December 15, 2018

Martin Paldam

Causality and kernel regressions

Illustrated by data for income and the Polity index

This note deals with the causal implications of kernel-regressions done on data *strings* made by stacking and sorting panel data.

It illustrates the argument by the 6,965 observations of pairs of income, *y*, and the Polity index, analyzed in Paldam and Gundlach (2018).[1] The two variables have the correlation 0.56. Competing theories claim that the correlation is due to either (1) $P \Rightarrow y$ or (2) $y \Rightarrow P$, as per the regressions in Table 2. In addition, some papers claim that the correlation is spurious, so that a variable *x* (such as culture) exists, where $x \Rightarrow y$ and $x \Rightarrow p$. The end of section 4 discusses this possibility.

Even a simple scatter-diagram of the data requires a choice of axes. If it drawn with *y* as the horizontal axis, the researcher suspects that *P* is function of *y*. The reverse drawing implies that the researcher suspects that *y* is a function of *P*. The same is the case for regressions, which see the reverse causality as a problem of simultaneity bias. Most researchers carefully stress that such suspicions are no proof.

A kernel regression is a special graphical regression analysis. Below I argue that when it is done on string data, it strongly reduces the potential for simultaneity:

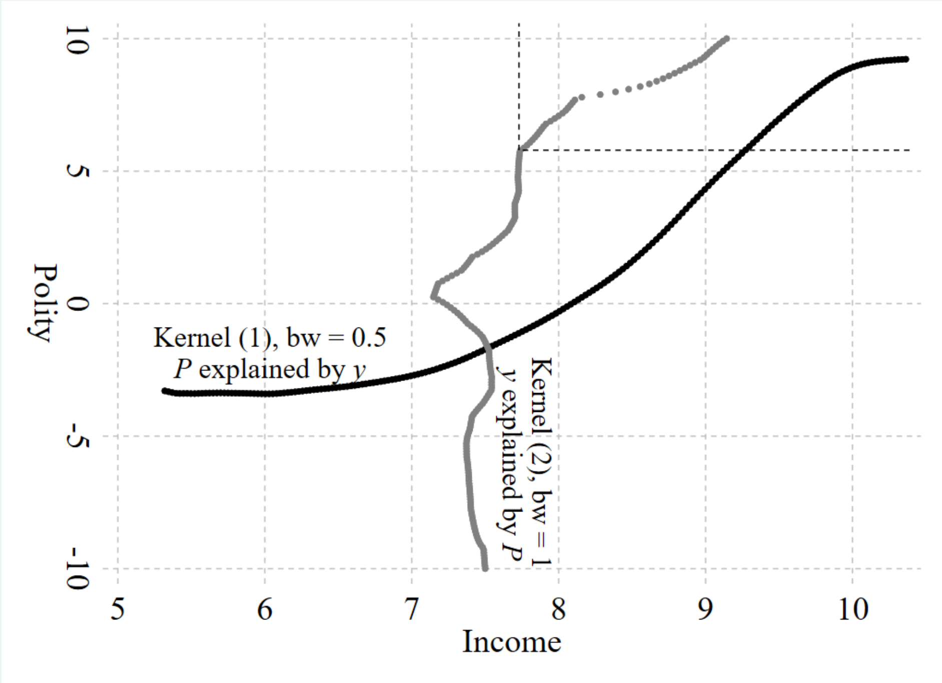
(1) $P_J = \kappa(\bar{y}_j, bw)$, where κ is the kernel with bandwidth *bw*, so that y_j is replaced by $\bar{y}_j$, the mean over all y's in the interval $[y_j – bw/2, y_j – bw/2]$. See kernel (1) on Figure 1.

(2) $y_j = \kappa(\bar{P}_j, bw)$, where κ is the kernel with bandwidth *bw*, so that P_j is replaced by $\bar{P}_j$, the mean over all *P*'s in the interval $[P_j – bw/2, P_j – bw/2]$. See kernel (2) on Figure 1.

A three-step process calculates the kernels – the first two makes the string: (step 1) The panel set (P_{it}, y_{it}) is pooled into (P_j, y_j), which (step 2) are sorted by *y* or kernel (1) or by *P* for kernel (2). (step 3) is the estimate of the kernel, which is a smoothed MA-process, with a fixed *bw* (bandwidth). Figure 1 show how different the two kernels are, except in the small area at the top right corner.

1. The data are Most of the analysis use the *Main* sample.

Figure 1. The reverse kernels (1) and (2), discussed in section 3



It is important for the analysis that the process generating the string gives an effective scrambling so that successive observations from the same country are spread out, and kernels come to contain data for many countries and years. Section 3 report scrambling tests, and concluding that the scrambling is rather efficient. In principle, each $P_j = P_{it}$ may influence y_{it} to y_{it+x}, where x is a number such as 5. Perhaps there may even be repercussions from neighboring countries. So, the potential for reverse causality in the $P_J = \kappa(\bar{y}_j, bw)$ relation depends upon the weight of the observations y_{it} to y_{it+x} in $\bar{y}_j$. Section 3 show that this weight is quite small. This is a main reason why the two kernel-regressions normally looks so different as seen on Figure 1.

If one of the two tells a much clearer story this must say something. If one looks as a weak reflection of the other, the intuition is that the clearest kernel is estimated according to the true causality, while the other is the reflection. One may further reduce the simultaneity bias by lagging the presumed explanatory variable, i.e., one may compare:

(3) $P_J = \kappa(\bar{y}_{j-1}, bw)$, where y is before P.

(4) $y_j = \kappa(\bar{P}_{j-1}, bw)$ where P is before y.

The rest of the paper proceeds as follows: Section 1 looks at the density distributions of the y and the P series, and section 2 show the reverse kernels estimated by the Stata defaults. Section 3 is the analysis of the efficiency of the scrambling. Section 4 show the sensitivity of the kernels to the bandwidth. Sections 2 and 3 discusses if the difference tells a story.

1. The two series: income, *y*, and Polity, *P*

Table 1 gives descriptive statistics for *y* and *P*, while Table 2 reports the reverse regressions. The very high *t*-ratios on the estimated coefficients, are due to the large value of *N*.

Figure 2 shown the density distribution for the *y*-variable. It is calculated by dividing the range in 20 intervals and counting the number observations in each interval. As 6,211/20 = 316, most interval holds many observations. However, for $y < 6$ and for $9.9 < y$ the data are thin.

Figure 3 reports that the Polity index has a rather strange distribution, with peaks at −7 and +10, where −7 is the score of most Communist countries before 1990, and +10 is the score of most Western democracies. The 237 observations for $P = 0$ are not part of the data analyzed in the rest of the paper, and hence displayed differently.

Table 1. Descriptive statistics for the two variables.

N = 6,211	Name	Mean	Std	Min	Max	Correlation
Income, ln to GDP per capita	*y*	7.974	1.106	5.319	10.363	0.559
Polity2 index	*P*	0.829	7.515	-10	10	

Table 2. A linear regression and the reverse

N = 6,211 Regression		(1) Constant *a*	(2) Slope *b*	(3) Reverse		(4) R^2
(1)	$P = a_1 + b_1 y$	-29.4 (-51.1)	3.795 (53.1)	$1/b_2$	12.164	0.312
(2)	$y = a_2 + b_2 P$	7.905 (674)	0.082 (53.1)	$1/b_1$	0.264	

Note: Parenthesis holds t-ratios.

Figure 2. The density distribution for *y* (the natural logarithm to GDP per capita

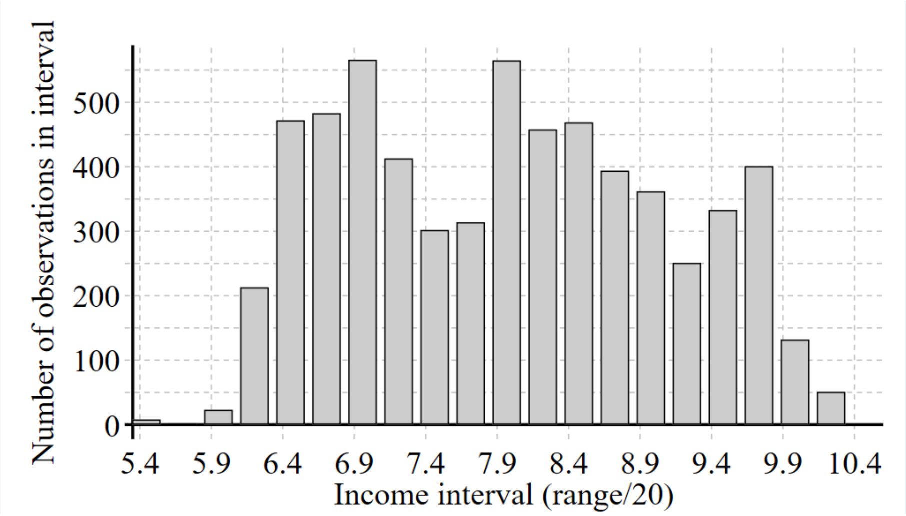


Figure 3. The density distribution for *P*

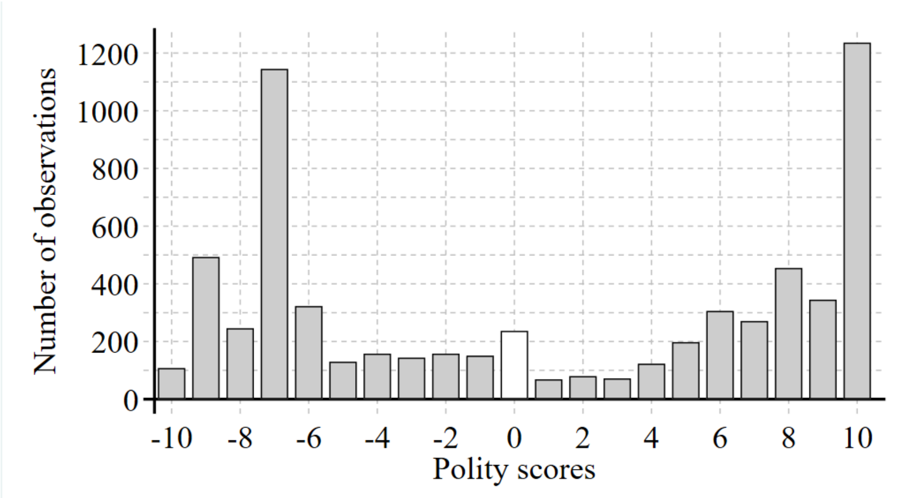


Table 3. A cross tabulation of the two variables

		Income						
		< 5.5	**5.5-6.5**	**6.5-7.5**	**7.5-8.5**	**8.5-9.5**	**> 9.5**	**Total**
Polity	**−10**		2	32	67	7		108
	−9		**114**	**134**	**183**	62		493
	−8		18	**100**	**107**	21		246
	−7		**127**	**597**	**250**	**171**		1,145
	−6		45	**141**	99	32		323
	−5		30	42	37	21		130
	−4		26	69	49	14		158
	−3		6	65	47	26		144
	−2		24	44	50	19	21	158
	−1		42	74	22	13		151
	1	1	21	24	19	4		69
	2		2	47	31			80
	3	1	8	10	34	19		72
	4	1	10	24	71	17		123
	5	3	30	47	93	25		198
	6		42	81	**102**	62	19	306
	7		4	89	**117**	61		271
	8		9	**116**	**149**	**153**	28	455
	9		9	33	**120**	**142**	41	345
	10			23	92	**511**	**610**	1,236
	Total	6	569	1,792	1,739	1,380	725	6,211

Table 3 is a cross-tabulation of the two density distributions. Cells with more than 100 observations are bolded. The two disjoint bolded areas in the table are due to the two-peaked distribution of *Polity*.

2. The stacked and sorted data: How well are strings scrambled?

The introduction argue that strings are effectively scrambled. This is understood in two ways:

(1) The strings does not contain many sequences that are also sequences in the original data, i.e., if $y_j = y_{it}$, then it should be rare that $y_{j+1} = y_{it+1}$, or $y_{j+1} = y_{it+2}$, or … $y_{j+1} = y_{it+5}$

Table 4 count the frequencies of such sequences. They are actually rare.

(2) The numbers of countries and years in each kernel are large.

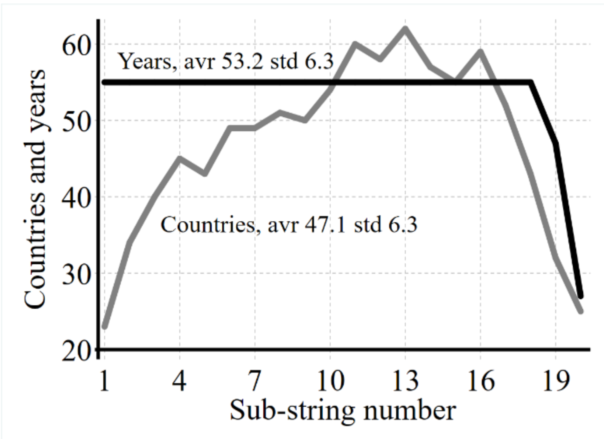
In Figure 5 *bw* = 0.27, which is 5% of the range. 5% of *N* = 6,965, is 348, so that the average number of observations is used for each kernel is 348, but it is fewer when the data are thin. So we divide the string into 22 sub-strings of 348 observations each, where the excess 4 observations are given to the last. Then I count the number of countries and years contained in each substring. The results are reported by Figure 4. There are almost 1/3 of the 155 countries and all years included in each sub-string except in the last two.

I think that both scrambling tests show that the scrambling is satisfactory.

Table 4. The scrambling tests done for string of all *N* = 6,965

Comparing *yj* and y_{j+n}	y_{j+1}	y_{j+2}	y_{j+3}	y_{j+4}	y_{j+5}	Sum
Same country	309	319	312	315	306	-
Next year	157	173	172	119	140	-
Both same	34	26	25	14	6	105
Both in %	0.49	0.37	0.36	0.20	0.09	1.51

Figure 4. The number of countries and years included in the 23 sub-strings



3 The two reverse kernels

Figures 5 and 6 show how the (*P*, *y*)-kernel and the (*y*, *P*)-kernel looks. They are done with the Stata-default: Epanechnikov's kernel and degree zero of polynomial smooth. The program has calculated the optimal kernel by a formula with some chosen parameters.

Figure 5. The (*P*, *y*)-kernel, using lpoly from Stata with the defaults

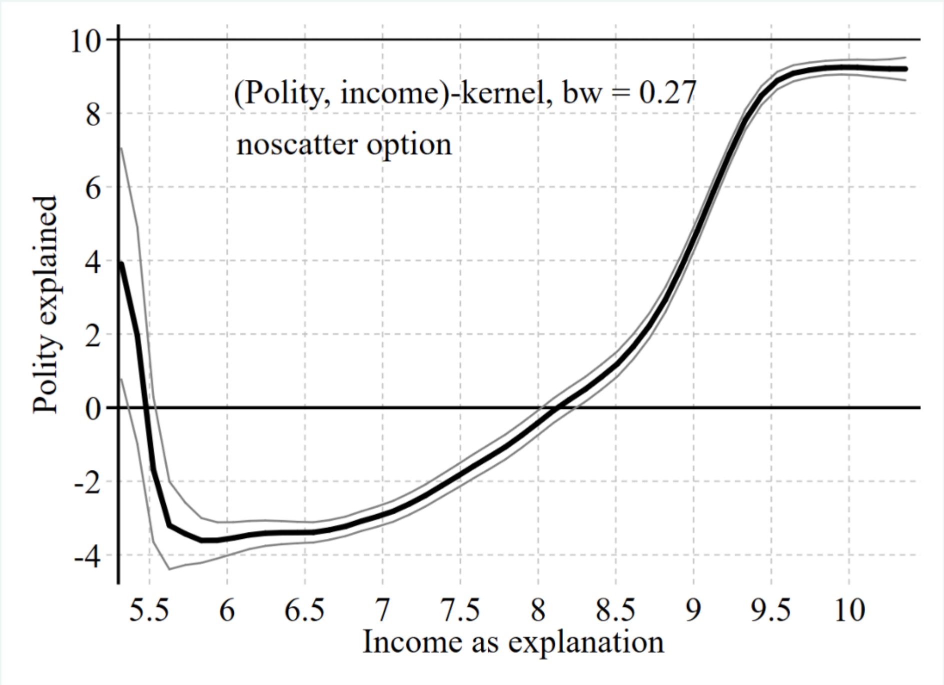


Figure 6. The (*y*, *P*)-kernel, using lpoly from Stata with the defaults

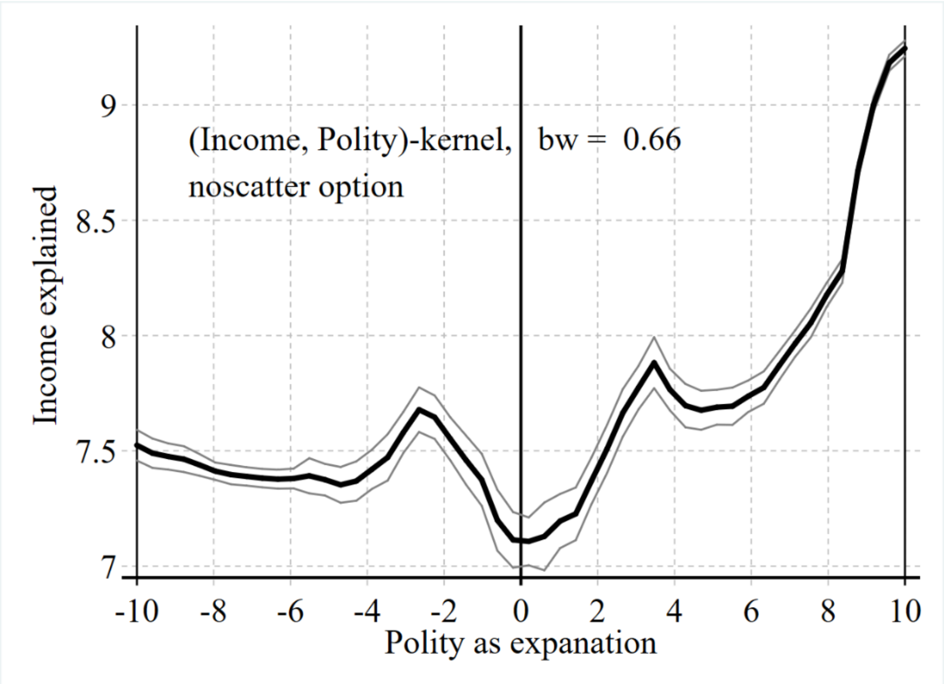


Figure 5 shows the (P, $\bar{y}$)-kernel, where y tries to explain P, while Figure 6 shows the (y, $\bar{P}$)-kernel, where P tries to explain y. The program has chosen the 'optimal' *bw* by a formula that chooses rather low values of the *bw* – just before the curve becomes really wobbly.

The theory that y causes P is the theory of the Democratic Transition. It suggest that the (P, y)-kernel should look as a transition curve. The curve on Figure 5 is an almost perfect transition curve. It is only problematic for $y < 5.8$, but from $y > 5.8$ the curve is as predicted by the theory. The problematic part of the curve occurs for the thin data at the start. As shown in section 3 it is not robust to the bandwidth.

The reverse theory is the Primacy of Institutions theory suggest that the kernel should have a positive slope as well. Figure 6 does not show a clear picture. There is no clear trend for $P < 6$, where the curve has waves that are difficult to explain. The rise for $P > 6$ is somewhat irregular. It is consistent with the idea that the (y, P)-kernel is a weak reflection of the (P, y)-kernel. However, the (P, y)-curve cannot be understood as a weak reflection of the (y, P)-curve.

The two kernels on Figure 1 are the same as the kernels on Figure 5 and 6, though they are drawn with about twice the *bw* for reasons explained in the next section.

4. The sensitivity of the two kernel to the bandwidth, *bw*

Figure 7 shows how the (*P*, *y*)-kernel looks for a broad range of *bw*'s. The transition path appears to be rather robust, while the problematic part of the curve below y = 5.8 is fickle. For all bandwidths, *bw* > 0.3, the low-end problem disappears.

Figure 7. The (*P*, *y*)-kernel (from Figure 5) for a range of bandwidths

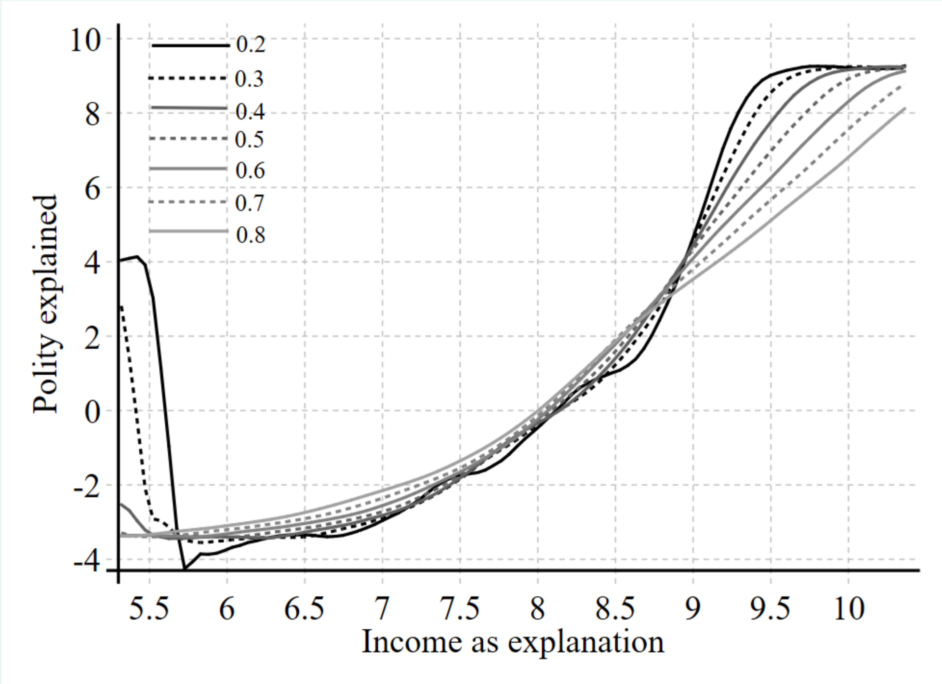
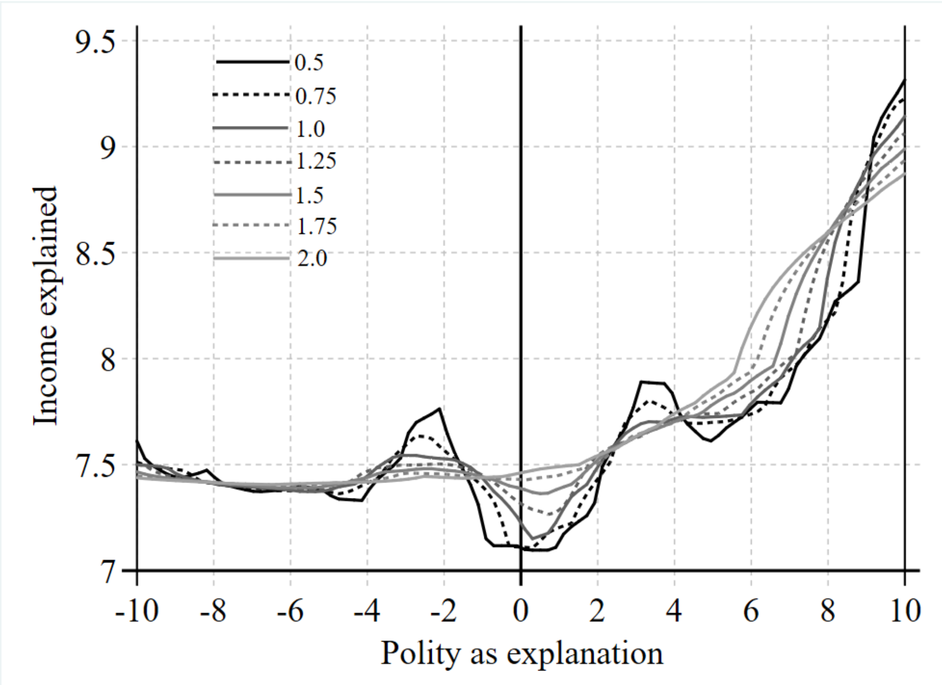


Figure 8. The (*y*, *P*)-kernel (from Figure 6) for a range of bandwidths



The convergence bend at the high end becomes gradually smaller, but it is visible until *bw* exceeds 0.6. For the *bw*-range of 0.4 - 0.6 the kernel looks like a perfect transition curve, and the positive slope for most of the range appears for all *bw*'s shown. I think that it is difficult to imagine that the data could give a better confirmation of the transition theory.

Figure 8 is the reverse (*y*, *P*)-kernel. It is obvious that it presents a picture that is much less clear and hence difficult to interpret. In particular, there is no robust connection from polity to income in the interval form −10 to plus 2. Once again, it is easy to interpret Figure 8 as a weak reflection of Figure 7, but not vice versa.

Finally, I return to the possibility that the correlation between *y* and *P* is spurious, i.e., that a third variable causes both *y* and *P*. If this is the case it is unclear if the (*y*, *P*)-kernel or the (*P*, *y*)-kernel should look better. One suspect that the two kernels will look much the same. The main reason why Figures 4 and 5 look so different is precisely that causality has a dominating direction. Thus, it is arguable that above kernels reject that the (*y*, *P*)-relation is spurious.

5. Conclusion

Thus, I conclude that that a comparison of the (*P*, *y*)-kernel and the (*y*, *P*)-kernel give evidence in favor of the Democratic Transition, and hence support for the idea that *y* causes *P*.

The evidence appears reasonably strong, but it is reassuring that it is supported by a formal causality test in Gundlach and Paldam (2009).

Reference:

Gundlach, E., Paldam, M., 2009. A Farewell to Critical Junctures. Sorting out the Long-Run Causality of Income and Democracy. *European Journal of Political Economy*. 25(3), 340-54

Paldam, M., Gundlach, E., 2018. Jumps into democracy. Integrating the short and long run in the Democratic Transition. *Kyklos* 71(3), 456-81