Supplementary Material

to the article by Peter Turchin and Andrey Korotayev "Population Dynamics and Internal Warfare: A Reconsideration"

England 1450-1800: Population data

Population numbers for the period 1540-1800 were taken from Table A.9.1 in Wrigely et al. (Wrigley et al. 1997). The quinquennial data of Wrigley et al was resampled at a decadal intervals. For the period 1450-1525 population data were taken from Hatcher (Hatcher 1977), also sampled at 10-y intervals (all data analyzed here were sampled at 10-y intervals). The value for 1530 was interpolated.

Detrending the English population data

Agrarian revolution in England started during the seventeenth century (Grigg 1989, Allen 1992, Overton 1996). We can trace this revolution using data on long-term changes in grain yields (Grigg 1989, Overton 1996). Average wheat yields in the thirteenth century were around 10 bushels of grain per acre. Yields declined slightly during the fourteenth and fifteenth centuries to 8 bushels per acre (perhaps as a result of global worsening of the climate). Even as late as 1580s, the yields were still at their late medieval level. During the seventeenth century, however, yields began improving, increasing to ca. 15 in 1700 and 20-21 in the early nineteenth century (Grigg 1989:69). Net yields (subtracting seed corn) were lower. For example, the typical late medieval seeding rates were 2 bushels per acre; thus, the net yield was only 6 bushels per acre. Net yields from Grigg and Overton are plotted in Figure 1. To capture the rising trend, I fitted the data after 1580 with a straight line (see Figure 1a, note the log-scale). It is possible that the jump in yields was more discontinuous than is suggested by a straight line. However, the authorities arguing for an abrupt improvement in yields disagree about its timing. Thus, Allen (1992) presented evidence suggesting that most of increase in yields was accomplished by 1700, while Overton (1996) argued that the decisive breakthrough actually occurred after 1750. Given this controversy and the degree of scatter in the yield data I decided against fitting a more complex nonlinear relationship.

We can obtain an approximate estimate of the carrying capacity by assuming that it was proportional to the net yield. Assuming the total potentially arable area of 12 mln acres (Grigg 1989) and that one individual (averaging over adults and children) needs a minimum of one quarter (8 bushels or 2.9 hectoliters) of grain per year, I calculated the carrying capacity of England shown by the broken line in Figure 1 (by coincidence 1 bushel of net yield per acre translates exactly into 1 million of carrying capacity).

We can now detrend the observed population numbers by dividing them with the estimated carrying capacity. The detrended population, which can also be thought of as "population pressure on resources" is defined as $N'(t) = N(t)/K(t)$. Note that the critical assumption here is that K is proportional to the net yield, Y; since Y is the only quantity varying with time in the formula, other components (total arable area, consumption minimum) being constant multipliers, K will wax and wane in step with Y. In other words, the exact values of constant multiplies do not matter, since we are interested in relative changes of population pressure. Note that the estimate of K is based not on the area that was actually cultivated (this fluctuated up and down with population numbers), but on the potentially arable area. The latter quantity fluctuated little across the centuries (for example, as a result of some inundation of coastal areas during the Middle Ages or more recent reclamation using modern methods) and can be approximated with a constant without a serious loss of precision.

A test of the appropriateness of this detrending was obtained by regressing the estimated population pressure on real wages reported by Allen (2001). There was a very close inverse relationship between these two variables, and not a very good one if we were to use the non-detrended population numbers. As Figure 1b shows, population pressure and inverse real wage fluctated virtually in perfect synchrony.

England 1450-1800: sociopolitical instability data

For the period 1492-1800 I used the list of civil wars and rebellions compiled by Tilly (1993:Table 4.2). The list reports on revolutionary situations in all British polities. Since my focus is on England, I excluded all rebellions in Ireland, as well as in Scotland prior to the unification under the Stuarts. For the period prior to 1492, I used the compendium of Sorokin (Sorokin 1937: Appendix to Part III), which essentially added the data on the Wars of the Roses. The complete list is given in Table 1.

Table 1. Civil wars and rebellions in England 1450-1800.

years	*Description*
1455-6	The Wars of Roses: 1st phase
1460-5	The Wars of Roses: 2nd phase
1467-71	The Wars of Roses: 3rd phase
1483-5	The Wars of Roses: 4th phase
1495	Rebellion of Perkin Warbeck
1497	Insurrection in Cornwall
1536-7	Pilgrimage of Grace
1549	Kett's rebellion
1554	Wyatt's rebellion
1569	Rebellion of catholic lords of the North
1639-40	The Bishops' Wars
1642-7	Civil War
1648-51	Second Civil War
1655	Penruddock rising in Salisbury
1660	Monk's coup; restoration of James II
1666	Revolt of Scottish Covenanters
1679	Revolt of Scottish Covenanters
1685	Monmouth and Argyll rebellions
1687-92	Glorious Revoultion, with intervention by France
1715-6	Jacobite rebellion in Scotland
1745-6	Scottish rising (Jacobite pretender)

Smoothing socio-political instability

I constructed an index of sociopolitical instability by assigning "1" to years with rebellion or civil war and "0" to years without (Boswell and Chase-Dunn 2000). To translate this discontinuous index into a smoothly varying one, I used the kernel regression. The kernel regression is a nonparametric function estimator. The degree with which the estimated curve interpolates vs. smoothes over a scatter in the data is determined by a single parameter *h* called the bandwidth (Härdle 1990). I used an exponentially weighted kernel (that is, the contribution of a data point to the smoothed point declines exponentially with distance between the two points). The choice $h = 50$ was determined by a prior empirical observation (see Turchin 2003: Section 9.1.2) that the times of trouble in European polities tended to be "lumpy" on a human generation scale. That is, revolutions and bouts of civil war tend to skip generations: if fathers participate in bitter internal fightings, their sons tend to value stability at almost any cost, while the grandsons exhibit a renewed willingness to revolt. Interesting as this pattern may be, it is a different phenomenon from the one we are investigating (the average periodicity of this lumpiness is two human generations, or 50 years, compared to 200-300 year secular cycles). Using the bandwidth of 50 y smooths out any bigenerational cycles that may be present in the data, and allows us to focus on the secular oscillations.

China: Population data

The situation with population data for China is complex. On one hand, the central authority in China (when it existed), conducted detailed censuses for tax purposes. On the other hand, corrupt or lazy officials often falsified or fabricated population data (Ho 1959). Conversion coefficients between the number of taxable households and the actual population are often unknown, and what is worse, these coefficients probably changed from dynasty to dynasty. The area controlled by the state also continually changed. Finally, it is often difficult to determine whether the number of taxable households declined during the times of trouble as a result of demographic change (death and emigration), or as a result of the state's failure to control and enumerate the subject population. Thus, there is a certain degree of controversy among the experts as to the precise levels that population numbers achieved at the highs and lows (Ho 1959, Durand 1960, Song et al. 1985). However, the controversy primarily concerns the absolute population levels, and there is a substantial degree of agreement on the *relative* changes in population density (which are, of course, of primary interests to our purposes).

The most detailed trajectory of population dynamics in China, known to me, was published by Chao and Hsieh (Chao and Hsieh 1988). These authors give estimates of Chinese population numbers at irregular time intervals. In order to make the data suitable for time-series analysis, I interpolated Chao and Hsieh data using an exponential kernel with bandwidth of 10 years, and then subsampled the resulting smoothed trajectory at 10 year intervals. Setting h = 10 y, same as the sampling interval, results in a minimal smoothing of the data.

After 1000 the trajectory becomes clearly nonstationary, and requires detrending. For this reason, I focus on the pre-1000 data (analysis of the second millenium data will be reported in Turchin and Nefedov, work in progress). After the fall of the East Han dynasty and before the Sui re-unification China was divided among a number of warring states. I excluded this period because the demographic-structural theory is state-centered (rather than focusing on state systems, as China was during the Han-Sui interregnum). This gave me two periods (using centuries as convenient break-points): 200 BCE – 300 CE and 600 – 1000 CE.

China: Instability data

The index of sociopolitical instability in China comes from a remarkable publication by J. S. Lee (Lee 1931), who during the 1920s set out to calculate the frequency of internecine wars in Chinese history (ranging from fairly localized uprisings to wide-spread rebellions and civil wars). For the period of interest to us (up to year 1000) Lee largely extracted his data from the Tih Wang Nien Piao by Chih Shao-nan. Checks with independent sources demonstrated the high accuracy of this source (Lee 1931:114). Lee presented the data as counts of internecine wars per 5-year interval. I smoothed his data using an exponential kernel with bandwidth h = 30 y, and resampled the data at 10-y intervals. I reduced the bandwidth (compared to h = 50 y used in the analysis of the English data) because the Chinese data did not appear to exhibit bi-generational cycles. In general, Chinese dynamics operated on a faster time scale, so using a bandwidth of 50 y would result in oversmoothing (I redid the analyses with h = 50 y and the results were qualitatively the same).

Rome: population and instability data

Population history of the Roman Republic and Empire remains a highly contentious topic (Scheidel 2001). Archaeological data, however, begins to throw light on this obscure aspect of Roman history. Recently Lewit (1991) integrated the results from numerous archaeological sites within the Western Empire and presented summaries indicating the proportion of archeological sites occupied in a 50-year period for Britain, Belgica, Northern and Southern Gaul, Northern and Southern Spain, and Italy. The data suggest that there were two periods of settlement expansion and two periods of settlement abandonment. I constructed a crude index of population dynamics by averaging provincial occupation curves (see Figure 4 of the main publication).

Data on internal warfare in the Roman Empire was published by Sorokin (1937:Table 26). The data points are given for each 25-y interval. Smoothing the data using kernel with $h = 50$ y reveals two periods of intense sociopolitical instability. One is the first century BCE, which was a period of transition between the Republic and Empire. During the first half of the Principate (the Early Roman Empire), after internal warfare subsided, population exhibited a long period of sustained growth. The population peak was achieved around 200. The second period of instability occurred during the third century, when the Empire was convulsed by a series of internal wars, which were accompanied by population decline. Another period of stability and population growth occurred during the first half of the Dominate (the fourth century). After the decline and fall of the Roman Empire in the West, population decreased. Note that Sorokin's index of internal warfare underestimates the extent of actual sociopolitical instability during the fifth century, because he treated barbarian invasions as external warfare.

The population data are too crude to analyze using standard time-series methods (the main problem is the length of the sampling period, 50 years). Thus, I did not fit models with the population index as the dependent variable, but only used it as an independent variable in the analysis of sociopolitical instability. The population data were smoothed using an exponential kernel with bandwidth $h = 30$ y, and resampled at 10-y intervals.

Analysis: regressions

Prior to analysis I log-transformed all data: $X(t) = \log N(t)$ and $Y(t) = \log W(t)$ where $N(t)$ and $W(t)$ are population and internal war data. As explained above, English population data were detrended by calculating population pressure. English internal war data were also non-stationary (see Figure 1b of the main publication). I detrended instability data by calculating $Y'(t) = Y(t) - (a_0 + a_1 t)$, where a_0 and a_1 are parameters of linear regression of $Y(t)$ on t. (This is equivalent to dividing the untransformed data by the temporal trend.)

I fitted a simple time-series model to the data, the linear autoregressive process

$$X(t) = a_0 + a_1 X(t-\tau) + a_2 Y(t-\tau) + \varepsilon_t \qquad (1)$$

(and an analogous model for $Y(t)$). Here a_i are parameters to be estimated, and ε_t is an error term, assumed to be normally distributed (Box and Jenkins 1976). The time delay was chosen as $\tau = 30$ y, which approximates a human generation length. This particular time delay is also a reasonable choice that optimizes the tension between *redundancy* and *irrelevance* (see Turchin 2003: Section 7.2.2). To check how my conclusions were affected by the specific value of the time lag, I fitted all models using an alternative choice of $\tau = 20$ y, and obtained essentially same results.

As another check I fitted a model that used a quadratic polynomial:

$$X(t) = a_0 + a_1X(t–\tau) + a_2Y(t–\tau) + a_{11}[X(t–\tau)]^2 + a_{22}[Y(t–\tau)]^2 + a_{12}X(t–\tau)Y(t–\tau) + \varepsilon_t$$

The purpose of this model was to determine whether the process had a strong nonlinear component. There was statistical evidence for nonlinearity in some series, but using the quadratic model instead of the linear one (where it fit better) did not change any conclusions discussed below, so I do not report these results here.

To quantify any reciprocal effects of population and instability on each other I employed the stepwise regression. Thus, in order to estimate the effect of instability on population change, I first regressed $X(t)$ on $X(t–\tau)$, and then tested whether adding the term $Y(t–\tau)$ significantly reduced unexplained variance. The effect of population density on instability was investigated by regressing $Y(t)$ on $Y(t–\tau)$, and then adding the term $X(t–\tau)$. The F-statistics associated with each test are reported in Table 1 of the main publication.

Analysis: cross-validation

The ultimate test of any model is its ability to predict independent data (data that were not used to develop the model and estimate its parameters). To assess the ability of the demographic-structural model to predict out-of-sample data I split each data set into two equal-sized parts. Then I fitted model (1) to the first half (the "fitting set") and used the estimated coefficients to predict each data point in the second half (the "testing set"). Thus, the predicted population values, X^*, (the asterix denotes prediction) were calculated as follows:

$$X(t)^* = a_0 + a_1X(t–\tau) + a_2Y(t–\tau)$$

where $X(t–\tau)$ and $Y(t–\tau)$ are the observed values of the independent variables in the *second half*, while the parameters a_0, a_1, and a_2 were estimated using data in the *first half*. The correspondence between the observed $X(t)$ and predicted $X(t)^*$ was assessed by linear correlation.

After using the first half of the data set to predict the second, I reversed the procedure and used the second half to predict the first. This procedure allowed me to use the complete data set for testing the model performance. Finally, I repeated the complete procedure with the instability data (Y).

One possible objection to the procedure outlined above is that there is some positive autocorrelation between $X(t)$ and $X(t–\tau)$ due to the time-series nature of the data, and it is conceivable that the excellent correlations between the observed $X(t)$ and predicted $X(t)^*$ are entirely due to this "inertial" effect. To eliminate this possibility, I redid the analyses with a different dependent variable, $\Delta X(t) = X(t) – X(t–\tau)$. $\Delta X(t)$ is a measure of the rate of change, and by using it we break the autocorrelation arising from the time-series nature of the data. In fact, $\Delta X(t)$ is none other than the realized per capita rate of population change, which is the standard dependent variable in the analyses of population data (Turchin 2003). There can still be some predictive relationship between $\Delta X(t)$ and $X(t)$, so we need to compare two alternative models:

$$\Delta X(t) = a_0 + a_1X(t–\tau) + \varepsilon_t \quad (2)$$

which I call the *inertial* model, and

$$\Delta X(t) = a_0 + a_1X(t–\tau) + a_2Y(t–\tau) + \varepsilon_t \quad (3)$$

which I call the *interactive* model. The interactive model has an extra parameter, but in a cross-validation setting this does not matter (if the extra independent variable does not have a systematic influence on the dependent variable, then adding it to the model actually *decreases* to the ability of the model to predict out-of-sample data).

Results presented in Table 1 show that only in one case out of twelve (indicated by italics in the table) the inertial model does better than the interactive one. In all other cases the prediction accuracy is substantially increased by using the interactive model. In fact, in half of the cases the correlation coefficient between the observations and predictions made by the inertial model is not significantly positive. In the English case the intertial model does so poorly that the correlations between the predictions and observations are actually negative. Thus, knowledge of population dynamics significantly increases the ability to predict instability, and vice versa.

Table 1. Comparing out-of-sample predictive abilities of the inertial and interactive models

Source of data	Dependent variable	Correlation between predicted and observed			
		1st half => 2nd half		2nd half => 1st half	
		inertial	interactive	inertial	interactive
England	population	–0.57	0.94	–0.07	0.44
England	instability	–0.13	0.80	–0.53	0.89
Han China	population	0.45	0.57	*0.73*	*0.48*
Han China	instability	0.39	0.87	0.37	0.68
Tang China	population	0.56	0.80	0.61	0.90
Tang China	instability	0.57	0.78	0.66	0.92

Literature Cited

Allen, R. C. 1992. Enclosure and the yeoman. Clarendon Press, Oxford, UK.

Allen, R. C. 2001. The great divergence in European wages and prices from the Middle Ages to the First World War. Explorations in Economic History **38**:411-447.

Boswell, T., and C. Chase-Dunn. 2000. The spiral of capitalism and socialism: toward global democracy. Lynne Rienner, Boulder, CO.

Box, G. E. P., and G. M. Jenkins. 1976. Time Series Analysis: Forecasting and Control. Holden Day, Oakland, CA.

Chao, W.-L., and S.-C. Hsieh. 1988. History of Chinese Population (in Chinese). People's Publisher, Peking.

Durand, J. D. 1960. The population statistics of China, AD 2-1953. Population Studies **13**:209-256.

Grigg, D. 1989. English agriculture: an historical perspective. Blackwell, Oxford, UK.

Härdle, W. 1990. Applied nonparametric regression. Cambridge University Press, Cambridge, UK.

Hatcher, J. 1977. Plague, population and the English economy: 1348-1530. Macmillan, London, UK.

Ho, P. T. 1959. Studies on the Population of China: 1368-1953. Harvard University Press, Cambridge, MA.

Lee, J. S. 1931. The periodic recurrence of intrnecine wars in China. The China Journal **(1931: March-April)**:111-163.

Lewit, T. 1991. Agricultural productivity in the Roman economy A.D. 200-400. Tempus Reparaturm, Oxford, UK.

Overton, M. 1996. Agricultural revolution in England. Cambridge University Press, Cambridge.

Scheidel, W., editor. 2001. Debating Roman demography. Brill, Leiden, Netherlands.

Song, J., C. H. Tuan, and J. Y. Yu. 1985. Population Control in China: Theory and Applications. Praeger Scientific, New York.

Sorokin, P. A. 1937. Social and cultural dynamics. Vol. III. Fluctuations of social relationships, war, and revolution. American Book Company, New York.

Tilly, C. 1993. European revolutions: 1492-1992. Blackwell, Oxford, UK.

Turchin, P. 2003. Complex Population Dynamics: A Theoretical/Empirical Synthesis. Princeton University Press, Princeton, NJ.

Wrigley, E. A., R. S. Davis, J. E. Oeppen, and R. S. Schofield. 1997. English population history from family reconstruction: 1580-1837. Cambridge University Press, Cambridge, UK.

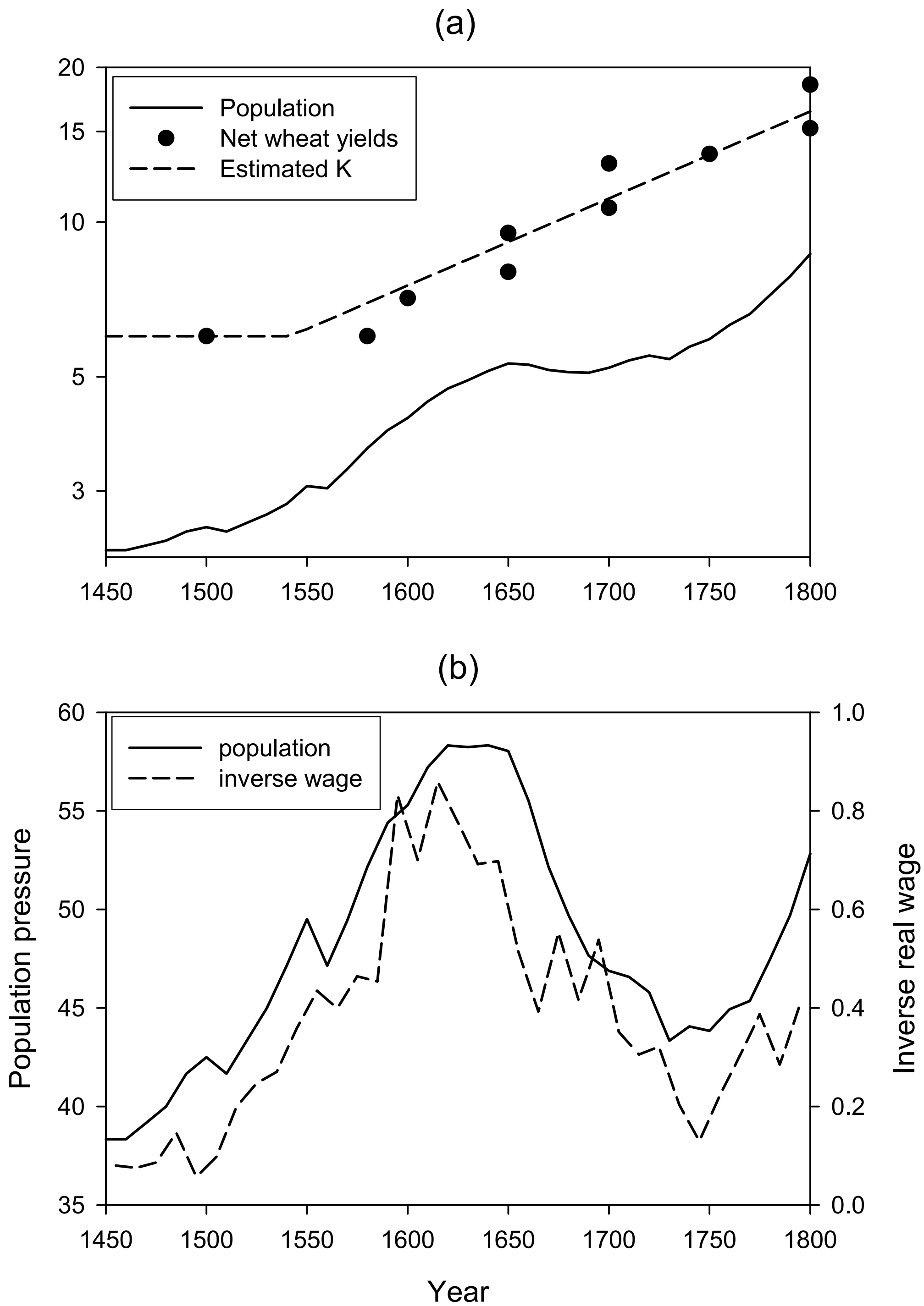

Figure 1. Detrending population trajectory for England. (a) Population numbers (in million), net yields (in bushels per acre), and the estimated carrying capacity (in million of people) in England from 1450 to 1800 (all variables plotted on a log-scale). (b) Detrended population ("population pressure") trajectory (solid curve) and inverse real wages (broken curve).