University of Sussex
Business, Management & Economics

Economics Department Working Paper Series

No. 60-2013

Quantile Kernel Regression for Identifying Excellent Economists

Richard S.J. Tol
Department of Economics, University of Sussex
Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands
Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands
Email: R.Tol@sussex.ac.uk

Abstract: Quantile kernel regression is a flexible way to estimate the percentile of a scholar's quality stratified by a measurable characteristic, without imposing inappropriate assumption about functional form or population distribution. Quantile kernel regression is here applied to identifying the one-in-a-hundred economist per age cohort according to the Hirsch number.

JEL Classification: A11

Key Words: quantile kernel regression; Hirsch number; economics

1. Introduction

The Hirsch number (Hirsch 2005) is an excellent measure of life-time achievement. The Hirsch number cannot fall over time and tends to increase. Any ranking based on the Hirsch number thus favours those with a longer career. This is fine for many purposes, but not if the aim is to identify excellent individuals in a cohort, e.g., for hiring scholars (Ellison 2010). The Hirsch rate (Burrell 2006;Liang 2006) – the Hirsch number over the number of active years – corrects for career length. However, the Hirsch rates assume a linear relationship between Hirsch number and active years. This may be problematic when comparing job candidates of different ages if the relationship is (locally) non-linear. This paper therefore proposes quantile kernel regression (Sheather and Marron 1990) as a method to find exceptional researchers. Kernel regression does not impose linearity or any other functional form. Quantile regression focuses the analysis on exceptional, rather than average, scholars. The proposed method is applied to a sample of 32,000 economists. For illustration, I am looking for the one-in-a-hundred economists in each age group.

As far as I know, I am the first to do apply quantile kernel regression to this problem.[1] Of course, this paper is not the first to seek to identify excellence (van Leeuwen et al. 2003). Percentiles are a natural way to identify excellence: A scholar is excellent if she is better than, say, 99 out of 100 of her peers. People may be close in rank but far apart in percentile, and vice versa. The "crown indicator" – a z-score (Moed 2010;van Raan 2005) – corresponds to percentiles only if the underlying distribution is normal[2] or can be transformed to normality (Lundberg 2007). Care needs to be taken when defining the peer population (Gingras and Lariviere 2011;Herranz and Ruiz-Castillo 2012;Leydesdorff and Opthof 2011;van Raan et al. 2010;Waltman et al. 2011a;Waltman et al. 2011b).Quantile kernel regression directly estimates the percentiles (the quantile part) as a function of the subpopulation characteristics (the regression part), and without imposing any particular distribution (the kernel part).

The paper proceeds as follows. Section 2 presents the methods. Section 3 shows the data. Section 4 discusses the results. Section 5 concludes.

2. Method

The intuition behind kernel regression is straightforward (Takezawa 2006). A standard ordinary least squares regression $y=X\beta+u$ with $u\sim N(0,\sigma^2)$ is equivalent to $y\sim N(X\beta,\sigma^2)$. That is, our prediction for y, $X\beta$, is the expected value of a probability density function. That density is the

[1] A Scopus search on "quantile" or "kernel" and "Journal of Informetrics"or "Scientometrics" returned only three papers (Beirlant et al. 2007;Hengl et al. 2009;Sarabia et al. 2012), none of which does what the current paper does.

[2] The standard interpretation of a z-score is a two-sided test. A z-score of 2 (or rather 1.96) implies a score that is either exceptionally high (or rather 39 out of 40) or exceptionally low (or rather 1 out of 40). It strikes me that we are more often looking for exceptionally good scholars.

density of *y* conditional on *X*. Below, we consider univariate regression only, so we replace matrix *X* by vector *x*; β becomes a scalar.

A conditional density function is defined as:

(1) $$f(y|x) = \frac{f(x,y)}{f(x)}$$

The generic definition of a univariate kernel density function is:

(2) $$\hat{f}(x) = \frac{1}{nh_x}\sum_{i=1}^{N} K\left(\frac{x-X_i}{h_x}\right)$$

where h_x is the so-called bandwidth, X_i are a series of observations i=1,2,…,N, and K is kernel function, which can be any function that integrates to one, with a first moment at zero and a finite second moment.

A bivariate kernel density is typically defined as:

(3) $$\hat{f}(x,y) = \frac{1}{nh_xh_y}\sum_{i=1}^{N} K\left(\frac{x-X_i}{h_x}\right) K\left(\frac{y-Y_i}{h_y}\right)$$

The conditional kernel density follows from substituting (2) and (3) into (1). The expected value of *y* conditional on *x* then follows from:

(4) $$\mathbb{E}(y|x) = \int_x \; y\hat{f}(y|x)\mathrm{d}y$$

Although kernel regression analysis is often focused on Equation (4), we in fact derive, as an intermediate step, the entire conditional distribution.[3] That is, we know not only the conditional mean (Equation (4)), but also the higher conditional moments, the mode, median and any percentile that may hold our interest.

The literature on kernel density estimation is focussed on two questions: What kernel function *K* to use, and with which bandwidth *h*? There is no objective answer to those questions. If we use a Gaussian kernel function, we want the kernel density to be as close as possible to the Normal distribution, and we define closeness as the mean integrated square error, then the optimal bandwidth is:

(5) $$h_x \cong 1.06\hat{\sigma}_x n^{-0.2}$$

See (Takezawa 2006). I follow these conventions below.[4]

[3] Note that, under certain assumptions, it is possible to simplify the equations and skip the estimation of the bivariate density.

[4] The literature on quantile kernel regression (Falk 1986;Parzen 1979;Sheather and Marron 1990;Takeuchi et al. 2006;Yu and Jones 1998) advocates a range of alternative bandwidths, derived from a combination of distance measures and target distributions.

For the bivariate kernel, I use the bivariate Gaussian with bandwidth

$$(6) \qquad \begin{bmatrix} h_x^2 & h_{xy} \\ h_{yx} & h_x^2 \end{bmatrix} \cong 1.06^2 \begin{bmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xy} \\ \hat{\sigma}_{xy} & \hat{\sigma}_y^2 \end{bmatrix} n^{-0.4}$$

3. Data

IDEAS/RePEc (Krichel and Zimmermann 2005) is an internet service for economists at http://ideas.repec.org/. It operates as an archive for working papers (almost 500,000 items from over 3,500 series). Publication lags are substantial in economics. IDEAS/RePEc serves as the main platform for the early availability of submitted articles. IDEAS/RePEc also operates archives for journal articles (almost 800,000 items from over 1,500 journals), books and book chapters (almost 30,000 items), and software components (almost 3,000 items). Paper series, individuals and departments are ranked on a range of indicators (Seiler and Wohlrabe 2012;Zimmermann 2007). There are nascent activities on curated bibliographies, academic genealogies, and collaboration networks. IDEAS/RePEc has profiles of over 43,000 economists, and a database of over 10 million references, 4 million of which are to items in the publication databases. I here exploit the latter two.

Data were collected on 9 and 10 November 2012. Data were scraped from the simplified citation profiles. The citation profiles have information on the number of works (journal articles, working papers, books, chapters), the number of co-authors, total citations, total self-citations (of any co-author), Hirsch number, and years since first publication. There is a separate page for each citation profile. I wrote a Matlab script, reproduced in Appendix A of (Tol 2013), to visit each page and extract the relevant data. The indicator used here is the Hirsch number, which has not been corrected for self-citations.

Data can be found at http://www.sussex.ac.uk/Users/rt220/CohortMatthewPublic.xlsx

More than 43,000 economists have registered with IDEAS/RePEc, but only 31,781 have a complete profile. Figure 1 depicts the numbers per cohort. Recall that cohorts are defined by years since first publication. "Young" economists are overrepresented, reflecting both a thriving discipline in a growing higher education sector and a growing importance of IDEAS/RePEc.

4. Results

Figure 2 shows a scatter plot of Hirsch numbers and number of active years. It also shows selected percentiles of the conditional kernel density function. The basic patterns are easy to understand. Hirsch numbers increase with seniority. The spread of the distribution too increases

with seniority as true quality is revealed over time. Figure 3 shows more detail: the histograms and the conditional kernel density functions at 5, 10, 15 and 20 years since first publication.

In their first year, about 0.3% of economists have a Hirsch number of 2. This increases to 3 for 0.2% of the sample after a year, for 1.5% after two years, for 6.2% after three years, for 16.3% after four years, to 30.5% after five years, to 85.3% after ten years, to 97.5% after fifteen years, and to 99.3% after twenty years.

To stand out after five years, you need a Hirsch number of 6 or higher – only 1.1% of economists achieve this. But 13.8% of economists reach this Hirsch number after 10 year. At that point, 1.6% of economists have a Hirsch number of 11 or higher. After 15 years, 17.5% of economists have a Hirsch number greater than 10, but only 1.3% score 21 or higher. After 20 years, 5.3% of economists have Hirsch number above 20; to be exceptional (1.3%), you need a Hirsch number of more than 25.[5]

The results thus identify exceptional scholars per cohort. Figure 3 shows the densities for cohorts that are not really comparable. Figure 2, on the other hand, allows one to compare people who might apply for the same job. Suppose that one is looking for someone with some ten years of experience, and strives to hire only the top 1%. That implies a Hirsch number of 12 or higher for someone with 10 years of experience, 11 or higher for 9 years of experience, and 15 or higher for 11 years of experience.

One may object that kernel estimators are needlessly complicated. If a lot of data is available, the empirical percentiles are just as informative. Kernel estimators, however, use information from adjacent observations (here, Hirsch numbers from cohorts nearby) and interpolate between observations. This is particularly advantageous when few observations are available. Figure 4 illustrates this with the cohort that published their first paper 25 years ago. Only 173 such economists have registered with IDEAS/RePEc. The histogram and the kernel density (as defined above) thus become multimodal.

This is necessarily the case for the histogram, but not for the kernel density. Equation (5) sets the bandwidth proportional to the number of observations (31,781) to the power one-fifth. Figure 4 shows two alternatives: In one, the bandwidth is proportional to the sample standard deviation; in the other, n in Equation (5) is set to 173. In both cases, the kernel density is smooth (unlike the histogram) and thus closer to our pre-conception of a population density function.

With 173 scholars, the top two roughly form the 99 percentile. See Table 1 for the precise numbers. Their Hirsch numbers are 55 and 49. The narrowest kernel density includes 48 too, so that the top 3 (out to 173) are in the top 1%. The widest kernel finds that the lowest Hirsch numbers are underrepresented and thus lowers the 99 percentile to a Hirsch number of 40. The middle kernel puts the 99 percentile at 45, including 4 economists in the top 1%.

[5] The current author falls in the top 20%.

While the kernel density estimators agree that the empirical percentile is overly strict, there is no agreement on by how much it should be relaxed. This is a common issue in tail estimation. There are few observations of the extremes, and thus little confidence in estimates. That said, even the widest kernel puts the threshold at 40, a Hirsch number that few economists can hope to achieve.

Table 1 also illustrates the advantage of kernel percentile over rank (and empirical percentile). Economists 3 (h=48) and 4 (h=45) are close in rank but further apart in percentile. Economics 2 (h=49) and 3 (h=48) are closer in percentile than in rank.

5. Discussion and conclusion

In this paper, I propose quantile kernel regression as a way of identifying excellent scholars by cohort. Like the crown indicator, the proposed method finds people who stand out – but percentiles have a natural interpretation whereas z-scores do not (unless the distribution is Normal). Like the Hirsch-rate, the proposed method distinguished between people of different age – but kernel regression does not impose linearity. I illustrate the proposed method with a large sample of economists.

The results appear reasonable, but need to be tested still against data from other disciplines, against alternative assumptions on kernel regression, against alternative non-parametric methods, and against parametric methods for quantile regression. This is deferred to future research.

References

Beirlant, J., W.Glaenzel, A.Carbonez, and H.Leemans (2007), 'Scoring research output using statistical quantile plotting', *Journal of Informetrics*, **1**, (3), pp. 185-192.

Burrell, Q.L. (2006), 'Hirsch Index or Hirsch Rate? Some Thoughts Arising from Liang's Data', *Scientometrics*, **73**, (1), 19-28.

Ellison, G. (2010), *How Does the Market Use Citation Data? The Hirsch Index in Economics*, Forthcoming, American Economic Journal: Applied Economics. Working Paper **3188** ,CESifo, Munich.

Falk, M. (1986), 'On the estimation of the quantile density function', *Statistics and Probability Letters*, **4**, (2), pp. 69-73.

Gingras, Y. and V.Lariviere (2011), 'There are neither " king" nor " crown" in scientometrics: Comments on a supposed " alternative" method of normalization', *Journal of Informetrics*, **5**, (1), pp. 226-227.

Hengl, T., B.Minasny, and M.Gould (2009), 'A geostatistical analysis of geostatistics', *Scientometrics*, **80**, (2), pp. 491-514.

Herranz, N. and J.Ruiz-Castillo (2012), 'Sub-field normalization in the multiplicative case: Average-based citation indicators', *Journal of Informetrics*, **6**, (4), pp. 543-556.

Hirsch, J.E. (2005), 'An Index to Quantify an Individual's Scientific Research Output', *Proceedings of the National Academy of Science*, **102**, 16569-16572.

Krichel, T. and C.Zimmermann (2005), 'The economics of open bibliographic data provision', *Economic Analysis and Policy*, **39**, (1), 143-152.

Leydesdorff, L. and T.Opthof (2011), 'Remaining problems with the "New Crown Indicator" (MNCS) of the CWTS', *Journal of Informetrics*, **5**, (1), pp. 224-225.

Liang, L. (2006), 'h-Index Sequence and h-Index Matrix: Constructions and Applications', *Scientometrics*, **69**, (1), 153-159.

Lundberg, J. (2007), 'Lifting the crown-citation z-score', *Journal of Informetrics*, **1**, (2), pp. 145-154.

Moed, H.F. (2010), 'CWTS crown indicator measures citation impact of a research group's publication oeuvre', *Journal of Informetrics*, **4**, (3), pp. 436-438.

Parzen, E. (1979), 'Nonparametric Statistical Data Modelling', *Journal of the American Statistical Association*, **74**, (365), 105-121.

Sarabia, J.M., F.Prieto, and C.Trueba (2012), 'Modeling the probabilistic distribution of the impact factor', *Journal of Informetrics*, **6**, (1), pp. 66-79.

Seiler, C. and K.Wohlrabe (2012), 'Ranking economists on the basis of many indicators: An alternative approach using RePEc data', *Journal of Informetrics*, **6**, (3), pp. 389-402.

Sheather, S.J. and J.S.Marron (1990), 'Kernel Quantile Estimators', *Journal of the American Statistical Association*, **85**, (410), 410-416.

Takeuchi, I., Q.V.Le, T.D.Sears, and A.J.Smola (2006), 'Nonparametric quantile estimation', *Journal of Machine Learning Research*, **7**, pp. 1231-1264.

Takezawa, K. (2006), *Introduction to Nonparametric Regression* Wiley, Hoboken.

Tol, R.S.J. (2013), 'The Matthew effect for cohorts of economists', *Journal of Informetrics*, **7**, (2), pp. 522-527.

van Leeuwen, T.N., M.S.Visser, H.F.Moed, T.J.Nederhof, and A.F.J.van Raan (2003), 'The Holy Grail of Science Policy: Exploring and Combining Bibliometric Tools in Search of Scientific Excellence', *Scientometrics*, **57**, (2), 257-280.

van Raan, A.F.J. (2005), 'Measuring science: Capita selecta of current main issues', in *Handbook of quantitative science and technology research*, H.F. Moed, W. Glaenzel, and U. Schmoch (eds.), Springer, Berlin, pp. 19-50.

van Raan, A.F.J., T.N.van Leeuwen, M.S.Visser, N.J.van Eck, and L.Waltman (2010), 'Rivals for the crown: Reply to Opthof and Leydesdorff', *Journal of Informetrics*, **4**, (3), pp. 431-435.

Waltman, L., N.J.van Eck, T.N.van Leeuwen, M.S.Visser, and A.F.J.van Raan (2011a), 'Towards a new crown indicator: An empirical analysis', *Scientometrics*, **87**, (3), pp. 467-481.

Waltman, L., N.J.van Eck, T.N.van Leeuwen, M.S.Visser, and A.F.J.van Raan (2011b), 'Towards a new crown indicator: Some theoretical considerations', *Journal of Informetrics*, **5**, (1), pp. 37-47.

Yu, K. and M.C.Jones (1998), 'Local linear quantile regression', *Journal of the American Statistical Association*, **93**, (441), pp. 228-237.

Zimmermann, C. (2007), *Academic Rankings with RePEc*, Department of Economics Working Paper Series **2007-36** ,University of Connecticut, Storrs.

Table 1. The top 5 economists of the 25-year cohort: rank, Hirsch number and alternative estimates of their percentile (empirical distribution; kernel density with narrow, middle and wide bandwidths).

Rank	Hirsch	Empirical	Kernel		
			Narrow	Mid	Wide
5	39	97.1	97.3	98.0	98.8
4	45	97.7	98.3	98.9	99.4
3	48	98.3	98.8	99.3	99.6
2	49	98.8	99.2	99.5	99.6
1	55	99.4	99.6	99.8	99.9

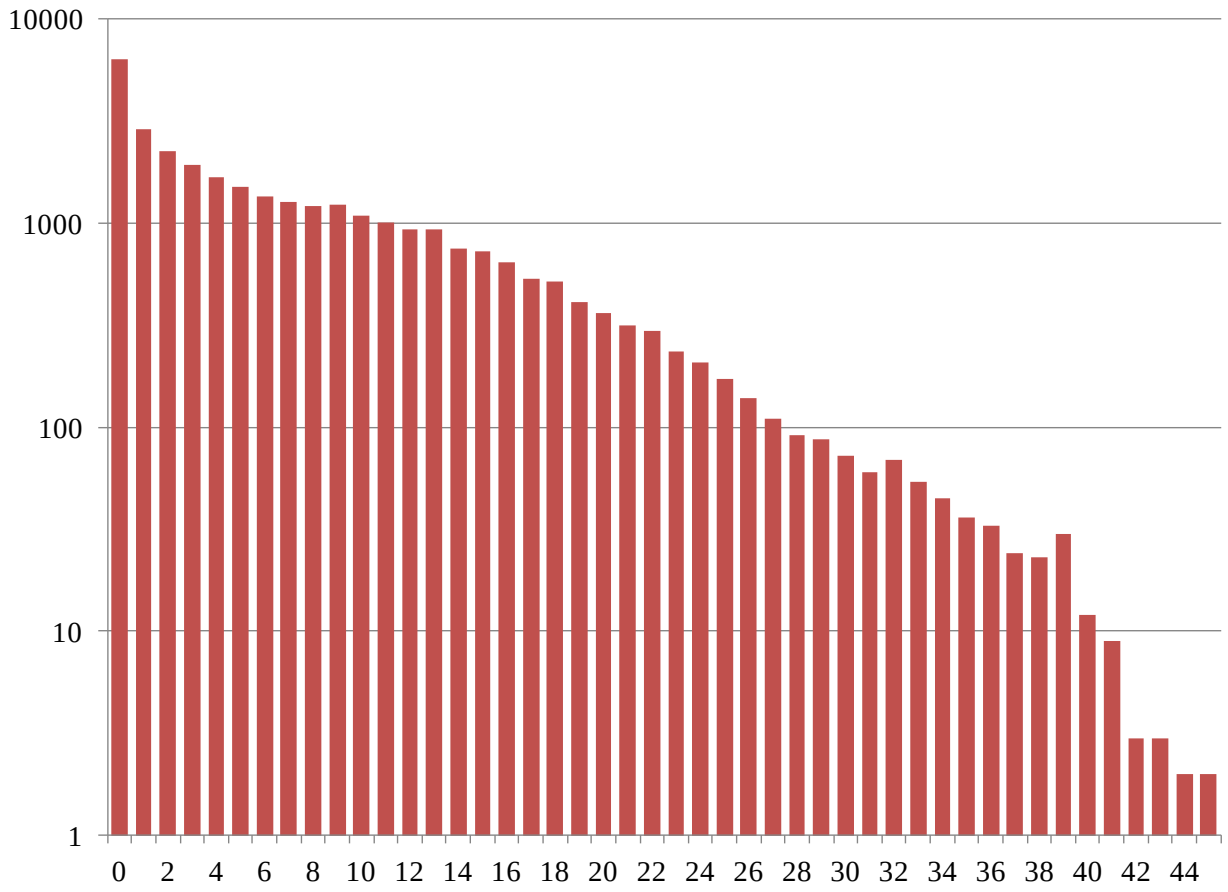

Figure 1. Number of registered economists by cohort.

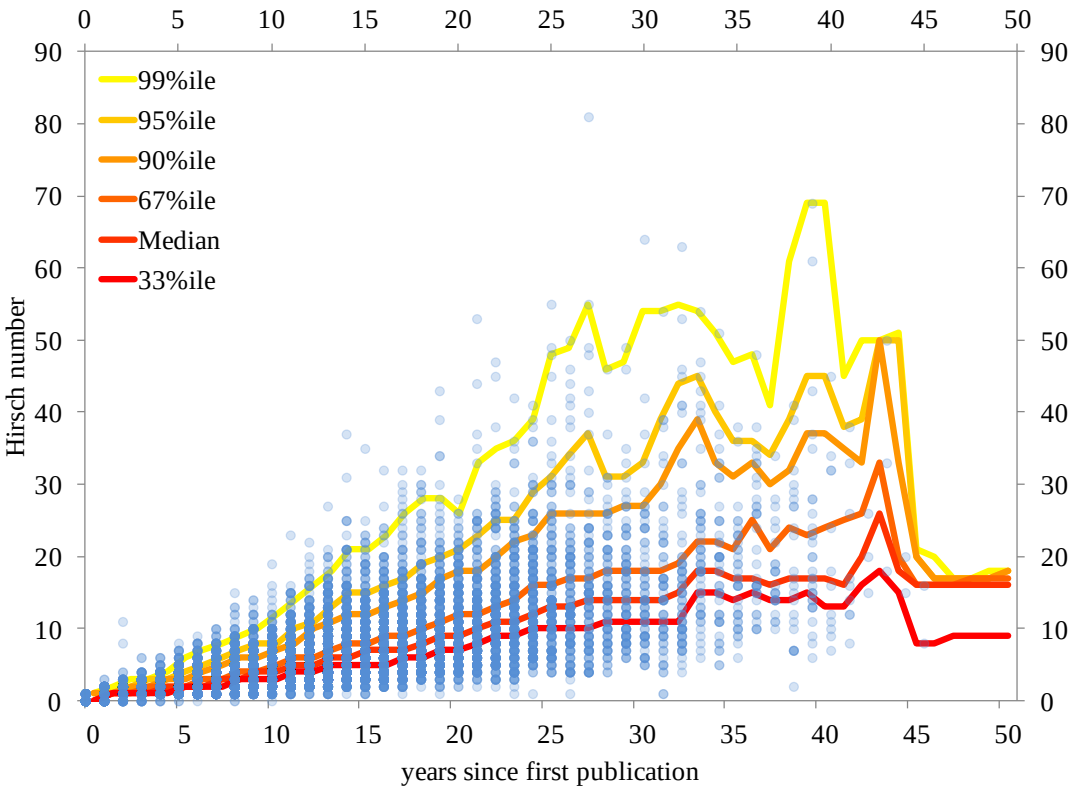

Figure 2. Hirsch number by years since first publication, observed (dots) and selected percentiles of the conditional kernel density.

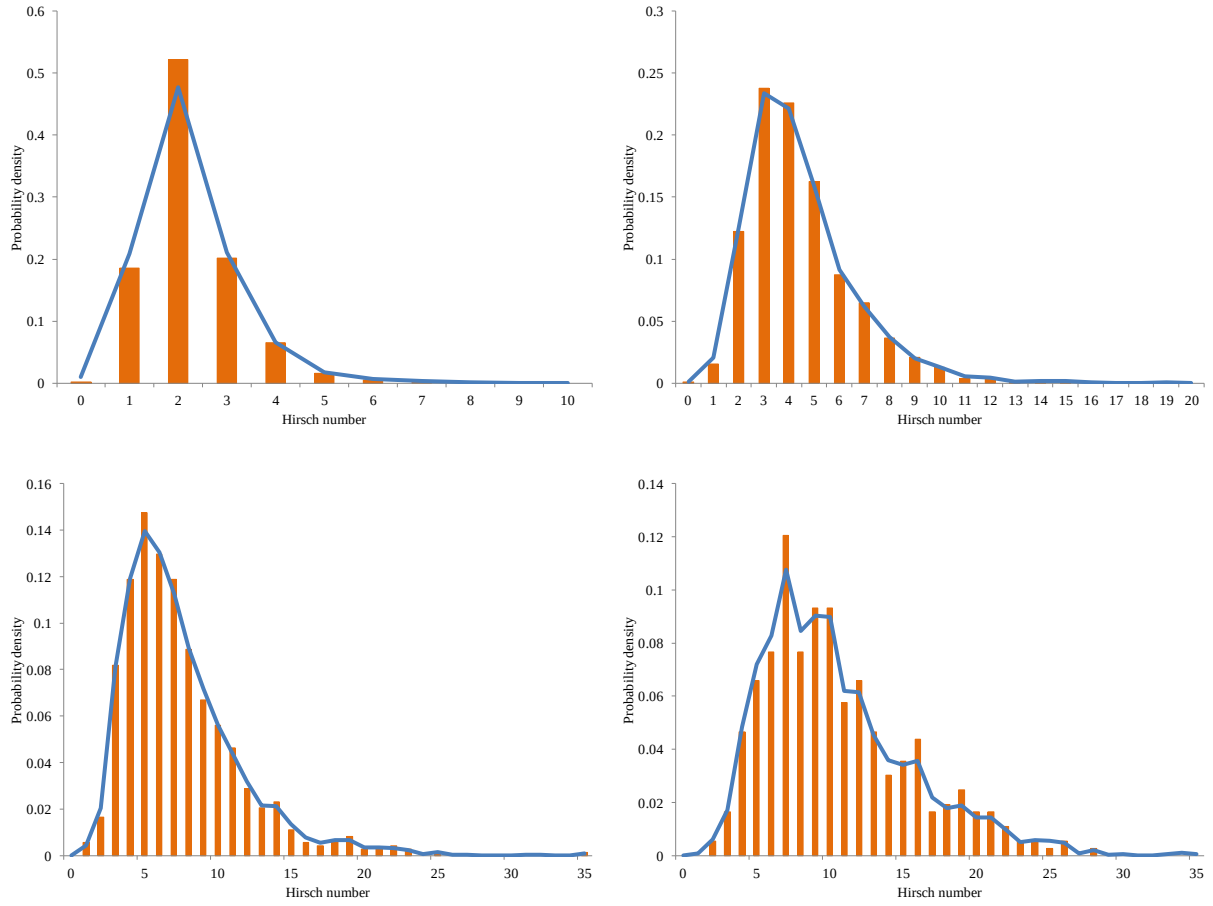

Figure 3. The histogram and conditional kernel density of the Hirsch number for economists who first published 5, 10, 15 and 20 years ago.

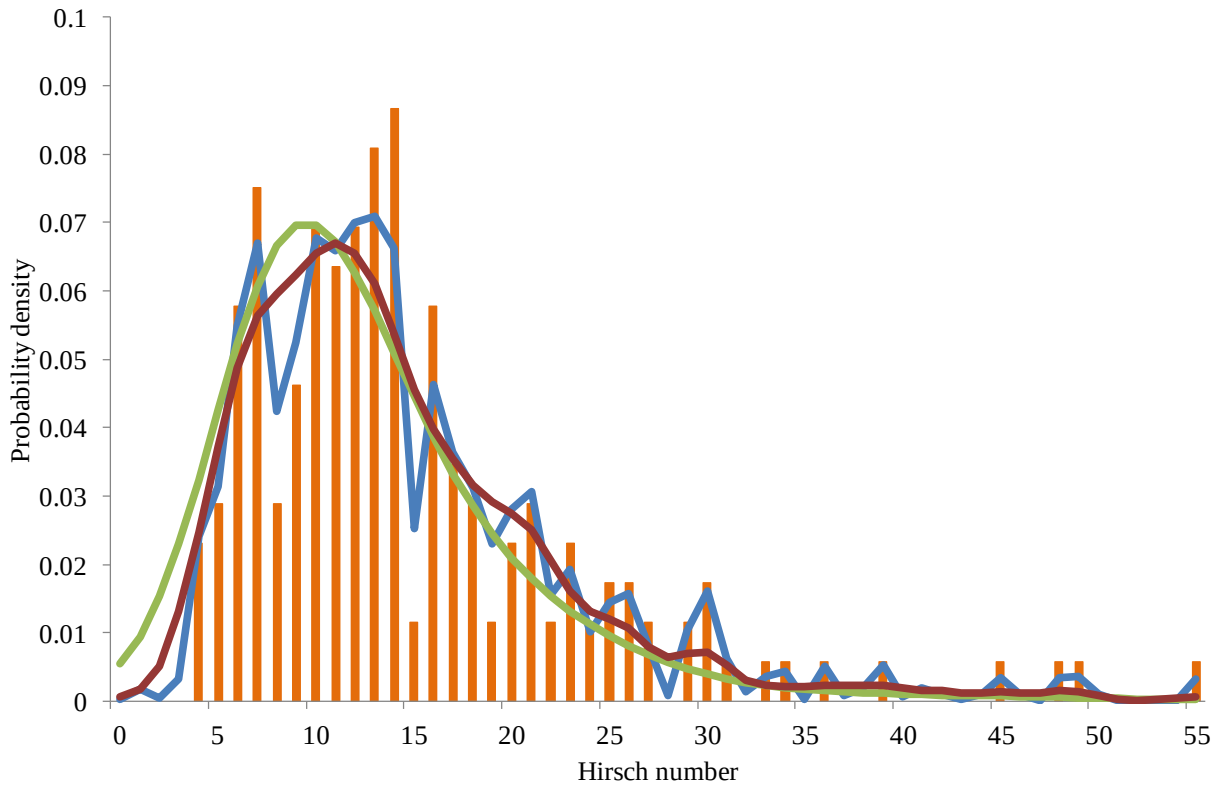

Figure 4. The histogram and the conditional kernel density, for three alternative bandwidths, of the Hirsch number for economists who first published 25 years ago.