explanation of Zipf's law is so obvious that it is often overlooked. Imagine a set of objects all of which have the same size. Imagine that these objects can grow or decline by doubling or halving. For each object, flip a coin to determine if it grows or declines. Keep doing this for a few iterations and you will see that the chance that an object grows bigger and bigger is lower and lower because of the possibility each time that the object can decline rather than grow. This is a simple branching process that essentially says that the number of branches leading to large sizes relative to the total gradually decreases as the process continues. If you want to justify this for yourself, take say 10 objects and work through the process. This is quite laborious, but after a dozen or so trials, it becomes clear that the objects are beginning to order themselves into a small number of large ones and a large number of small ones. This process is even clearer if objects disappear when they fall below their lowest size with new objects at this smallest size replacing them. A particularly nice illustration of this process is given by Gabaix (1999) where he shows this even more clearly when the average growth of the system is constrained to a constant each time the objects are grown. In short, what this process is suggesting is that if things start small, the chances of them all growing big is very low because at each stage on their way to bigness, there is an even chance that they will get smaller.

Many plausible models are based on this kind of multiplicative growth where the notion of exponential growth arises from the process. The standard model in economics, which was developed by Gibrat in 1931 for the growth of firms, is essentially one of proportionate but random growth (for a history and detailed explanation, see Sutton 1997). What the model assumes is that growth is comprised of an average rate λ which applies to all the objects in the population—cities, firms, incomes, whatever—and a deviation from this average ε_{it}, which is a normally distributed random variate applicable to each object i at time t. The model can be stated very simply as $P_{it} = \lambda P_{it\text{-}1} + \varepsilon_{it}P_{it\text{-}1}$ where P is the population of object i at times t and t-1. The model simplifies to $P_{it} = (\lambda + \varepsilon_{it})P_{it\text{-}1}$ where if we start with the population at time $t = 0$ as P_{i0}, and apply this equation, we generate a growth series that we can write as $P_{it} = (\lambda + \varepsilon_{it})\ (\lambda + \varepsilon_{it\text{-}1})\ (\lambda + \varepsilon_{it\text{-}2})\ .\ .\ .\ (\lambda + \varepsilon_{i1})P_{i0}$. If we examine the logarithm of $(\lambda + \varepsilon_{it})$, assuming that ε_{it} is small relative to λ, then it is clear that we can simplify this as ε_{it} (using Taylor's expansion). We can thus write the series in log form as $\log P_{it} = \log P_{i0} + \varepsilon_{it} + \varepsilon_{it\text{-}1} + \varepsilon_{it\text{-}2}\ .\ .\ . + \varepsilon_{i1}$.

This is none other than the expression that leads to a normally distributed set of objects, in this case a normal distribution of the logarithms of the populations or, when transformed back, a lognormally distributed set of populations. We have already illustrated what a lognormal distribution looks like in Figure 2A, which shows the frequency of urban areas in Great Britain by size. This distribution is not a power law, hence our earlier caveat indicating that such distributions are likely to be lognormal, although one might be forgiven for assuming that its long tail could be approximated by such a law. Moreover, it arises as a result of a geometric process where growth is compounded in contrast to a normal distribution that is the result of additive growth. Gibrat called this the Law of Proportional Effect that is the result of a process where the statistics of the growth rates—the mean and the variance—are constant for all sizes of event. Its approximation using a power law assumes that the smaller events are disregarded and that the power law is fitted to the long tail, sometimes called the heavy or fat tail that occurs in the skew of the distribution to the right. It is not surprising that researchers have disregarded the lognormal distribution, for fitting Zipf's law just to the long tail of the distribution has been most successful. For example, Krugman (1996) took 130 metropolitan areas listed in the 1993 Statistical Abstract of the United States and demonstrated that nearly all the variance in the ranked city size distribution could be explained by $P_r = P_1 r^{-\alpha}$, where the value of the parameter was $\alpha = 1.003$, uncannily near the mythical value of unity!

During the last 10 years there has been an explosion of models built around Gibrat's law that show the links between proportionate random growth, random walks, and Brownian Motion in generating power laws for a whole range of physical and social phenomena (Stanley et al 1996). The solution to generate a power law from a multiplicative process is really rather simple: introduce a process that essentially cuts off the thin tail and leaves the distribution simply described by the fat tail, which then follows a power law or some related exponential. We have already anticipated this to an extent in our informal discussion earlier where we suggested that if an object got below a certain size, it disappeared and was replaced by a new object at the minimum threshold size. This is akin to not letting the object get below a certain size and it is this mechanism that, when added to the Gibrat model, leads to distributions that follow a power law. The most effective demonstration of this is provided by Levy and Solomon (1996a, 1996b) who initially presented the model as one that mirrored an economic market. They justified the lower bound for problems of this kind by arguing that income or wealth could not fall below a given limit due to subsidies in the economy. It is harder to say the same about cities, but Blank and Solomon (2001) suggest that there may be service thresholds below which there are no economies of scale, and cities would not exist. In fact what they also suggest is an extension to their model that embraces birth and death processes not unlike ideas originally developed by Simon (1955).

There are many developments at present with this style of model. One of the most intriguing issues is that mathematical analysis is throwing up all kinds of results that enable parameter values to be approximated from plausible assumptions about how such systems work. Gabaix (1999), for example, shows how this kind of model is consistent with various economic growth models, while Sornette and Conte (1997) argue that this is but one of a very large number of multiplicative processes with repulsion that are related to random walks, Levy flights, diffusion, and Brownian Motion. Distributions such as those shown in Figure 2 suggest all kinds of ad hoc approximations. Laherrere and Sornette (1998) discuss a class of stretched exponentials, showing how these fit various ranked distributions from cities to earthquake magnitudes. Malacarne, Mendes and Lenzi (2001) demonstrate that the Zipf law modified by Mandelbrot (1966) as $P_r = P_1(\mathrm{c} + r)^{-\alpha}$, which they call the q-exponential, fits various data from cities in Brazil and the United States particularly well. Recently an even more intriguing suggestion has been made by Reed (2001). He shows that if a birth process is assumed for Gibrat's law, and if it is assumed that events are distributed exponentially according to their age, then in the steady state, the distribution becomes what he calls 'double Pareto'; this means that the smaller events follow a different power law from the larger events with the thin tail being simulated by a positive power, the fat tail by a negative power.

Conte or Cont?

Lenzi or Pedron?

Our short review of where this field now stands barely does justice to the wealth of ideas currently being generated from both physics and, to a lesser extent, economics. Many of these ideas are being unpacked even further to link to new dynamic models based on conventional economic and physical theory that suggests how such power laws emerge. These models are much more explicit and causally complex than Gibrat's law, which we might take almost as the default or null hypothesis in this domain. These new models relate strongly to fractals, complexity and chaos, where various mechanisms are postulated as to how power laws emerge from simple assumptions. In what follows, we show that despite the stability implied by these distributions from time period to time period, it is critical to explore the underlying dynamics of change. To this end and with the premise that the kind of growth that occurs is random and proportionate, we show how this is consistent with very volatile dynamics where cities and other types of events move up and down the hierarchy quite dramatically while still preserving the underlying regularity that is Zipf's law.

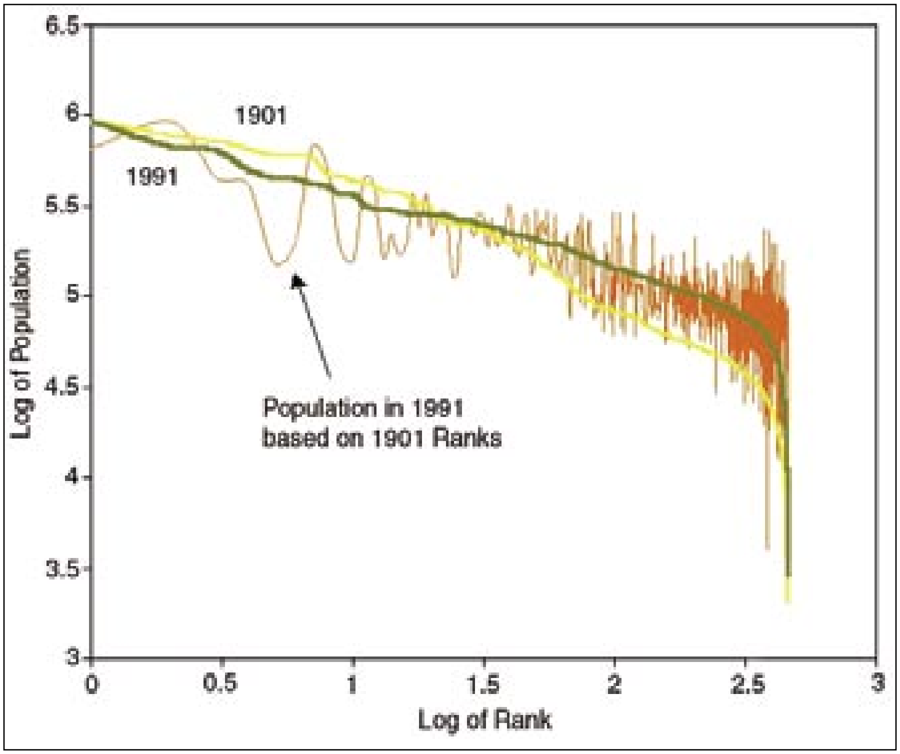


Figure 3 Rank-size relations for 1901 and 1991 with changes in rank

3 Searching for transitions in evolving systems

When we examine changes over time in the rank-size relationship for many different kinds of systems, we see a remarkable regularity in form. To anticipate our analysis, we show the log-log population rank-size relationship for 459 administrative units in Great Britain in Figure 3, where it is very clear that the form is quite similar. Yet this similarity over long time periods, indeed over decades and centuries even, disguises quite turbulent dynamics at the more micro level. Our causal knowledge of how places have changed over such extended time periods tells us that there is a lot more going on than meets the eye with respect to the kind of stability implied by the relations in Figure 3. We also show in this figure the changes in rank that have occurred over a century between 1901 and 1991, thus revealing substantial change.

To get a handle on such change, there are a variety of statistics that we can compute. If we normalize the populations for size, in the case of populations at times t = 1901 and t+1 = 1991, we can define the overall percentage shift in the population as $\Sigma_r \mid (P_{rt} / \Sigma_r P_{rt}) - (P_{rt+1} / \Sigma_r P_{rt+1}) \mid$. During this period, there is a large shift of some 46 per cent. This is in direct contrast to an average shift of 85 places in rank using the formula $\{\Sigma_i | r_{it} - r_{it+1} |\} / \mathrm{N}$ where r_{it} and r_{it+1} are the ranks of place i at times t = 1901 and t+1 = 1991. This average shift in rank is some 17 per cent of the overall number of ranks (459), and this is considerably less than implied by the absolute shift in populations (Batty 2001). In short, rank is a considerably more stable measure than absolute population from time period to time period, thus masking the micro-dynamics of settlement change. In the analysis that follows, we do not look any further at changes in the ranks of specific places but concentrate on trends and discontinuities that these kinds of statistics reveal through time.

We use a measure of difference between ranks due to Havlin (1995), not unlike the absolute average difference just stated but which has rather more tractable statistical properties associated with a measure of distance. This was originally used and extended by Vilenksy (1996) for the comparison of texts based on word frequencies, where it was found that books by the same author had rather more in common with each other in these terms than books by different authors. For any two times j and k, the distance is defined as $R_{jk} = \{\Sigma_i (r_{ij} - r_{ik})^2 / \mathrm{N}\}^{1/2}$. This gives the average shift in rank from time j to k. We refer to this as the Havlin Matrix **R** which is clearly symmetric with respect to time, though we are mainly interested in the data that relate to differences between forward time periods where $k > j$. From this very rich matrix of differences, we can develop various temporal analyses. Specifically, we are interested in the trend through time, in discontinuities in these trends which are associated both with dramatic changes in the rank of different places, and the possibility that places fall or gain in rank only to reverse themselves back towards their previous positions in time.

There are many ways in which we can measure different changes between ranks between different time periods. In the various data sets that we explore in this chapter, we look at the cumulative change in ranks based on the forward series R_{tt+1}, R_{tt+2} . . . $R_{tt+\mathrm{n}}$ as well as the backward series R_{t+nt}, R_{t+nt+1} . . . $R_{t+nt+\mathrm{n}-1}$. The plots that we make use a particular year as the numeraire in computing these forwards and backward series. We also look at the incremental forward series where the ranks change from time period to time period. This involves a comparison of R_{tt+1}, R_{t+1t+2} . . . $R_{t+n-1t+\mathrm{n}}$. We can of course examine many different orders of difference in the Havlin Matrix but we restrict ourselves to looking at first-order change based on $\Delta_{jk} = R_{jk} - R_{jk+1}$, $k > j$. We would usually expect to find that these differences were positive unless there were reversals in the general trend. From this difference matrix, we can compute the overall drift as $\Delta = \Sigma_{jk} \Delta_{jk} / \Sigma_{jk}|\Delta_{\mathrm{jk}}|$ where the range of the summation is defined according to the indices above. We have now introduced sufficient measures, and at this point we demonstrate the analysis on three very different data sets, each implying rather different kinds of dynamics.

4 Cities, countries and communications

The data set we have for urban populations is based on one constructed around British municipalities from the decennial censuses of population between 1901 and 1991. These data have been standardised to the 459 municipalities that existed in England, Scotland and Wales in 1901. Strictly speaking, it might be argued that these municipalities are not cities, but we prefer to use these labels as this reflects an exhaustive subdivision of the space economy that is not complicated by changing definitions of what a city is. The dynamics implied in this series are rather conservative. The British population has not grown all that much over the last 100 years relative to other growing systems: in 1901 it was around 37 million growing to 54 million by 1991. We would expect the dynamics implied by this growth to be fairly smooth from our causal knowledge of the British spatial system where most of the big cities were already established by the beginning of the last century. This is in quiet contrast to what has happened in other parts of the world.

In Figure 4, we plot populations for these 459 municipalities by their rank, and this shows remarkable stability in the aggregate form of this relationship. There are some crossovers, and the profile flattens slightly through time implying that the population is decentralising. In Figure 5, we show the Havlin plots which reveal consistent changes in rank but without any real surprises: no discontinuities and no obvious switches back to rank orders that have occurred earlier. We have also examined the first-order change through time based on examining the series $R_{1901,\,1911}$, $R_{1911,\,1921}$ . . . . This reveals substantial changes in rank order in the mid twentieth century during the period of economic depression and war,

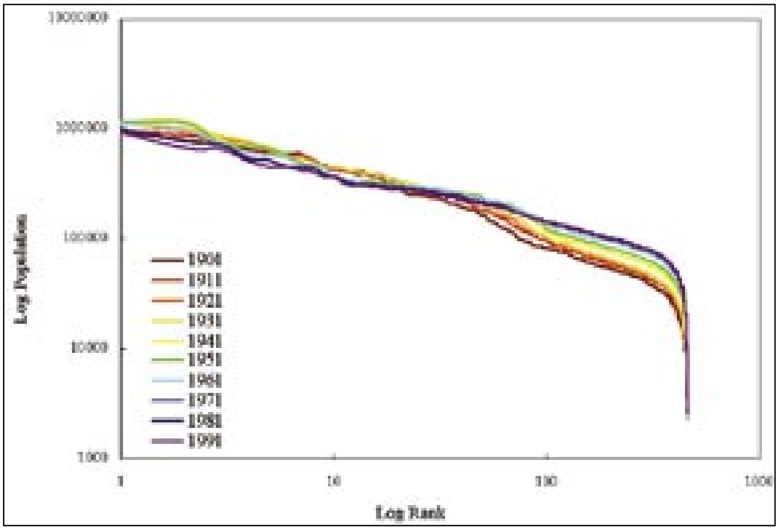


Figure 4 Log-log rank-size for British population 1901 – 1991

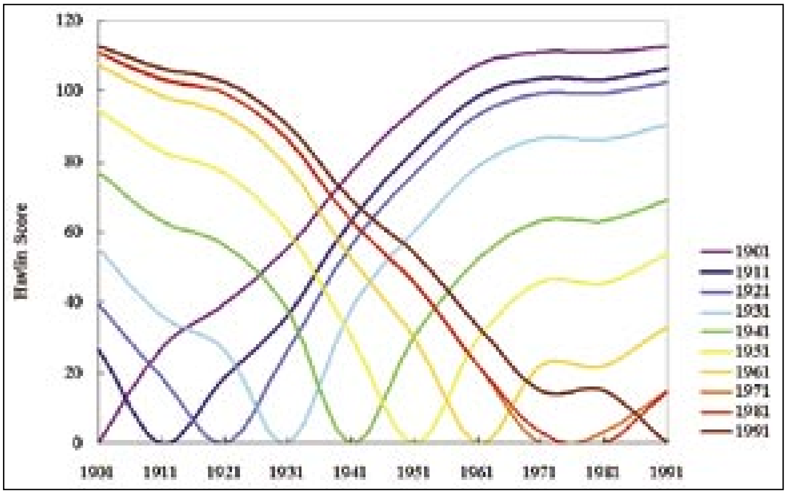


Figure 5 Havlin plots for rank differences in British population 1901 – 1991

although these are difficult to interpret without a detailed examination of the disaggregate patterns in different places.

Our second example is based on global populations in 210 countries. We have data on populations in these countries annually from 1980 to 2000 taken for ITU (International Telecommunication Union) published in 2001. These data are even smoother in a disaggregate way than the national British data as can be seen in Figure 6, where the log-log rank-size plots show very little major change during this period.

The Havlin plots in Figure 7 are quite regular with no surprising discontinuities. What this means is that the world-population system has hardly reordered itself at all over the last 20 years in terms of rank-size. The first-order change series $R_{1980, 1981}$, $R_{1981, 1982}$, . . . shows little change, even less than the British data, and this implies that as we aggregate, we lose details and we smooth variations. In fact, this is borne out very clearly if we compute the drift parameter Δ which is equal to unity, meaning that there are no reversals in overall rankings from time period to time period, although there are individual switches as the Havlin Matrix reveals. For the British population system, the drift parameter is almost unity at 0.995.

Our last example demonstrates a very different kind of dynamics for it deals with the recent rapid diffusion of mobile communications devices by country. The data are taken from the same sources as the global-population series and again show change from 1980 to the year 2000. We show log-log plots of the size and rank of mobile devices by country in Figure 8. These are considerably more skewed and have a much bigger thin tail than in the other two examples. In terms of growth dynamics, this would appear to be largely due to the fact that the system is growing rapidly and is still somewhat immature. This is clearly marked by the Havlin plots shown in Figure 9, which display considerable discontinuity in the early to mid-1990s when mobile phones were first penetrating many countries. In fact, the first-order change based on the series $R_{1980, 1981}$, $R_{1981, 1982}$ . . . is fairly smooth, but the changes in rank are very substantial with almost half the number of countries having changed rank by the year 2000. The drift parameter too reflects this turbulence with a value of 0.674, implying that 33 per cent of the system has moved back and forth between previous rank orders during the 20-year period.

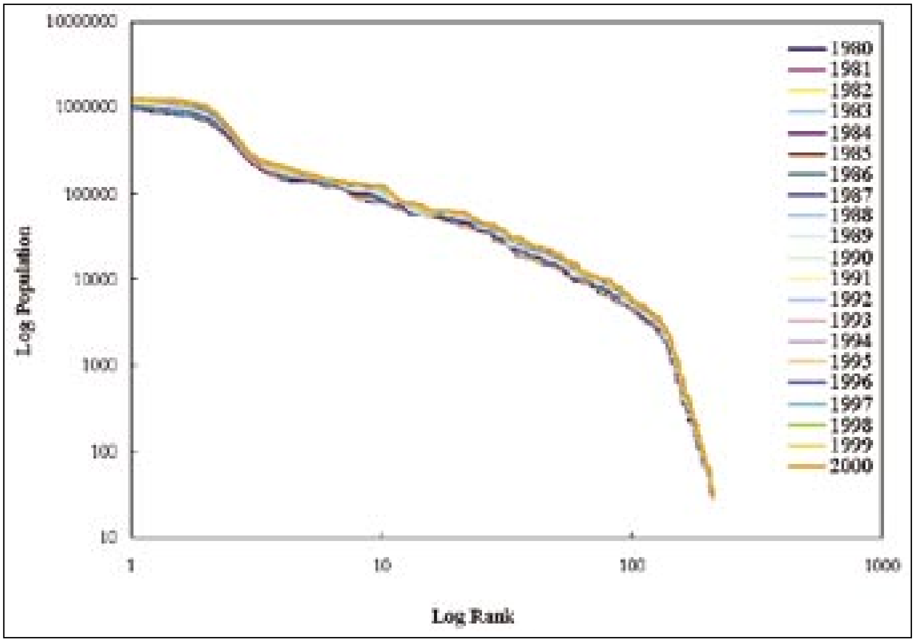


Figure 6 Log-log rank size for global country populations 1980 – 2000

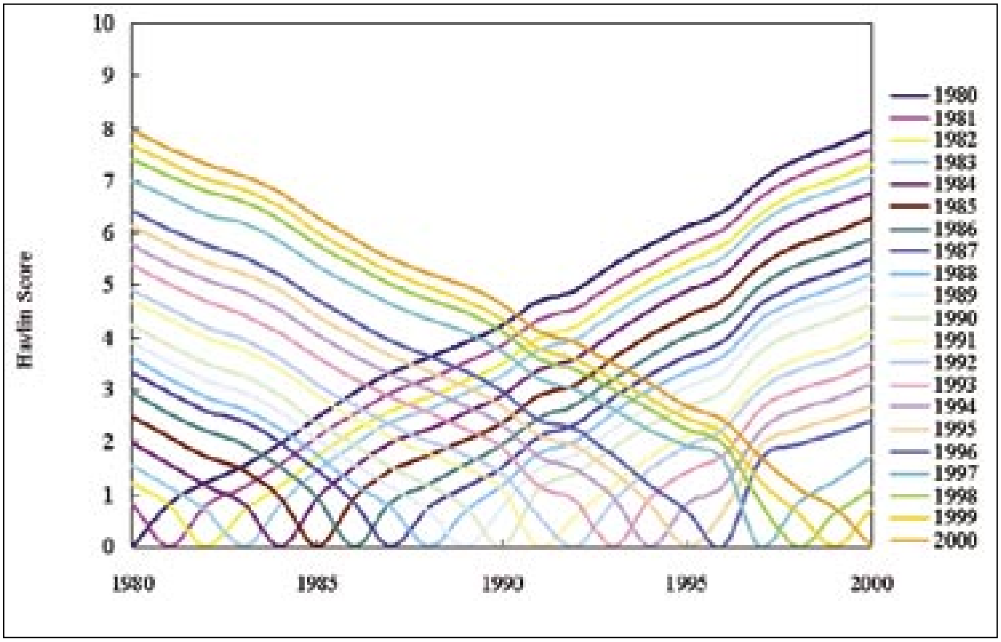


Figure 7 Havlin plots for rank differences in global populations 1980 – 2000

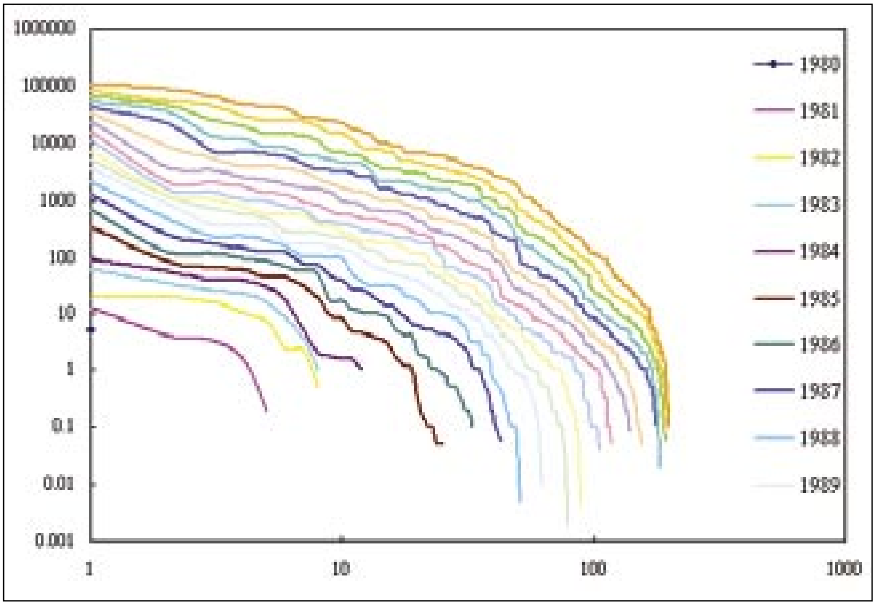


Figure 8 Log-log rank-size for penetration of mobile devices by country 1980 – 2000

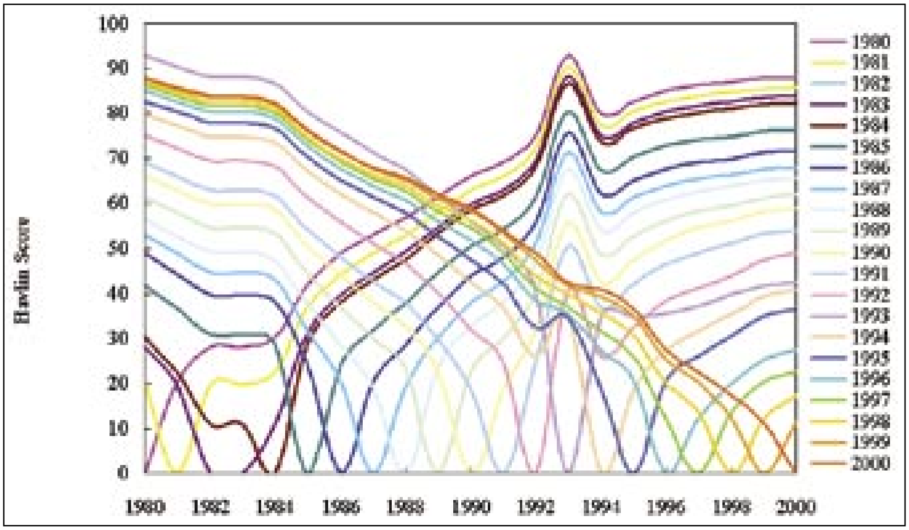


Figure 9 Havlin plots for rank differences in mobile devices by country 1980 – 2000

5 Next steps

This chapter has shown how aggregate spatial patterns marked by strong statistical regularities, implying stability at least superficially, are often simply masks for a deeper underlying dynamics. The rank-size rule is perhaps the best known of such regularities but to date, although there have been a flurry of explanations based on random growth theory, there has been little empirical exploration of what their dynamics means. We have not actually fitted any relationships here, for the various Zipf plots reveal very strong regularities in any case, and it is easy enough to guess that power laws will give very high correlations with the observed data. However, what we have shown is that by comparing different relationships through time using the Havlin Matrix, we are able to unpack the dynamics of change in more detail, illustrating that seeming regularity can be consistent with turbulence at a more disaggregate level.

There is much work still to do with the data that we have explored here. In particular we need to consider the extent to which we need to truncate the distributions by getting rid of the thin tails so that we can fit the best power laws. We also need to assess the extent to which such systems give Zipf parameters close to unity, and ways in which we might develop random growth models that reflect the kinds of systems that these rank-size profiles characterize. We also need to explore the extent to which we can say that the Havlin statistics reveal that similarities through time reveal changes belonging to one system or another. Just as it is possible to compare two different authors and works by the same authors using these statistics, we need to pose and answer the question as to whether systems that change through time are nearer to other systems at the same time or to themselves at earlier time periods. In this manner, we would hope to extend this kind of analysis into ways of classifying spatial systems into different types of structural dynamics.