The Rank-Size Rule in Europe: testing Zipf's law using European data

Graham Crampton, Economics Department, Reading University,
Whiteknights, Reading RG6 6AW, U.K.

Email: g.r.crampton@reading.ac.uk

Paper prepared for
ERSA2005 Conference, Amsterdam, August 2005, Paper 185, Theme K

Abstract
Keywords: Zipf; primacy; cubic.

The large literature on the rank-size rule of city sizes has received rather inconsistent treatment in the European continent. Part of the problem has been the fact that (unlike the U.S.) there are inconsistent Census dates and no uniform definition of what is meant by a functional urban area. This paper uses data from a French research project which provides physical urban area data for a number of European countries, down to quite small minimum urban sizes. This allows international comparison of the usual Pareto estimation parameters, and also some examination of whether square or cubic terms are significant. The nature and economic basis of such non-linearities in the logarithmic rank-size relationship are of interest. The paper examines the possibility of measuring the primacy of the urban size distribution using the second or third order polynomial fit in OLS estimations, and focuses on the way in which the slope of the Log(Rank) / Log (Size) relationship may differ within the estimated fit. As a result, the crude Zipf law estimate using the linear functional form may produce quite misleading results. We find that certain countries with strong regional governance or a history of regional city-states have `anti-primate' distributions, at least at the top end of the size distribution.

1. Introduction

We do not have space to attempt a review of the enormous body of research on the rank-size rule, but Parr (1969) is an early element of the theoretical discussion of city size hierarchy, and important recent contributions have come from Gabaix(1999), Reed (2002), and Dobkins/Ioannides (2000). The literature has recently `come to life' again, with contributions of both theoretical and empirical form. The principal theoretical insight has been the understanding of the extent to which the RSR is simply a statistical phenomenon with virtually no empirical (or economic, or regional science) content. In particular, Reed (2002) and Gabaix(1999) stressed that if urban growth rates obeyed the Gibrat Law, i.e. had a statistical distribution whose mean and variance did not vary with urban size, then the rank-size rule evolves dynamically purely as a side effect of the Gibrat law. Reed (ibid) also found that the rank-size relationship could slope with different signs, depending on which tail of the distribution one is examining. Estimations of the RSR function over time are also common (e.g. Guérin-Pace (1995)).

Much of the literature is based on US metropolitan area data, largely because consistent SMSA definitions have been reported for the urban areas (with core cities over 50,000 population) since 1950 as part of the US Census.

2. Comments on Definitions and Data

A recent French-based research project, GEOPOLIS, has used satellite photography to define a database of worldwide urban areas, using a morphological criterion. I am not aware of detailed description of the source in English, but Moriconi-Ebrard (2000) makes it clear (ibid, p.24) that a 200m. distance cut-off between built-up areas of an agglomeration plays the crucial role. That is, a metropolitan area is made up of local authority areas, and a local authority area is included if its largest built up component is not separated by more than 200m. from the built up part of the core city or other linked areas. This of course is *not* based on commuting links, the way the US definition is, or any other form of economic linkage. But it has the great advantage of being internationally uniform, and available for European metropolitan areas.

3. Estimates of Various Forms of the Rank-Size Form for different Countries.

We wanted to make use of the GEOPOLIS data base to produce simple RSR estimations for those countries in which adequate data had been made available, and add square and cubic terms to the specification. This would enable us to produce rather more realistic versions of `primacy' or `non-primacy' in the urban size distributions for individual countries. It turns out there were European 14 countries for which this could be done (sadly excluding the U.K.). They were Austria, Belgium, Croatia, Spain, France, Germany, Hungary, Ireland, Italy, Netherlands, Portugal, Romania, Slovakia, and Switzerland.

Following the more conventional form of estimation of the RSR, with the log(rank) as the dependent variable, and powers (square and cube) of the log metropolitan population as explanatory variables, we have:

$$LRANK = d_0 + d_1\ LMPOP + d_2\ SQLMPOP + d_3\ CULMPOP$$

The conventional estimations of the RSR have assumed c_2 and c_3 to be zero. We carried out estimations of the simple linear, the linear and square term, and the full cubic form, for all 14 countries. Results are given in Table 1, with the Durban-Watson statistic. The latter is more normally used in time-series econometrics, but we wished to make use of it to study certain properties of the residuals (below section X). We also note in Table 1 the number of observations on metropolitan areas for each country, and the minimum population size for that country. The collection of data on metropolitan area population was clearly arduous and time consuming for the GEOPOLIS team (and was not completed for Europe). Some of the countries, notably Austria, and the still more rural countries of Ireland and Slovakia had distributions going down to 10,000, but the size distribution for Spain goes down only to 80,000 and those for France and Germany to 50,000. For the bigger countries in the sample, the lower tail of the size distribution was neglected, no doubt for time and resource reasons.

For the simple linear estimates of the RSR

$$LRANK = b_0 + b_1\ LMPOP$$

the most interesting outcome is whether the indications are of a (linear) primate (i.e. big city - dominated) distribution. Let us examine how many of the coefficients of LMPOP in the linear form are significantly (at 99%) different from one. In the form of the specification we

have used primacy would be indicated by the absolute value of the coefficient b_1 of LMPOP less than one, and anti-primacy (or middle-domination for want of a better name) by a value more than one. Focusing just on the coefficients that are 99% significantly different from one, we find 5 primate and 2 anti-primate (middle-dominated) countries.

Table 1. OLS estimates of Rank-Size equations (Met Pop 2000), DepVar=LogRank

(Standard errors in parentheses; ** Significant at 99%, * at 95%)

Austria	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=66	12.4	-0.905**			0.84	97
		(0.020)				
Minpop=	21.9	-2.65**	0.0786**		0.72	99
10,000		(0.16)	(0.0071)			
	33.5	-5.74**	0.350*	-0.00783	0.88	99
		(1.97)	(0.17)	(0.0050)		

Belgium	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=50	10.6	-0.711**			0.38	94
		(0.025)				
Minpop=	20.3	-2.38**	0.0704**		0.30	96
16,000		(0.33)	(0.014)			
	91.2	-20.2**	1.54**	-0.040**	1.08	98
		(2.5)	(0.20)	(0.006)		

Croatia	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=44	12.4	-0.941**			0.59	99
		(0.016)				
Minpop=	17.7	-1.91**	0.0445**		0.57	99
10,000		(0.22)	(0.010)			
	-13.4	6.58	-0.724*	0.0230*	0.74	99
		(3.3)	(0.30)	(0.009)		

Spain	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=55	15.2	-0.981**			1.29	98
		(0.017)				
Minpop=	14.6	-0.889*	-0.004		1.32	98
80,000		(0.33)	(0.013)			
	-124.2	31.0**	-2.44**	0.062**	2.05	99
		(4.8)	(0.37)	(0.009)		

France	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=109	15.4	-0.992**			1.54	99
		(0.011)				
Minpop=	18.9	-1.56**	0.0226**		1.22	99
51,000		(0.17)	(0.007)			
	-30.7	9.99**	-0.87**	0.023**	0.88	99
		(1.99)	(0.15)	(0.004)		

Germany	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=158	15.9	-1.01**			0.20	99.2
		(0.007)				
Minpop=	23.3	-2.20	0.0471**		0.24	99.7
50,000		(0.081)	(0.0032			
	56.8	-10.0**	0.65**	-0.016**	0.61	99.8
		(0.83)	(0.066)	(0.0017)		

Hungary	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=57	14.4	-1.06**			1.51	95
		(0.034)				
Minpop=	29.2	-3.62**	0.110**		0.99	98
20,000		(0.30)	(0.013)			
	-102.6	29.9**	-2.71**	0.078**	1.08	99
		(3.4)	(0.29)	(0.008)		

Ireland	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=35	11.7	-0.903**			0.33	90
		(0.05)				
Minpop=	33.9	-4.95**	0.181**		0.49	98
10,000		(0.34)	(0.015)			
	63.4	-13.0*	0.90	-0.021	0.35	98
		(5.0)	(0.45)	(0.013)		

Italy	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=69	15.7	-1.019**			0.80	99
		(0.015)				
Minpop=	22.7	-2.11**	0.0424**		0.90	99
83,000		(0.32)	(0.012)			
	146.7	-30.8**	2.25**	-0.056**	1.20	99
		(5.8)	(0.44)	(0.011)		

Netherlands	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=44	15.3	-1.059**			0.39	96
		(0.031)				
Minpop=	27.9	-3.08**	0.0807**		0.28	97
52,000		(0.47)	(0.019)			
	-144.1	38.0**	-3.17**	0.085**	1.23	99
		(6.6)	(0.52)	(0.014)		

Portugal	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=29	10.6	-0.748**			0.78	85
		(0.059)				
Minpop=	41.5	-5.89**	0.210**		2.43	97
25,000		(0.53)	(0.022)			
	152.8	-34.1**	2.57**	-0.065**	1.55	97
		(9.3)	(0.78)	(0.022)		

Romania	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=58	16.4	-1.154**			1.02	94.5
		(0.037)				
Minpop=	17.2	-1.28	0.005		0.99	94
40,000		(0.69)	(0.29)			
	-225.6	58.0**	-4.80**	0.13**	0.79	97
		(9.9)	(0.80)	(0.02)		

Slovakia	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=73	15.9	-1.240**			0.52	97.6
		(0.023)				
Minpop=	11.5	-0.391	-0.0399*		0.67	97.7
10,000		(0.40)	(0.019)			
	-155.9	46.2**	-4.34**	0.13**	2.09	99
		(3.8)	(0.35)	(0.011)		

Switzerland	C	LPOP	LPOPSQ	LPOPCU	DW	AdjRsq%
n=46	13.2	-0.938**			0.67	98
		(0.02)				
Minpop=	6.50	0.24	-0.051**		1.04	98
20,000		(0.37)	(0.016)			
	-5.84	3.45	-0.328	0.008	1.00	98
		(6.44)	(0.555)	(0.016)		

Let us also discuss the interpretation of the square and cubic terms, for the specifications where they are included.

Figures 1 and 2 show versions of primacy in the urban size distribution that could be termed primacy `of the second order' and `of the third order'. Primacy `of the second order' means that the specification with the square term has the coefficient c_2 positive

$LRANK = c_0 + c_1 LMPOP + c_2 SQLMPOP$

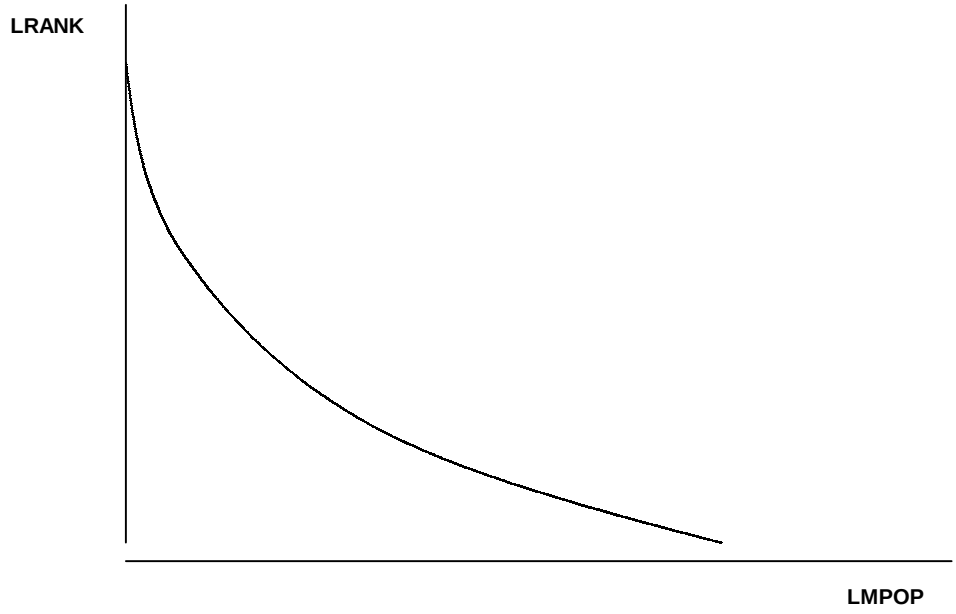

Figure 1 Primacy in urban size distribution measured by $c_2 > 0$

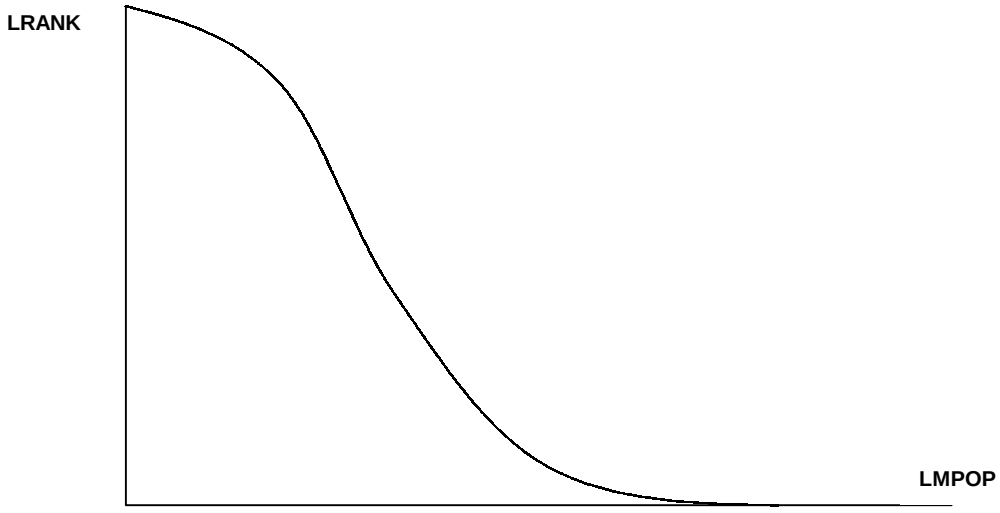

Figure 2 Primacy in urban size distribution measured by $d_3 > 0$

Primacy `of the third order' is indicated by $d_3 > 0$ in the estimation of

$LRANK = d_0 + d_1\ LMPOP + d_2\ SQLMPOP + d_3\ CULMPOP$

An examination of the OLS results in Table 1 shows the following, in terms of the signs of firstly c_2 and then d_3, summarised in Table 2.

Table 2 Indications of Primacy using Square and Cubic Terms in Rank Size Rule Estimations from 14 countries.

	Sign of $c_2 > 0$ (second order primacy)	Sign of $d_3 > 0$ (third order primacy)
Positive	10	6
Not significant (at 99%)	3	4
Negative	1	4

We see that the urban size distributions for the 14 countries come out primate (that is a positive coefficient c_2 by 10 to 1 in the second order estimations, and 6 to 4 in the third order estimations (that is a positive d_3). The 6 countries with primacy `of the third order' using $d_3$ in the cubic equation are Spain, France, Hungary, Netherlands, Romania, Slovakia. The 4 `anti-primate' (middle-dominated) countries with significantly negative d_3 come out as Belgium, Germany, Italy, Portugal.

It is notable that Belgium and Portugal came out in the simple linear run as primate (i.e. the opposite to the cubic) but with a very low Durbin Watson statistic, suggesting a weak specification and clusters of positive and negative errors. With the cubic equation, the D-W statistic is higher (closer to 2), indicating weaker clustering of residuals, but only for Romania does it (nearly) suggest no significant clustering at all.

4. Some Further Comments on Primacy

It is tempting to take a little further the concept of `primacy of the third order'. One possible extra step is to use the cubic estimation form, and compute what the slope is `at the top of the size distribution', that is when log(MPOP) takes its maximum value for that particular distribution. A straightforward differentiation of the cubic form gives:

$$\frac{dLRANK}{dLMPOP} = d_1 + 2d_2 LMPOP + 3d_3 (LMPOP)^2$$

and if we use the estimates of the d-coefficients, and the maximum value of LMPOP for that distribution, we readily arrive at the slope. The example shown in Fig. 2 above shows the cubic form with a slope well below one at the right hand intercept, which could be interpreted as a new form of primacy `at the top of the size distribution'. We are not then limited by the simplest rank-size log-linear function which constrains the log-slope to be constant and uses the slope below one as the indicator of primacy.

Out of the 14 countries, we ignored 3 countries for which we found that the forecast slopes `at the top of the distribution' came out positive (these were Hungary, Netherlands and Romania). For 3 other countries, the cubic coefficient d_3 was insignificant (these were Austria, Ireland, Switzerland), and it seemed appropriate to use the square functional form to compute a slope estimate. One of these (Ireland) produced a positive estimated slope.

Table 3 Indications of Primacy using estimated slopes of dLRANK/dLMPOP evaluated at LMPOP maximum.

Country	Primate at top of distribution?	Estimated slope at Log(MPOP) max
Germany	No	-1.51
Switzerland	No	-1.25
Belgium	No	-1.17
Italy	No	-1.16
Slovakia	Yes	-0.73
Portugal	Yes	-0.68
Croatia	Yes	-0.393
Austria	Yes	-0.386
France	Yes	-0.14
Spain	Yes	-0.04

The suggestion from Table 3 is that the most primate distributions `at the top end' are to be found in France and Spain, and the least primate are in Germany and Switzerland. It is interesting that of the most mature and well-developed European economies, Germany and Switzerland have the strongest system of regional governance. Belgium could also be mentioned in connection with strong regional autonomy, and Italy has its long history of city-states and battle for national identity and independence in the 19th century. Italian colleagues also often stress that the economic capital (Milan) and the political capital (Rome) are in many ways rivals, and indeed almost symbolise the two `old' Italies. In contrast, Austria and

France are often mentioned with reference to extreme primacy in urban size distributions, and even Spain, despite its two leading cities of very similar size, has a striking lack of regional cities of substantial size. We find it useful to fit a curvilinear relationship to the size distribution, and to draw primacy conclusions for a particular point of the distribution, namely the top. We could equally well examine other sections of the distribution.

5. Conclusions

We have used the GEOPOLIS data base of urban sizes for 14 European countries to compute for the year 2000 estimates of the Zipf law, including the traditional linear specification and square and cubic terms. The latter allow us to generate measures of primacy in the metropolitan size distribution that may vary down the Log(Rank) / Log (population) curve, unlike the traditional linear estimation of the rank-size rule. We find quite a broad spread of primate and `anti-primate' (or middle-dominated) distributions, with federal governance systems or recent national unity and independence being associated with an anti-primate distribution.

There are two principal weaknesses of the data set. Firstly, all of the urban area definitions are based on physical separation using satellite photography. Many would consider that a set of metropolitan areas defined as physical built-up areas, but without any reference to function (including commuting flows) was unsatisfactory.

Secondly, resource constraints in the GEOPOLIS project meant that not all European countries were included; and of those that were, the truncation size at the lower end of the distribution varied. Smaller countries (e.g. Ireland) had their size distribution detailed down to a size 10,000, whereas larger countries (e.g. Spain and Italy) had a truncation at population 80,000. This affected the findings using non-linear functional forms, given that the detail of the `lower tail' was lost if the truncation was at quite a big size. However, we are still able to obtain some interesting results, contrasting urban size distributions typical of unitary versus federal states in Europe.

There is a Europe-wide data set of urban areas for those over 200,000 population, which was discussed at an earlier RSA conference (Crampton (2003)), this has the weakness of completely omitting the smaller end of all the distribution. The current paper (although it gets no closer to a full Europe-wide urban size hierarchy down to a small size) focuses on gaining some insights from what the GEOPOLIS project gave us. We have not in this paper made any

use of the time series of (fixed boundary) sizes going back to 1950. That raises a different range of issues over using fixed-boundary metropolitan areas back into time periods when the urban area definition may be less appropriate.

Finally, since the author is U.K. based, a comparable set of U.K. urban areas using the same GEOPOLIS definitions would be interesting, though labour-intensive to generate.

References

Dobkins, L.H. and Y.M. Ioannides (2000) Dynamic evolution of the US city size distribution, in *The Economics of Cities*, J-F. Thisse and J-M. Huriot, eds., (Cambridge: C.U.P.).

Gabaix, X. (1999) Zipf's Law for cities: an explanation, *Quarterly Journal of Economics*, **114**, pp. 739-767.

Guérin-Pace, F. (1995) Rank-size distribution and urban growth, *Urban Studies*, **32**(3), pp. 551-562.

Le Galès, P. (2002) *European Cities* (Oxford: Oxford University Press).

Lotka, A.J. (1924) *Elements of Physical Biology* (New edition: *Elements of Mathematical Biology*, New York, 1965).

Moriconi-Ebrard, F. (2000) *De Babylone à Tokyo* (Paris: Ophrys).

Parr, J.B. (1969) City hierarchies and the distribution of city size, *Journal of Regional Science*, **9,** pp. 239-253.

Reed, W.J. (2001) The Pareto, Zipf and other power laws, *Economics Letters*, **74**, pp. 15-19.

Reed, W.J. (2002) On the rank-size distribution for human settlements, *Journal of Regional Science*, **42**(1), pp. 1-17.

Appendix. Calculation of Estimated dLRANK/dLMPOP evaluated at the population of the largest metropolitan area [ln(Metpop Max)].

Value of Coefft.	Coefficient	Country	Primacy (P or AP)	Compare with linear
-20.2	D1EST			
1.54	D2			
-0.04	D3			
15.3	LMMAX			
Est slope	-1.167	Belgium	AP	opposite
-34.1	D1EST			
2.57	D2			
-0.065	D3			
14.7	LMMAX			
Est slope	-0.680	Portugal	P	same
6.58	D1EST			
-0.724	D2			
0.023	D3			
13.5	LMMAX			
Est slope	-0.393	Croatia	P	same
31	D1EST			
-2.44	D2			
0.062	D3			
15.4	LMMAX			
Est slope	-0.0402	Spain	P	Linear =1
9.99	D1EST			
-0.87	D2			
0.023	D3			
16.1	LMMAX			
Est slope	-0.139	France	P	Linear =1
-10	D1EST			
0.65	D2			
-0.016	D3			
16.1	LMMAX			
Est slope	-1.512	Germany	AP	Linear =1
29.9	D1EST			
-2.71	D2			
0.078	D3			
14.6	LMMAX			
Est slope	0.647	Hungary	Positive	
-30.8	D1EST			
2.25	D2			
-0.056	D3			
15.1	LMMAX			
Est slope	-1.156	Italy	AP	Linear =1
38	D1EST			
-3.17	D2			
0.085	D3			
15	LMMAX			
Est slope	0.275	Netherlands	Positive	

58	D1EST			
-4.8	D2			
0.13	D3			
14.6	LMMAX			
Est slope	0.972	Romania	Positive	
46.2	D1EST			
-4.34	D2			
0.13	D3			
13	LMMAX			
Est slope	-0.73	Slovakia	P	opposite
Austria				
Ireland	**used**	**square**		
Switz		**function**		
-2.65	D1EST			
0.0786	D2			
14.4	LMMAX			
Est slope	-0.386	Austria	P	same
-4.95	D1EST			
0.181	D2			
13.8	LMMAX			
Est slope	0.0456	Ireland	positive	
0.24	D1EST			
-0.051	D2			
13.8	LMMAX			
	-1.17	Switzerland	AP	