The Economic Origins of the Territorial State
Supplemental Online Appendix

Scott F Abramson
The University of Rochester
sabramso@ur.rochester.edu

Supplemental Appendix

Overview of Appendix

The appendix to "Economic Origins of the Territorial State" includes a number of robustness checks and additional results not included in the main manuscript. It is divided in three sections.

1. The first section of the appendix provides additional graphical and statistical results indicating that the distribution of state size is, across all time periods, log-normal.

    - I reproduce Figure II from the main text but instead of pooling all observations I do it for the years 1100, 1200, 1300, 1400, 1500, 1600, 1700, and 1790, separately. The figure shows graphically that the untransformed distribution is heavily skewed and that the log-transformed data take on a roughly normal distribution
    - I produce Q-Q plots comparing the untransformed and log transformed the distributions of state size which furthermore indicate the distribution of state size is log-normal.
    - I conduct a Box-Cox transformation to see if state size is better characterized by an alternative power-transformation. This procedure produces a parameter estimate of $\lambda = -.099$ which is very close to zero, indicating a log-transformation is appropriate.

2. The second section of the appendix provides evidence that there was no change in the distribution process by which states were formed and failed over time. I use a series of OLS, change-point, and EM methods to show this.

    - *Change Point Analysis of the Number of States Over Time*

- I conduct an F-test based method of change-point detection which provides evidence that no structural break in the data describing the number of European states took place after 1455, and places the greatest likelihood of a structural break in the year 1210.
- I conduct the Bayesian change-point analysis proposed by Park (2010) which finds evidence of just a single structural break in the number of states over time, again in the year 1210
- I conduct a maximum likelihood change-point analysis which identifies a slightly larger number of breaks but, nevertheless, finds no evidence of a change-point in the number of states after 1260.

- *OLS Time-Series Methods*
  - As an alternative to assessing the possibility of a structural break in the number of states, I model this quantity as an AR process. I show that the change in the number of states follows a series random, mean zero, set of shocks.
  - I partition the time series into two periods - one before 1500 and one after - and reestimate the above model. I show that the results are nearly identical.
  - I conduct a change-point test on the AR model and find no evidence in favor of any structural break in the process determining changes in the number of states.

- *OLS Panel Methods*
  - Rather than looking at just changes in the longitudinal component describing the number of states, we might want to see if there was a change in the data generating process determining state size. To do this, I first estimate a simple panel model estimating the "time effect," capturing the first moment of the

distribution in each year we observe data.

  - Then, I show that these time effects for every year after 1500 - the period we expect big changes in the distribution of state sizes - are statistically indistinguishable from each other.
  - To gain a picture of changes in the second moment of the distribution I examine changes in the residuals from these regressions. While the number of outlying cases increases after 1450, these constitute far less than 5% of the data. Moreover, in line with the theory presented in the paper these cases are the peripheral, outlying, states that much of the literature has relied upon to inductively build theory.

- *EM Algorithm*
  - To assess how the distribution of states changed over time I next treat the problem as one of latent classification and use a simple Expectation-Maximization (EM) algorithm to classify each cross-section. What results is a choice of three ordered groups. Again, however, all years from 1485 onwards are classified as belonging to the same group. Again, in the period we expect large changes in the distribution of state sizes, we find no evidence of substantial change.
- *Difference Over Time in The Effect of State Size on Survival*
  - I compare the time-varying coefficients from the survival models presented in Figure III. I show that the relationship between size and survival is statistically indistinguishable across any pair of time periods.

3. The third section in the appendix details the construction and logic of the instrument and, in addition, provides a series of additional robustness checks to the instrumental variables analysis.

- I begin by outlining the paleo-climatic sources used to construct the instrument. Then, using 20th century data on outcomes directly related to wheat production I show that the logic of the instrument holds.
- I provide instrumental variables results which show that the main effects are robust to the inclusion of regional effects and the exclusion of Central Europe.
- I outline the method used to construct grid-square estimates of per-capita income and show that the main results are robust to the inclusion of these estimates.
- I show that the main results are robust to controlling for the number of cities in each grid-square.
- Using the sensitivity analysis proposed by Conely et al (2012) I find that the instrumental variables results are robust to a substantively large violation of the exclusion restriction.
- I show that the effect of urban population operates through a mechanism linking urban growth to political fragmentation through the presence of particular social groups, specifically those associated with commerce and industry. To accomplish this I exploit data on the presence of proto-industrial activity to show via the mediation procedure proposed by Imai, Keele, and Tingley (2010), Imai, Keele, Tingley et al. (2011), Imai, Tingley, and Yamamoto (2013) to show that the effect of urban population operates through the presence of commercial groups associated with these activities.

The Log-Normality of State Size

In the main text of the paper I provide evidence that the distribution of state size is log-normally distributed. In this section of the appendix I provide further support for this assertion.

Reproducing Figure II At Century Intervals

To begin I simply recreate Figure II, but instead of pooling the data across all years I look at just the years 1100, 1200, 1300, 1400, 1500, 1600, 1700, and 1790. Figure A1 plots in the top row the untransformed distribution at these points in time and in the lower row the log-transformed distribution. Again, what we see is that, across time, the untransformed distribution is heavily skewed. However, when we log-transform the entire data, in each period, they take on a roughly normal distribution.

QQ Plots

To gauge the appropriateness of the log-normal distributional assumption I plot the true sample quantiles of the untransformed and log-transformed data against quantiles from a hypothetical normally distributed random variable with the same mean and variance as the observed data, respectively. If state size were normally distributed we would expect the sample quantiles to be on the line $y = x$. Figure A2 gives these Quantile-Quantile (QQ) plots of both the untransformed state size in the upper panel and the log-transformed QQ plot in the lower panel. As expected when we take the log-transformed state size the QQ plots rest almost perfectly on the 45 degree line, a good indication that the data is, indeed, log-normal.

Box-Cox Transformation

We can consider another transformation of the data which like the log-transformation is designed to preserve rank, properly weight outlying cases, and create an approximately normal distribution, the Box-Cox power-transformation.[1] This transformation is as follows:

[1] Box and Cox, 1964

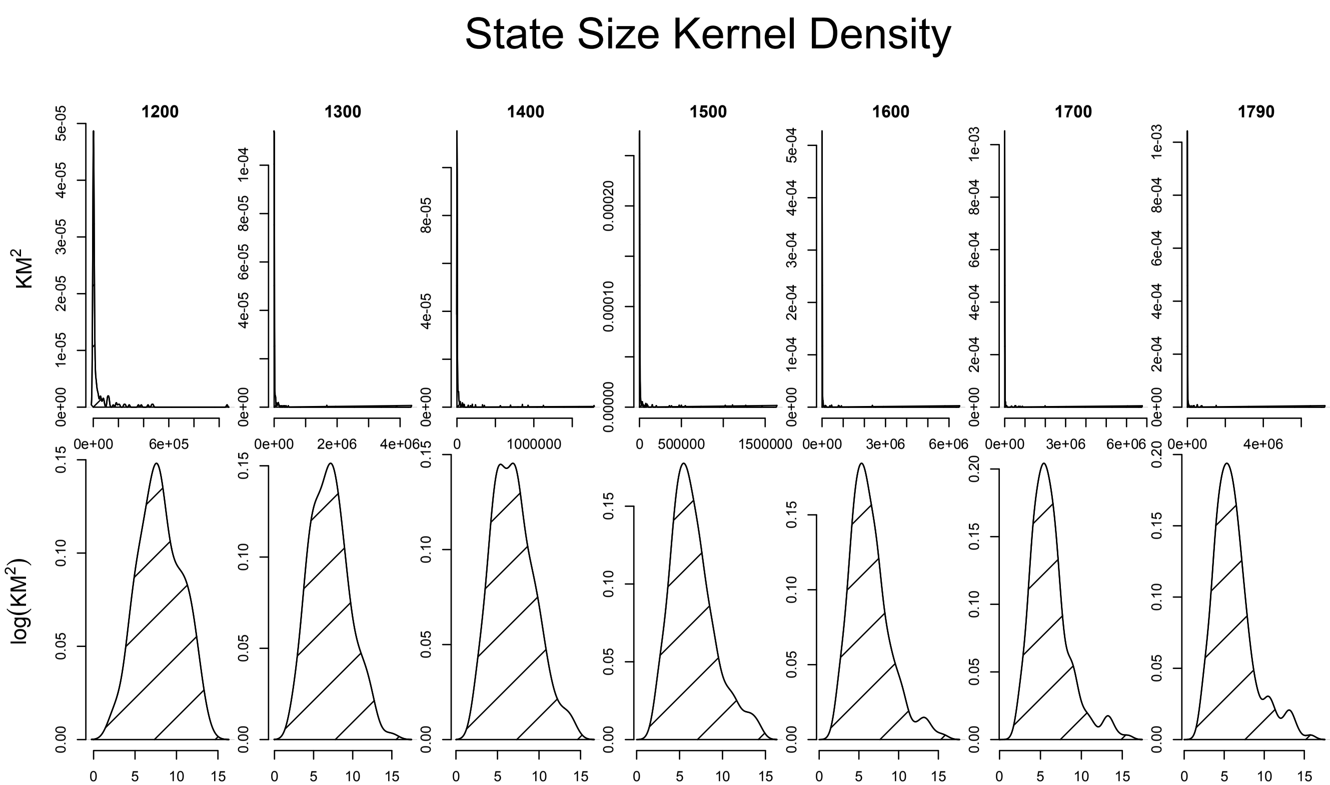


Figure A1: The top row plots kernel density estimates for the untransformed data on state size. The bottom row plots the same but for the log transformed data.

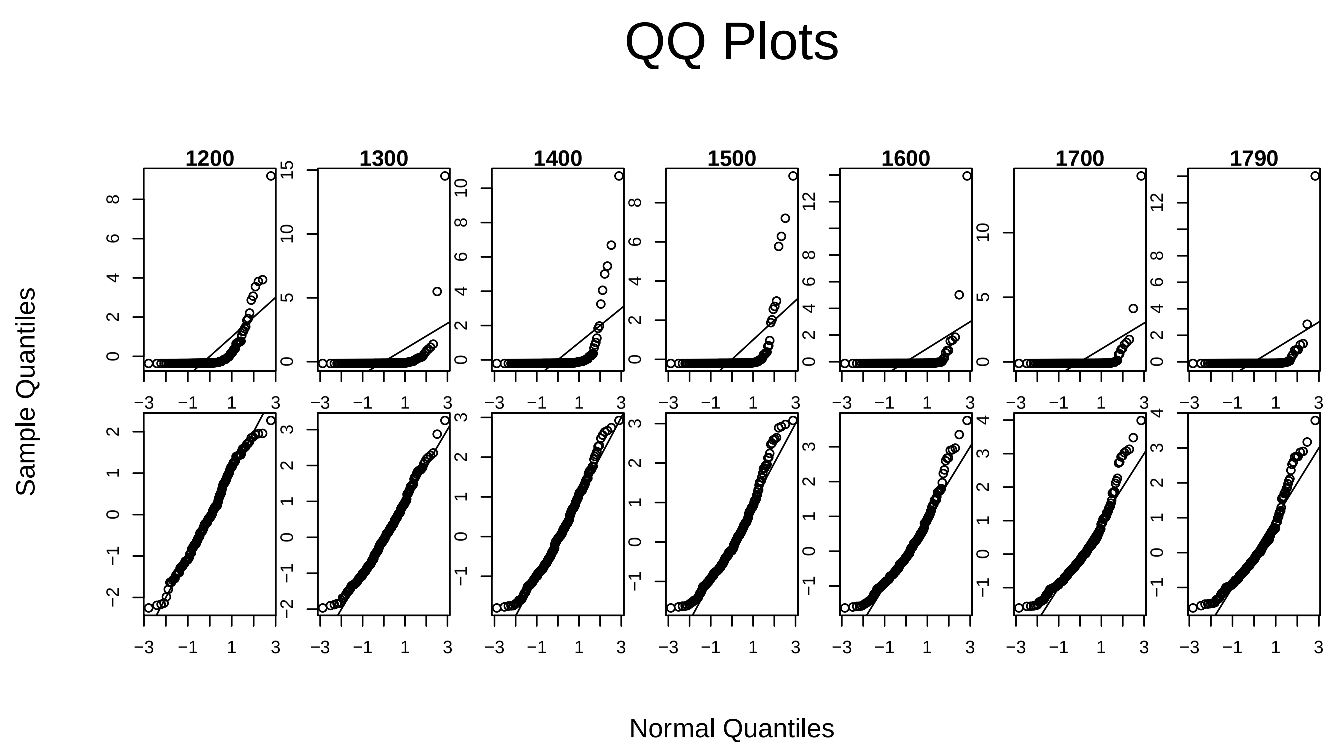


Figure A2: The top row plots for the untransformed data theoretical normal quantiles on the x axis against the sample quantiles on the y. The bottom row plots the same but for the log transformed data. Note that the transformed data sits nearly perfectly on the line $x = y$ whereas the untransformed data does not.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0. \end{cases} \tag{1}$$

Using the data pooled across all time periods I estimate λ to be -0.099, very close to a value of zero. This provides some suggestive evidence that a log normal transformation is appropriate. None of the substantive empirical conclusions drawn from this chapter are altered if I utilize a Box-Cox instead of the logarithmic transformation.

The Invariant Distribution of State Size

In this section of the appendix I provide additional evidence that the data generating process which determined the number and size of states remained constant during the period war-making theories would predict large changes. I do this in five steps.

1. I conduct three different change-point tests for a structural break in the number of states across time. I find no evidence in favor of a break during the "Age of the Territorial State."
2. I use regression based time series methods to show the process determining changes in the number of states was constant over time and best described as a series of mean zero random shocks.
3. I use panel methods to assess how the distribution of states - both the first and second moments - changed over time, again.
4. I use an EM approach to classifying each cross-section of the panel as coming from a particular latent class.
5. I show the relationship between state size and survival was constant across time periods.

Change Point Method 1

Consider a standard linear model $y_t = \alpha_t + \mu_t$ where y_t is the observation of our outcome (the number of European states) at time t, α_t some fixed mean, and μ_t is a random disturbance iid $N(0, \sigma^2)$. First, we will consider a test on structural change in this model where we are concerned with the null hypothesis of "no structural change." That is, $H_0 : \alpha_t = \alpha_0 \, \forall \, (t = 1...., n)$. We want to test this null hypothesis against the alternative that α the fixed component varies over time. This alternative can be formulated in terms of the model in the following way.

$$\alpha_t = \begin{cases} \alpha_A & \text{if } (1 < t < t_0) \\ \alpha_B & \text{if } (t_0 < t < n) \end{cases}$$

Where t_0 is a structural break in some interval (k,n-k). Chow (1960) provides a test on the structural change for the case where the potential change point t_0 is known. He suggests that we fit two separate regressions for the two subsamples defined by t_0 and to reject the null hypothesis when

$$F_{t0} = \frac{\hat{\mu}^T \hat{\mu} - \hat{\epsilon}^T \hat{\epsilon}}{\hat{\epsilon}^T \hat{\epsilon}/(n - 2k)}$$

is sufficiently large. Where $\hat{\mu}$ are the residuals fitted from the model with the entire data and no break and $\hat{\epsilon} = (\hat{\mu}_A, \hat{\mu}_B)$ are the residuals from the models where the fixed parameters $(\alpha_{A/B})$ are estimated separately. The test statistic F_{t0} is asymptotically χ^2 with k degrees of freedom and, if we make the assumption of normality, F_{t0} has an exact F distribution with k and $n - 2k$ degrees of freedom. Since the above method requires knowing where the structural break t_0 takes place, structural break tests of this sort have been extended somewhat naturally towards the estimation of F statistics for all possible change points within some interval $\{\underline{t}, \bar{t}\}$. Then, based upon these we can reject the null of no structural break at a given point in time $t \in \{\underline{t}, \bar{t}\}$ if any of these test statistics gets too large. I show in Figure A3 the estimated F-stats from the procedure outlined

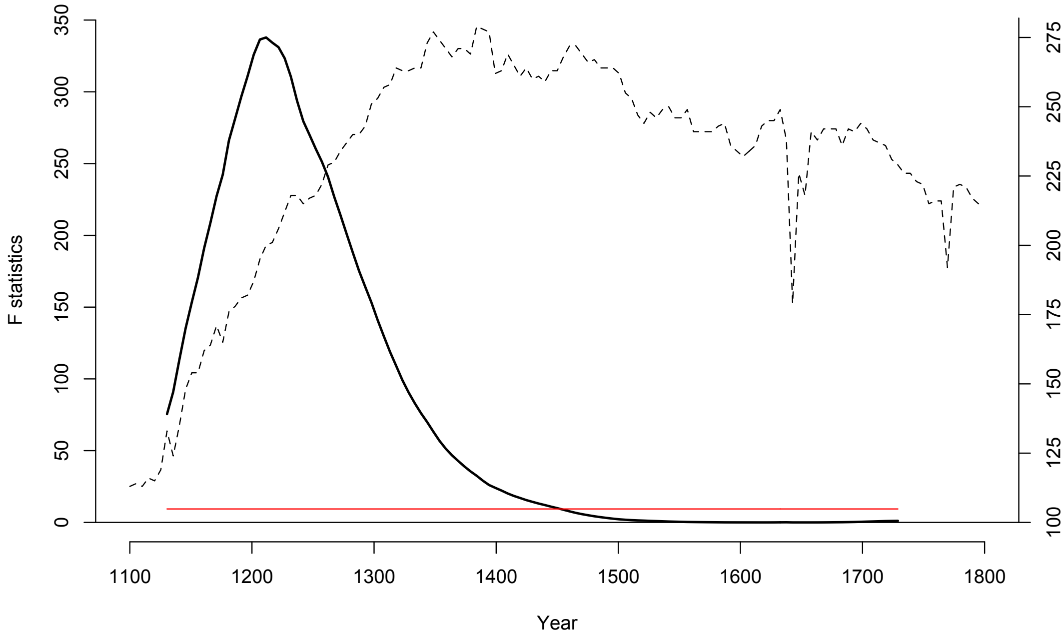


Figure A3: This figure plots on time on the x-axis against the F-statistic calculated for a break-point at that point in time. Values above the horizontal red line are those which achieve significance at the $p < .05$ level. The dashed line gives the number of European states.

above. On the x-axis is time and on the y-axis is the F-statistic associated with a structural break at that point in time. F-statistics above the horizontal red line are those that achieve significance at the $p < .05$ level. We see that in no year after 1440 is there any evidence of a structural break and that this method places the greatest likelihood (the largest F-statistic) of a structural break in the year 1210. In all, this method provides evidence that there was no break in the period associated with large scale change in the production of military violence.

Change Point Method 2

As a second way of assessing how the processes driving the number of states changed over time I adopt the method proposed by Park (2010) to identify structural breaks in count processes like the number of states, treating the number of states as coming from a number of possibly distinct Poisson processes and then identifying the point in time where the data generating process transitions from one state to another. Following Frühwirth-Schnatter and Wagner (2006) Park transforms a Poisson process into a linear regression model with a log Exponential(1) error distribution by exploiting the assumption that the time between successive events is independent and follows an exponential distribution. Taking the logarithm of interarrival times, they link the length of time between the $j-1$th event and the jth event within time interval t, τ_{tj} as follows:

$$
\begin{aligned}
& p(y_t|\lambda_t) = e^{-\lambda_t}\lambda_t^{y_t} \\
& \tau_{tj} \sim \mathcal{E}xp(\lambda_t) = \frac{\mathcal{E}xp(1)}{\lambda_t} \\
& \log \tau_{tj} = \mu_m + \varepsilon_{tj}, \varepsilon_{tj} \sim \log(\mathcal{E}xp(1))
\end{aligned}
$$

Where $\mathcal{E}xp()$ is the Exponential distribution and μ_m the regression parameter characterizing the hidden state m at time t. Using the approximation of the log Exponential(1) distribution proposed by Kim, Shephard, and Chib (1998) and Chib (1998) 's method to identify changepoints Park's procedure can recover point estimates and credibility intervals for μ_m in each state as well as the transition matrix

$$
\begin{pmatrix}
p_{11} & p_{12} & 0 & \dots & 0 \\
0 & p_{22} & p_{23} & \dots & 0 \\
 & & \multicolumn{2}{c}{\dots\dots\dots\dots} & \\
0 & 0 & 0 & p_{m-1,m-1} & p_{m-1,m} \\
0 & 0 & \dots & 0 & 1
\end{pmatrix}
$$

Where the probability of switching to regime j from state i is defined as $p_{ij} = P(m_{t+1} = j|m_t = i)$ and regime transitions are constrained to only temporally forward switches.

I begin with an arbitrary large number of possible changepoints - seven in this instance - and use the Bayes factor selection criteria outlined by Chib (1995) and Park (2010; 2012) to choose the correct number of breaks. For each of the seven models I adopt uninformative Beta priors on the location of each changepoint reflecting an equal duration of each regime given the number of states in the model. So, since we have 139 observations for the one changepoint case I adopt

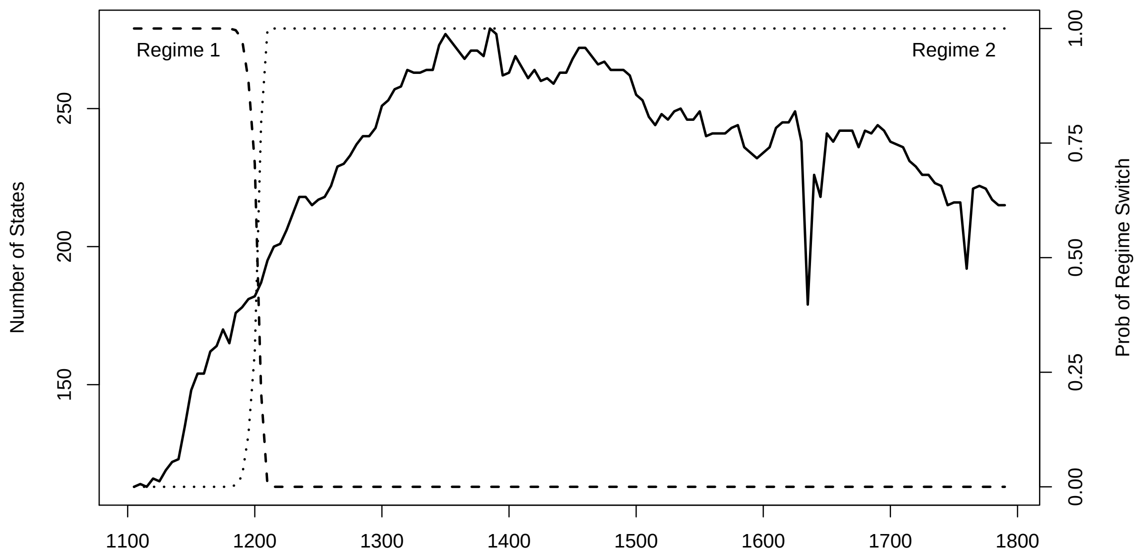


Figure A4: The Number of Independent European States

This figure plots the number of states across time. The sharp drop and recovery in this figure between 1620 and 1650 is the consequence of the Thirty-Years war. Similarly, the decline in 1760 is associated with the temporary decline in the number of units coded as independent because of the Seven Years war. The dashed lines represent the posterior probabilities of each interval falling within a given regime. Beginning with seven possible regimes, an optimal classification of two regimes, with a break at 1210 is chosen.

$\mathcal{B}eta \sim (6.95, 0.1)$, for the two changepoint case $\mathcal{B}eta \sim (4.33, 0.1)$, and so forth. For the Poisson parameter I choose gamma priors: $\lambda_t \sim \mathcal{G}amma(1, 1)$. In each model MCMC chains are run 100,000 times after discarding the first 100,000 runs.

Figure A4 plots the number of states across time as well as the estimated structural breaks derived from the procedure described above. It shows that instead of declining over time, the number of independent units *increased*, expanding rapidly between the twelfth and thirteenth centuries, peaking in the late fourteenth century, and declining slightly in the period after that.[2] Did this reduction from a late fourteenth century peak constitute a radical shift in the number of states within the European system? Again, the break-point procedure results in the choice of a single structural break dated at 1210. Figure A4 plots on the right hand axis the posterior probability of a change in regime, demonstrating the break at 1210. The mean of the first period is estimated to be 130.1 with a 95% credibility interval of $[124.3, 135.7]$ and the mean for the second period is estimated to be 227.8 with a 95% credibility interval of $[225.1, 230.7]$.

Changepoint Method 3

As a further robustness check for the presence of changes in the data generating process that determined the number of states, I exploit a third, maximum likelihood, method of searching for changepoints in time series data. Again consider some time series $y_{1:n} = (y_1, ...y_n)$. A changepoint is said to exist when there exists a time, $\tau \in \{1, ..., n-1\}$, such that the statistical properties of $(y_1, ...y_t)$ and $(y_{t+1}, ...y_n)$ are different in some way. Extending this logic to multiple changepoints we can describe a series of changepoints m and their position in time, $\tau_{1:m} = (\tau_1, ..., \tau_m)$. Where the position of each changepoint is an integer between 1 and $n-1$, such that $\tau_0 = 0$, $\tau_{m+1} = n$, and the changepoints are ordered such that $\tau_i < \tau_j$ iff $i < j$. As a consequence, the m changepoints will divide the data into $m+1$ segments where the ith segment contains the data $y_{(\tau_{i-1}+1):\tau_i}$. Each partition of the data is characterized by a parameter θ_i and an associated with a probability density function $p(y_{(\tau_{i-1}+1)|\theta_i})$. This allows us to write a log likelihood function.

[2]Stasavage (2012) looking at a subset of 168 city-states shows a similar pattern.

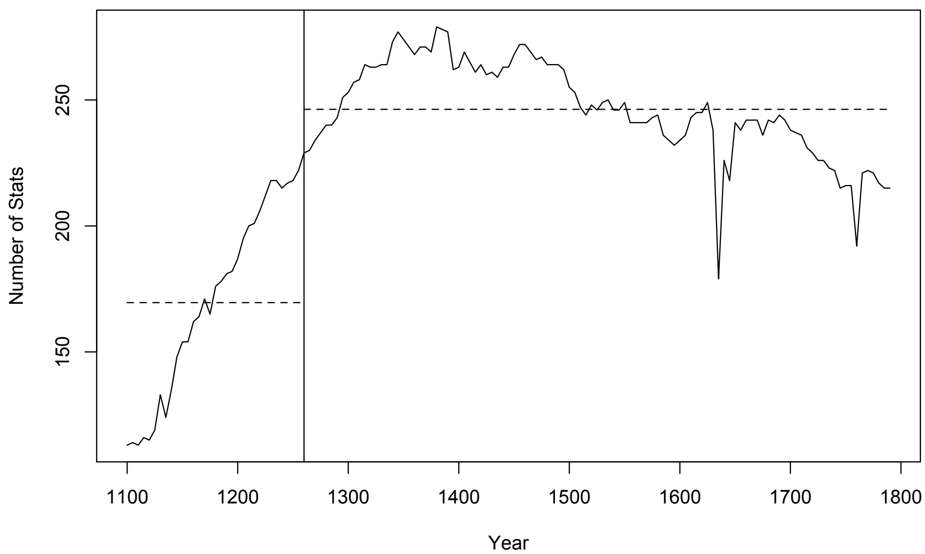


Figure A5: This figure plots the number of states with the estimated breakpoints from the maximum likelihood procedure outlined in this section. The mean for each segment is given by the dashed line.

$$ML(\tau_m) = \sum_{i=0}^{m+1} \ln p(y_{(\tau_{i-1}+1)|\theta_i})$$

The problem is then how to select over the possible combination of break points, $\tau_{1:m}$. To do this we begin by minimizing the following

$$\sum_{i=0}^{m+1} [C(y_{(\tau_{i-1}+1)|\theta_i})] + \beta f(m)$$

Where C is simply the negative of the log likelihood function defined above and $\beta f(m)$ is some cost function, which will penalize additional breakpoints. In line with the Bayesian version of this model presented in the earlier subsection, we will use the BIC penalty, although substantively there is no change if we use AIC or SIC penalties. A number of search algorithms have been developed to optimize the above over all possible combinations of possible cutpoints. As t gets large this problem becomes increasingly computationally intense and, as such, many use algorithms like that of Edwards and Cavalli-Sforza (1965) and Scott and Knott (1974) to approximately minimize the above equation. However, since my t is relatively low (140 periods) I use the segment neighborhood algorithm proposed by Auger and Lawrence (1989) which computes this minimization problem exactly. Furthermore, rather, than make distributional assumptions about $p()$ (and since the number of state sizes is unlikely to follow a normal distribution) I use non-parametric cumulative sum test statistic developed by Page (1954).

Following the procedure outlined above the results we get are roughly similar to those presented above. I plot the estimated breakpoints in Figure A5. Again, we see that there is no estimated break during the period associated with large scale changes in military technologies and tactics. However, this procedure does estimate a set of best-fit breakpoints at slightly different dates than the Bayesian method detailed in the previous section. The best fit includes a single break at 1260. Again, while these are slightly different than those identified by the previous methods, they nevertheless suggest that the data generating process determining the number of states in the European system did not change substantively during the period of the military revolution.

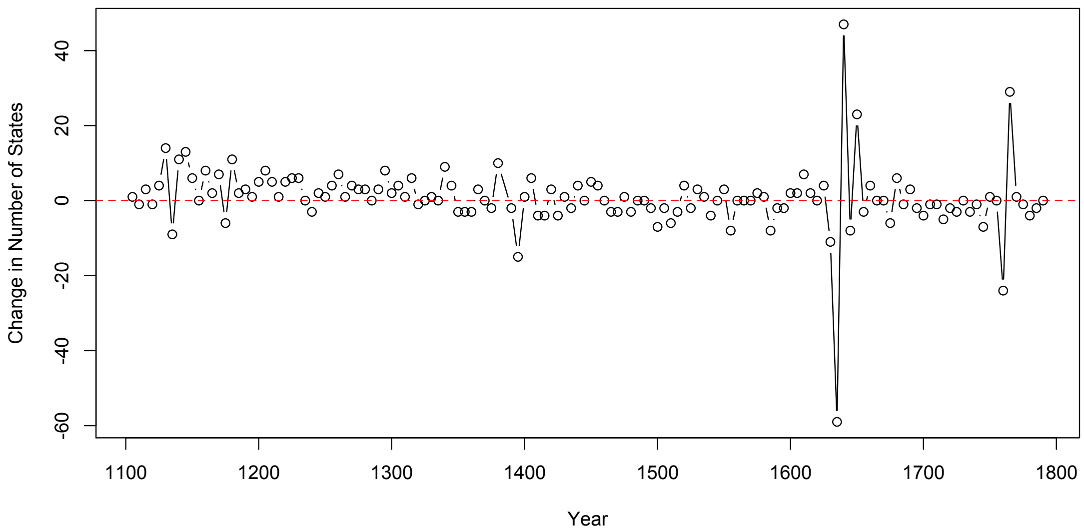


Figure A6: The change in the number of states over time.

OLS Method 1.

Are the observed changes in the number of states the consequence of deterministic trends or, rather, some random process. To assess this we can estimate the model $\Delta y_t = \lambda y_{t-1} + \epsilon$, where we are treating the change in the number of states Δy_t as a function of the last period's outcome y_{t-1} and some random noise. We are interested in two features of this model. First, if $\hat{\lambda}$ is positive it means that the series is exploding, if it is estimated to be negative it is mean-reverting, and if it estimated to be zero then it is a constant process. Second (because we are estimating this model without an intercept) if the residuals $\hat{\epsilon_t}$ are mean zero it indicates that the change in states in any given period is a random shock centered around zero. As it turns out, we estimate $\hat{\lambda} = 0.002$ with a standard error 0.003 - a fairly precisely estimate zero - and $\overline{\hat{\epsilon_t}} = 0.389$ with a standard deviation of 8.76. These results indicate that the change in the number of states is characterized by constant mean zero series of shocks.

How do these random shocks then effect the overall number of states? To see this we can simply rearrange terms to yield a model of the form $y_t = \beta y_{t-1} + \epsilon_t$, where $\beta = 1 + \lambda$. Mechanically, estimating this model yields identical estimates of $\hat{\lambda}$ and $\overline{\hat{\epsilon_t}}$. Here it is clear, however, that the total number of states is a function of the past number of states and a mean zero random shock. However, since we see that $\hat{\lambda}$ is statistically indistinguishable from zero these shocks persist across time.

The next natural question is to see if the distribution of ϵ - the random component determining state size - and the estimate of λ - the parameter describing the mean-revertive nature of our data - changed across time. Again, there are two ways of assessing this. First, I simply divide the series in two at the point we could expect a change to occur - here at 1500 - and reestimate the model for each period. The results for the two periods are substantively and statistically similar. For the period from 1100-1500 I estimate $\hat{\lambda}$ to be 0.006 with a standard error of 0.0024 and $\overline{\hat{\epsilon_t}} = 0.45$ with a standard deviation of 4.94. For the period from 1500-1790 I estimate $\hat{\lambda}$ to be -0.004 with a standard error of .007 and $\overline{\hat{\epsilon_t}} = 0.31$ with a standard deviation of 12.19. So, we see that the point estimates for $\hat{\lambda}$ and $\overline{\hat{\epsilon_t}}$ are roughly equivalent across periods, though the variance of $\overline{\hat{\epsilon_t}}$ is larger in the second period than in the first.

A more principled way of assessing whether or not there was a change in this relationship is, again, to estimate a changepoint model as before but now, instead of estimating a break in the linear model $y_t = \alpha_t + \epsilon_t$ where we are estimating a break in a fixed mean that may change over time, we estimate a break in the relationship $y_t = \beta_t y_{t-1} + \epsilon_t$. Again, we can rearrange terms so that this model equals $\Delta y_t = \lambda_t y_{t-1} + \epsilon_t$. To see if there is a break in this model, I implement the F-statistic based method outlined in the previous section. The results of this procedure, presented graphically in Figure A7, indicate that in no period is the statistical evidence in favor of a break in the series.[3]

In sum, using these regression based techniques, I have provided evidence that the statistical process by which the number of states was determined before 1790 was characterized by a strong auto-regressive component and a series of stochastic shocks centered around zero. Moreover, I have provided evidence that this process was constant across time. In other words, when proponents of

[3] I also implement the Bayesian changepoint model presented in the previous section. It produces the same result of no changepoint. The ML method does not allow for the easy incorporation of covariates and is principally designed to tests in mean shifts in a series.

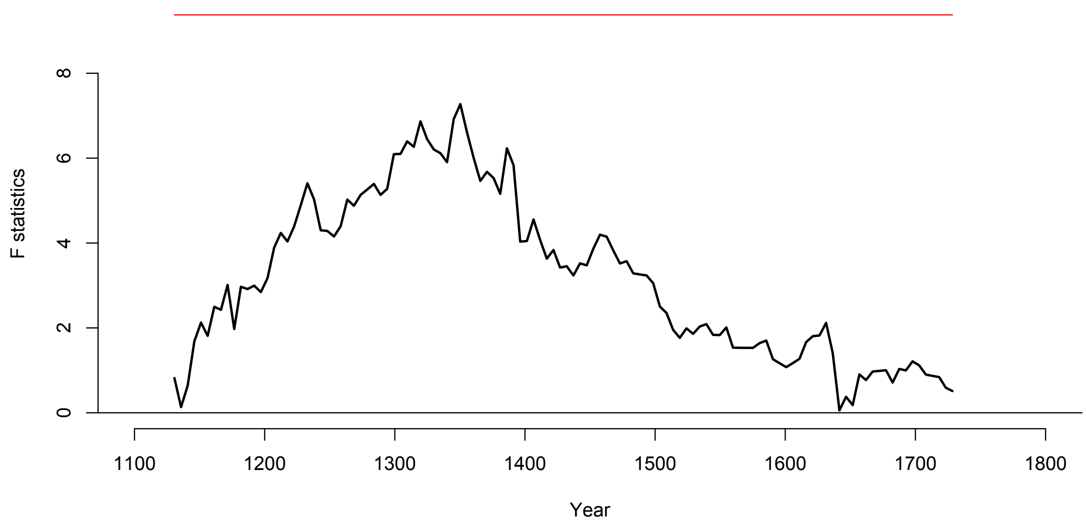


Figure A7: This figure plots the F-statsistics associated with a break-point in the auto-regressive process $\Delta y_t = \lambda y_{t-1} + \epsilon$. Values above the red horizontal line are statistically significant at $p < .05$. No value from this procedure is significant at this level.

war making theories would expect large changes in patterns of state making to emerge, we observe none.

OLS Method 2

Thus far we have only exploited information about the mean number of states to assess whether or not there was a change in the international system such that certain types of states (namely, large territorial states) could persist when other, smaller entities could not. Rather than simply exploiting the time-series component of the data to assess this, we can use the full longitudinal dataset on state size. To do this I start estimating the model $lnArea = \alpha_t + \epsilon_{it}$, where I am treating the log of state size for each country as a function of a year specific mean, α_t, and some random disturbance ϵ_{it}. Again, our estimate of the time effect, $\hat{\alpha}_t$, captures the first moment of the distribution of state size and the residuals, and $\hat{\epsilon}_{it}$ will tell us about the second moment.

We want to know, again, if the distribution of state size fundamentally changed over time. To

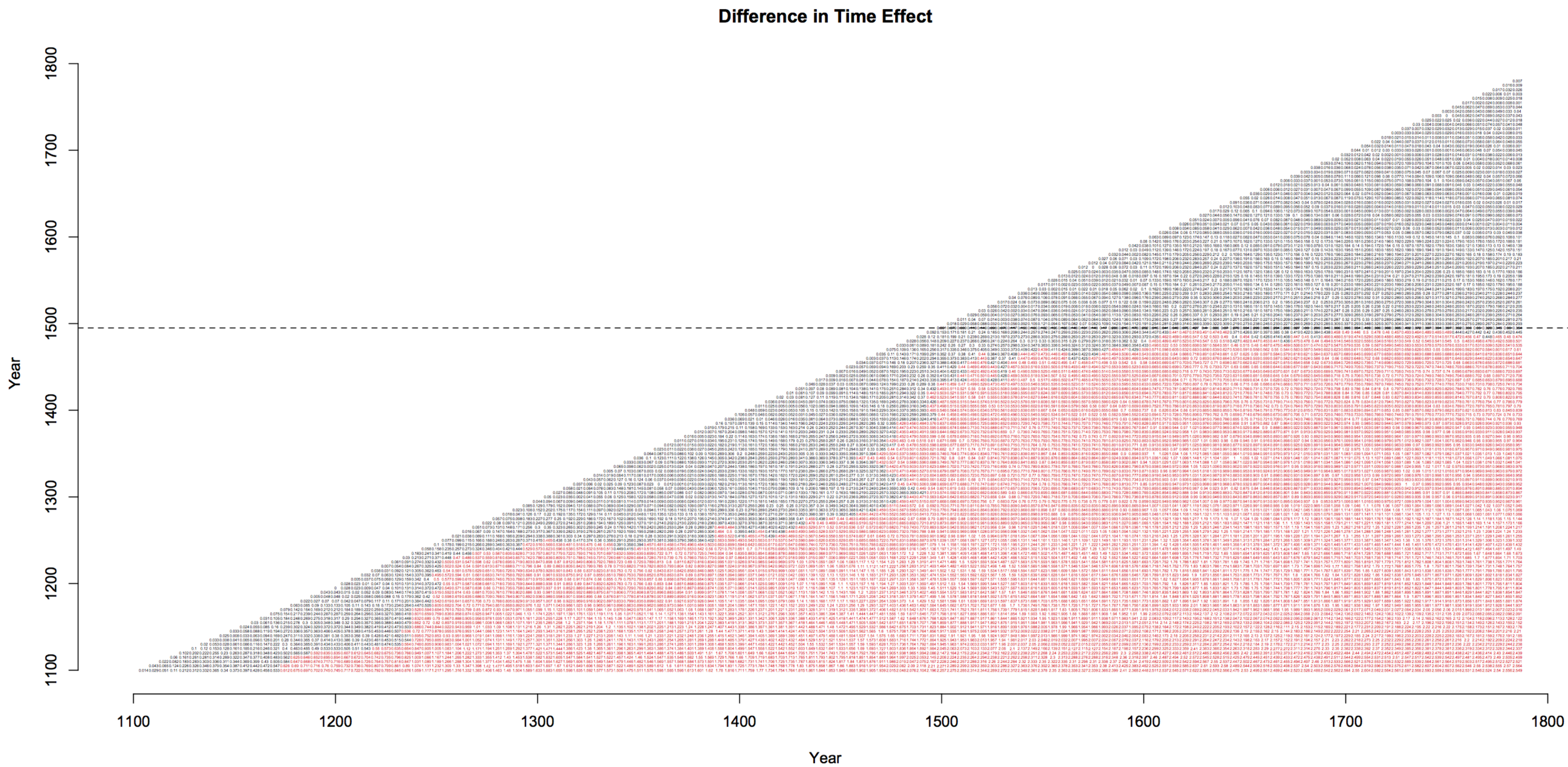


Figure A8: This figure plots $|\alpha_t - \alpha_{t'}|$ - the difference between time effects - for each pair t, t' from 1100-1790. Those differences that are red are statistically significant at the $p < .05$ level. No pair after 1500 are statistically significant.

begin, we can compare each set of time effects to see how the first moment changed across time. Figure A8 plots $|\alpha_t - \alpha_{t'}|$ - the difference between time effects - for each pair t, t' from 1100-1790. The differences colored black are those that do not achieve statistical significance at the $p < .05$ value. The differences that are colored red are those that are statistically different from zero at this level. What we see is that no pair of time effects after fifteen hundred are statistically distinguishable from each other. Indicating that, indeed, the average state size has remained roughly constant in the period described as the "age of the territorial state."

Still, we might think the second moment of the distribution changed. To asses this, in Figure A9 I plot the box-plot of residuals from the above aggression against time. Consistent with the analysis in the main text we see, that the number of outlying large states began to increase after 1455. However, for 95% of the data the distribution has remained, again, roughly constant. Still, we might be interested in the particular cases represented by these outliers. The grey points on these plots represent the following states: Russia, The Ottoman Empire, The Mongol Empire, France, England, Spain, Castile, Aragon, Denmark, Sweden, and Norway. These outlying states, in large part, explain the increase in skew in the residuals. That is, the observed skew is driven by already large peripheral states that became larger. For some, particularly Russia and the Scandinavian Countries this growth was not wholly at the expense of existing states but rather extending their control to newly populated regions.[4] Still others, like England, France, and Castile, for example, build large states by subjugating new territories to their rule. Nevertheless, these peripheral states, while historically important, were not not the norm. And, moreover, the low levels of commercialization in these places relative to the European core, as my results show, explain the consolidation into increasingly large states.

EM Method

Another way of thinking about how the distribution of states has changed over time is to model each state in a given year as being drawn from a common latent group. Treating the units within

[4] Though, the Muscovite princes did, ultimately, consolidate territory held previously by the Mongols under the loose rule of other Russian principalities, e.g. Kievan Rus, The Republic of Novgorod, Smolensk, Chernigov, amongst others.

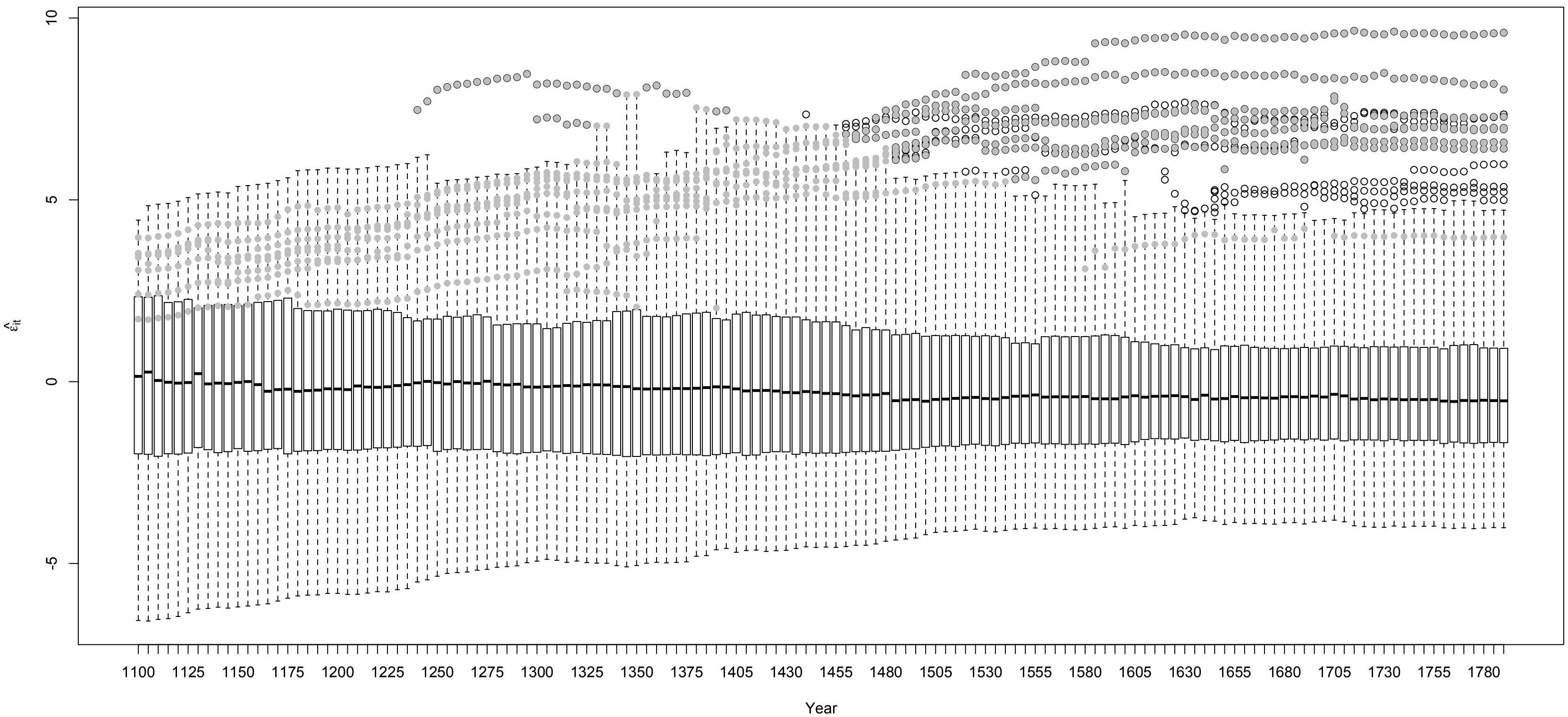


Figure A9: This figure gives the box-plot of the residuals from the regression $lnArea = \alpha_t + \epsilon_{it}$. The grey points represent the following states: Russia, The Ottoman Empire, The Mongol Empire, France, England, Spain, Castile, Aragon, Denmark, Sweden, and Norway.

every cross-section as coming from a common latent distribution, we can then use an expectation-maximization algorithm to classify each year's latent class. Formally we can write the log-likelihood as

$$\log L = \sum_{t=1}^{T}\sum_{i=1}^{N_t} \log h(y_{it}|\psi) = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K} \pi_k \log f(y_{tn}|x_{tn}, \theta_k)$$

Where π_k is the component class weight given to years 1 through T such that $\sum_k \pi_k = 1$, θ_k is the component specific parameter vector for the density function $f(\cdot)$, and $\psi = (\pi_i, ...\pi_k, \theta_1, ..., \theta_k)$. The posterior probability that cross section t belongs to class j is given by

$$P(j|t) = \frac{\pi_j \Pi_{n=1}^{N_t} f(y_{it}|\theta_j)}{\sum_k \pi_k \Pi_{n=1}^{N_t} f(y_{it}|\theta_k)}$$

A simple E-M algorithm classifies the latent distribution from which each time period is being drawn where the E- Step is defined as

$$\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{i=1}^{N_t} P(k|t) \log h(y_{it}|\psi)$$

and the M step is the maximization of this quantity with respect to ψ. This is implemented using the `FlexMix` package in `R`

Beginning with a large number of possible latent groups, again choosing seven, and using a BIC minimization criterion to select the number of latent groups, I find that there are three classes. Note that this procedure is not guaranteed to return an ordered classification of time periods. That is, the years 1500 through 1700 could be classified as coming from the same latent distribution as 1100 through 1200; there is *no* requirement that there is *any* ordering. Nevertheless, when estimated, this procedure results is an ordered set of classifications which are presented in Figure A10. Three classes emerge, the first from 1100 to 1240, a second from 1240 to 1485 , and a third from 1485 to 1790. In line with the previous analysis we again see that there is no change in the distribution of state sizes over the entire period associated with large-scale military innovations. Rather, over the entire period where small states are predicted to be driven from the international system, by this

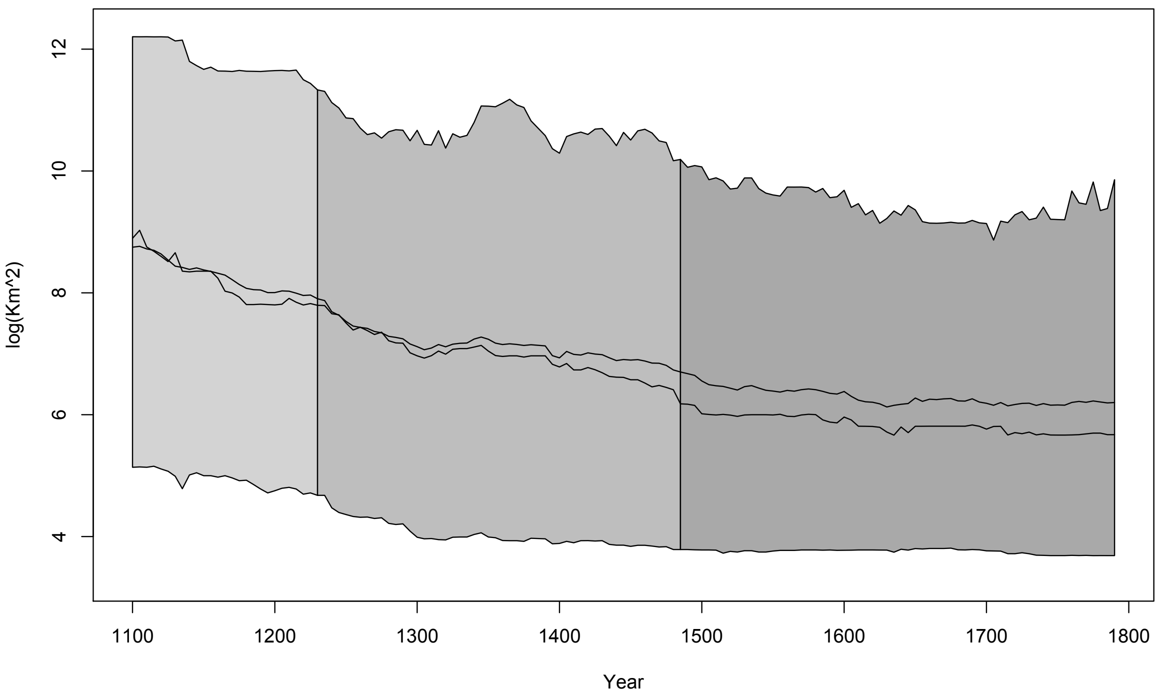


Figure A10: This figure gives the mean, median, 90th and 10th percentile state size across time. It then shades the latent classification derived from the EM approach. Three classes emerge from this procedure. The first from 1100 to 1240, a second from 1240 to 1485, and a third from 1485 to 1790.

measure, the data generating process determining the size of states remained constant.

The Time Invariant Effect of State Size on Failure

The following table gives the difference in means and Z-statistic for each of the time-varying parameters from the Cox proportional hazard model presented in the main text. No pair of parameters are statistically distinguishable form each other at conventional levels.

	γ_{1100}	γ_{1200}	γ_{1300}	γ_{1400}	γ_{1500}	γ_{1600}	γ_{1700}
γ_{1100}	0	.	.	.	.	.	.
	(0)	.	.	.	.	.	.
γ_{1200}	0.027	0	.	.	.	.	.
	(0.603)	(0)	.	.	.	.	.
γ_{1300}	0.036	0.009	0	.	.	.	.
	(0.809)	(0.206)	(0)	.	.	.	.
γ_{1400}	0.042	0.015	0.006	0	.	.	.
	(0.94)	(0.338)	(0.132)	(0)	.	.	.
γ_{1500}	0.049	0.022	0.013	0.008	0	.	.
	(1.11)	(0.508)	(0.302)	(0.17)	(0)	.	.
γ_{1600}	0.073	0.046	0.037	0.031	0.024	0	.
	(1.65)	(1.048)	(0.842)	(0.71)	(0.54)	(0)	.
γ_{1700}	0.07	0.043	0.034	0.028	0.02	-0.003	0
	(1.572)	(0.97)	(0.764)	(0.632)	(0.462)	(-0.078)	(0)

Table A1: This table presents the difference between each of the time varying parameters with Z-statistic in parentheses.

Instrumental Variables Construction and Robustness Checks

The Construction and Logic of the Instrument

The instrument is constructed in two steps:

1. I take spatially referenced temperature data from two paleo-climatological sources, measured at half-degree by half-degree latitude/longitude intervals. The first measure from Mann, Zhang, Rutherford et al. (2009) records temperature anomalies for the past 1500 years. A temperature anomoly captures the deviation at each point from the 1961-2000 mean temperature. I then construct a measure of absolute temperature by adding back the 1961-2000 baseline mean temperature as calculated from Jones, New, Parker et al. (1999)'s twentieth century data. This yields a half degree by half degree grid of temperatures for every year over the past 1500 years. Hundred year averages of these yearly measures are then taken.

2. Next, I take these estimates, measured at fixed intervals, and construct a smoothed measure of temperature for the entire continent. From this continuous measure the average for each grid-square is taken yielding an estimate of temperature across our fixed but arbitrary pieces of geography. All of these operations are taken using the interpolation and zonal averaging tools found in ArcGIS 10.

I employ two data sets from the FAO. The first, the "Agro-climatically attainable yield for rain fed wheat," is from the Global Agro-ecological Assessment for Agriculture in the 21st century. It captures the ability of land to produce wheat absent of modern irrigation techniques. I estimate the optimal climate to grow wheat (at around 10.5 °C). A clear a parabolic relationship between temperature and this FAO measure is observed simply by plotting it against average annual temperature between 1960 and 2000.

Regressing the FAO measure of wheat suitability on the absolute deviation from 10.5 degrees we see that, indeed there is a negative relationship between the two. The results from this regression are summarized in first column of Table A2. The effect of a one degree deviation from the optimal temperature is substantial, decreasing the FAO measure by .61 units. This is a particularly large

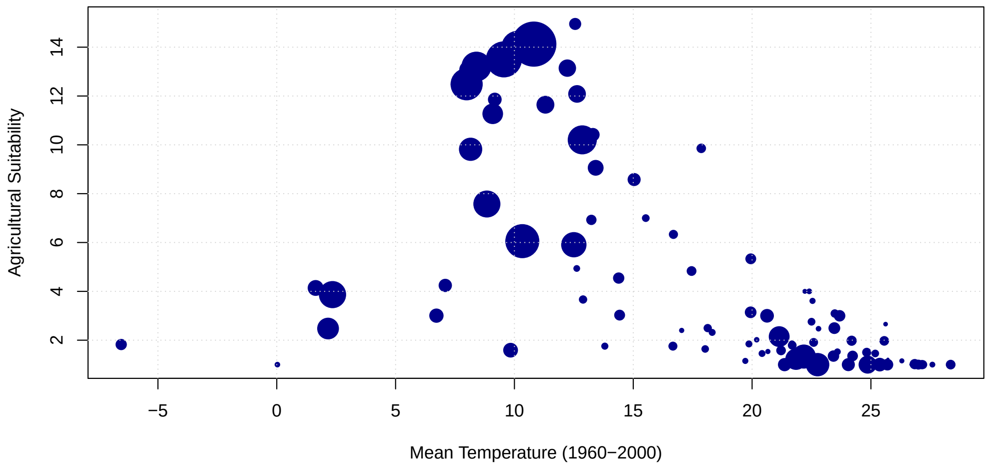


Figure A11: The FAO wheat suitability index is plotted on the y-axis against average annual temperature on the x-axis. The FAO measure is the "Agro-climatically attainable yield for rain fed wheat" is from the Global Agro-ecological Assessment for Agriculture in the 21st century. It captures the ability of land to produce wheat absent of modern irrigation techniques. A clear parabolic relationship with a peak at approximately 10.5 °C is observed. The radius of each circle is proportional to the average wheat yield between 1960 and 2000

effect since the FAO measure is on a fourteen point scale. Moreover, an large amount of the variation in the FAO wheat suitability measure is explained by deviation from this optimal growing temperature, the R^2 statistic is calculated to be .55. In addition, regressing average annual wheat yields between 1960 and 1990 on deviation from the optimal growing temperature again shows a similarly robust relationship. A one degree deviation from the optimal temperature has a large effect on average annual wheat yields – approximately 1600 hectograms per hectare.

	FAO Wheat Suitability	Avg. Wheat Yield
Intercept	9.89***	36901***
	(.49)	(2659)
Growth Temperature	-.61***	-1612.7***
	(.05)	(254)
R Squared	.55	.30
N	119	94

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table A2: The relationship between temperature and the suitability to produce wheat. The first column regresses the FAO wheat suitability index against the absolute average distance from 10.5 °C between 1960 and 2000. The second column regresses the average wheat yield on this measure.

Regional Effects

Table A3 gives empirical results including regional dummies for the following regions: The British Isles and France, Iberia, Italy, Scandinavia and the Baltics, Eastern Europe including Russia, and the Ottoman Territories which includes the Balkans. I then divide out Central Europe, comprising contemporary Germany and Austria, from the rest of peripheral Europe. Across specification, the main results hold and are of roughly the same magnitude.

Controlling for Income

To construct a grid-square level estimate of per capita income, I take the data defining historical incomes for contemporary states from Maddison (2006) and assign these values to the centroid of each contemporary state. Then, using these points I estimate smooth-spline estimate of income and measure the average value to each grid-square. All operations are done in ArcMap10. Since Maddison (2006) only provides estimates for the years 1000, 1500, 1600, 1700, and 1800. I linearly interpolate income for the years 1200-1400 based on their 1000 and 1500 income levels.

If the instrument affects state formation processes through other economic channels, inclusion of this measure of income should control for these channels. These results, given in Table A4, show that the effect urban growth remains qualitatively unchanged when I include this measure of per capita income. The results hold, but are slightly smaller, when I exclude the linearly interpolated years 1200-1400 from the analysis.

Controlling for The Number of Cities

To assess whether political fragmentation is being driven by the number of distinct cities on a given unit, thus increasing the potential for city-states, or by changes in the economic potential of territory, I re-run the main set of statistical models but now controlling for the number of cities. That is, by controlling for the number of cities on a given territory we are making comparisons between changes in urban population across units that have the same number of cities within them. These results are presented in Table A5. Across specification the results are positive and statistically significant, indicating that the effect of changes in urban population are not operating through the creation of more city-states. In fact, when we control for the number of cities on a given unit the effect of changes in urban population on the number of states within a given grid-square increases in comparison to the base-line models.

Sensitivity Analysis

Following Conley, Hansen, and Rossi (2012) I conduct a sensitivity analysis using their union of confidense intervals method and allow for a direct effect of the instrument on the outcome over a range of symmetric bounds. Formally, we estimate via a 2SLS-like procedure $Y = X\beta + Z\gamma + \epsilon$, where Z is my instrument and X is my variable of interest and allow the parameter γ - capturing the direct effect of the instrument on the outcome - to take on a range of values $\gamma_0 \in [-\gamma, \gamma]$. The following graphs give the resulting confidence rustling from this procedure and treating γ as a function of my initial 2SLS estimate of β. I conduct this exercise for both the within estimator and the pooled estimator. For the within estimates we see that our initial estimate would become statistically indistinguishable from zero when $\gamma = .36\beta$. For the pooled estimates this is true when $\gamma = .26\beta$.

Mechanisms

This section provides some suggestive empirical evidence that the process of urbanization caused the fragmentation of political rule by creating new social groups associated with the revival of commerce and urbanization. Again, I use the data on proto-industrial activity presented in the first section as a proxy for these actors' presence. First, I simply describe the statistical relationship between proto-industry and political fragmentation in the 1500. Next, I treat the presence of the actors proxied by the existence of proto-industry as a mediating variable affecting the number of states forming on a given piece of geography. That is, I estimate the relationship between urbanization in 1200 and political fragmentation as it operates through the existence of proto-industry.

To begin, I estimate the relationship between the presence of proto-industry before 1500 and the number of states on the geographic units defined above. The results, estimated via negative binomial regression, are presented in Table A6 and indicate a strong, positive, and statistically significant relationship between the existence of proto-industry and the number of states forming on a given piece of territory. To gauge the magnitude of this relationship, in Figure A13 I plot the predicted number of states on each unit as derived from the odd columns of Table A6, for the observed values of iron and textile centers and then their sum. The predicted increase in the number of states is stark, with one state expected on a unit with zero proto-industrial centers and twenty at eight, the maximum observed total proto-industrial centers. The magnitude of this relationship is substantial, indicating that in those places where the presence of groups involved in craft manufacturing were also those places where by 1500 the most new states had formed.

As a next step, I estimate proportion of the effect of urbanization on political fragmentation that operates through the creation of social groups proxied by the existence of early manufacturing. To accomplish this, I adopt the methodology proposed by Imai, Keele, and Tingley (2010), Imai,

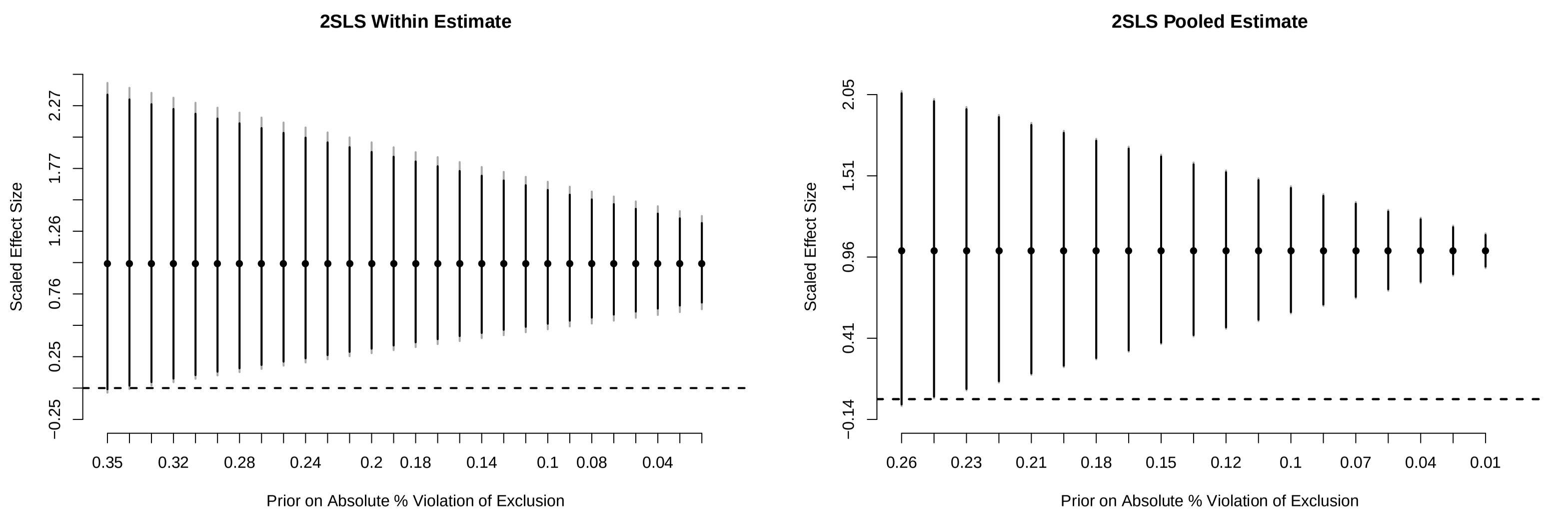
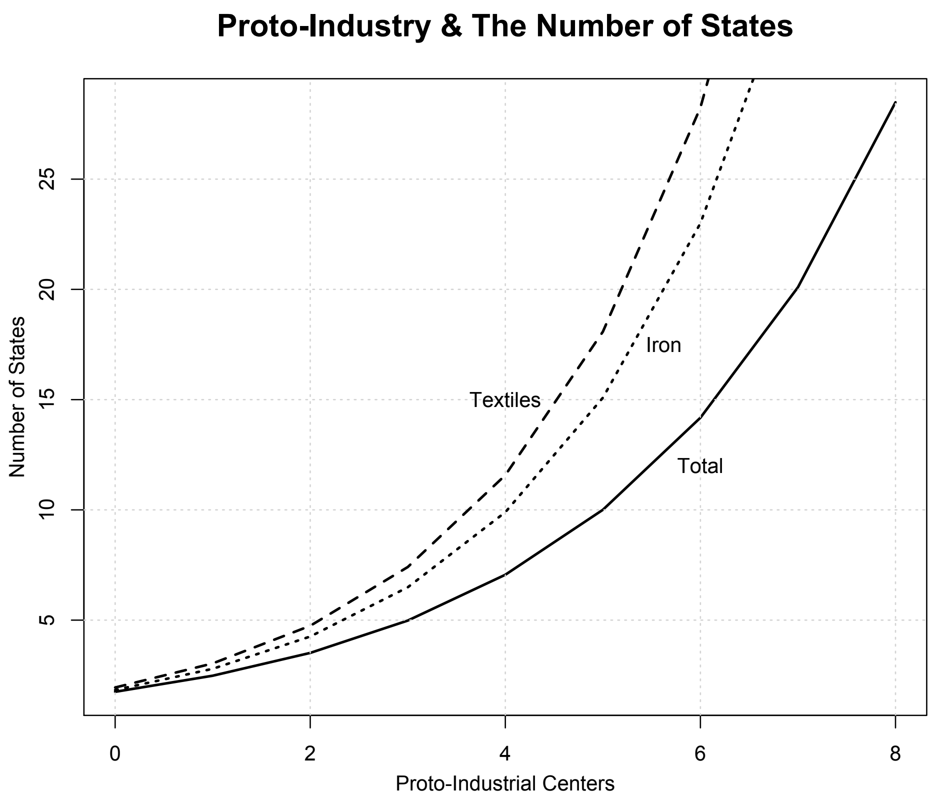


Figure A12: These plots provide results from the union of confidence interval method for conducting sensitivity analysis of the exclusion restriction. The x-axis gives the symmetric interval of γ as a fraction of the estimated effect size β. The y-axis gives the resulting confidence band.

Figure A13: The Existence of Proto-industry and Political Fragementation



This figure gives the predicted number of states forming on a given geographic unit as a function of the number of proto-industrial centers existing before the year 1500.

Tingley, and Yamamoto (2013) to conduct causal mediation analysis.[5] Because the "treatment" - urban population in 1200 - is not randomly assigned the mediating effects I estimate here should not be interpreted causally. Still, conducting this exercise and estimating this mediating relationship gives us some picture of the proportion of the effect of urbanization as it operates through the existence of groups involved in manufacturing.

Again, to avoid post-treatment bias I am forced to examine the relationship between the "treatment," urban population in 1200, the mediator, proto-industry before 1500, and the number of

[5] We implement the procedure as detailed in Imai, Keele, and Tingley (2010). Formally, the mediation effect is defined as $\zeta = \mathbb{E}\left[Y(t, M(m+1)) - Y(t, M(m))\right]$, the direct effect, $\delta = \mathbb{E}\left[Y(t+1, M(m)) - Y(t, M(m))\right]$ and the total effect $\tau = \mathbb{E}\left[Y(t+1, M(m+1)) - Y(t, M(m))\right] = \zeta + \delta$.

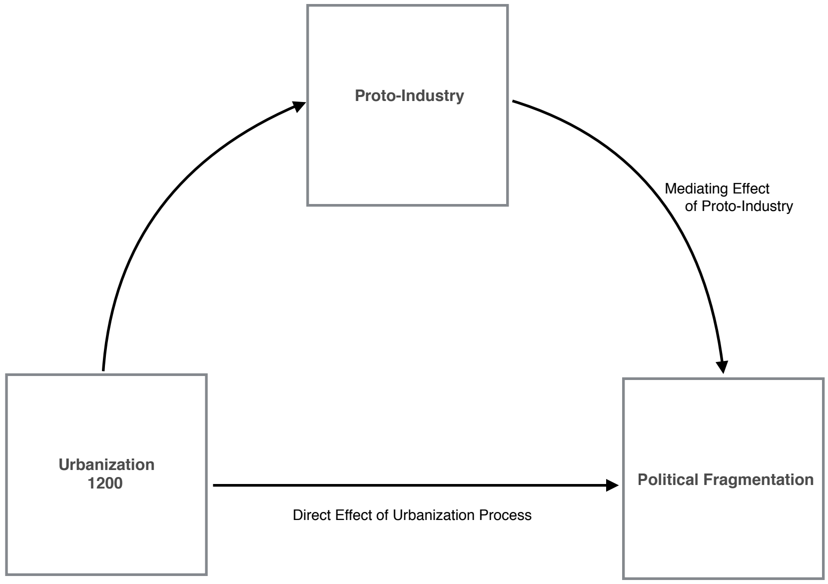


Figure A14: The Mediating Effect of Proto-Industrial Development on Political Fragmentation
This graph describes the mediating effect of proto-industrial development on the number of states forming in the year 1500. The treatment is taken to be early urbanization, measured in the year 1200, which has some direct effect on the process of state formation (in 1500) as well as some indirect, mediating, effect as it operates through the creation of groups involved in craft manufacturing.

states on a given unit at 1500. Figure A14 describes these relationship graphically. Implementing the procedure detailed in Imai, Keele, and Tingley (2010), Table A7 gives two types of estimates of this mediating relationship. First, I estimate both the mediating and direct effects using ordinary least squares. Second, because the outcome is a count measure I use a Poisson model.[6]

There exists a consistently positive and statistically significant relationship between urbanization and political fragmentation as it operates through each measure of proto-industrial activity. Moreover, these results indicate that about forty percent of the total relationship between urban population in 1200 and the number of states forming on a given geographic unit is estimated to operate through the presence of proto-industry. To the degree that the existence of iron metallurgy

[6]Count-models that account for over-dispersion are not currently available in the R package `mediate`.

and textile production proxy for the presence of artisan and other commercial groups, this result demonstrates that the process of uneven urbanization caused, in part at least, political fragmentation by creating new social groups with the material resources capable of asserting independence.

References

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 211–252.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 1313–1321.

———. 1998. Estimation and comparison of multiple change-point models. *Journal of econometrics* 86 (2):221–241.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi. 2012. Plausibly exogenous. *Review of Economics and Statistics* 94 (1):260–272.

Frühwirth-Schnatter, S., and H. Wagner. 2006. Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika* 93 (4):827–841.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychological methods* 15 (4):309.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105 (4):765–789.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (1):5–51.

Jones, PD, M. New, DE Parker, S. Martin, and IG Rigor. 1999. Surface Air Temperature and Its Changes Over the Past 150 Years. *Reviews of Geophysics* 37 (2):173–199.

Kim, S., N. Shephard, and S. Chib. 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies* 65 (3):361–393.

Maddison, A. 2006. *The world economy*, vol. 1. OECD.

Mann, M.E., Z. Zhang, S. Rutherford, R.S. Bradley, M.K. Hughes, D. Shindell, C. Ammann, G. Faluvegi, and F. Ni. 2009. Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science* 326 (5957):1256.

Park, Jong Hee. 2010. Structural change in US presidents' use of force. *American Journal of Political Science* 54 (3):766–782.

———. 2012. A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models. *American Journal of Political Science* .

Stasavage, David. 2012. Was Weber Right? City Autonomy, Political Oligarchy, and the Rise of Europe. *Tech. rep.*, Working Paper.

Instrumental Variables Estimates By Region

log(Urban Population)	0.43	0.42	0.46	-0.81	0.47	4.41	1.17	0.28	0.28	0.58	0.68
	[0.32,0.54]	[0.32,0.53]	[0.33,0.60]	[-11.56,9.95]	[-2.48,3.43]	[-2.41,11.24]	[0.52,1.82]	[0.25,0.32]	[0.25,0.32]	[0.31,0.84]	[0.25,1.11]
Just Core				Yes	Yes	Yes	Yes				
Just Periphery								Yes	Yes	Yes	Yes
Region Effects	Yes	Yes	Yes								
Time Effects		Yes	Yes		Yes	Yes			Yes	Yes	
Fixed Effects							Yes				Yes
Lat/Long						Yes				Yes	
River Density			Yes								
F Stat on Excluded Instrument	71.32	74.38	51.85	0.31	2.69	2.56	37.55	318.44	319.81	18.53	14.09

Table A3: This table presents additional specifications of the baseline empirical model. The first three columns present results controlling for region effects. Columns 4-7 estimate the effect of urban population on the number of states forming on a given unit, excluding territory outside of the core. Columns 8-11 do the same, excluding the core. Ninety-five percent confidence intervals are in brackets. Standard errors are clustered by unit.

Instrumental Variables Estimates Controlling For Income

	All Years			*> 1400*		
log(Urban Population)	0.42	0.39	0.78	0.20	0.17	0.53
	[0.35,0.48]	[0.32,0.45]	[0.35,1.21]	[0.16,0.24]	[0.13,0.21]	[0.14,0.92]
log(Income)	-1.02	-0.41	-1.94	0.63	1.35	0.04
	[-1.46,-0.58]	[-1.13,0.31]	[-3.97,0.10]	[0.18,1.08]	[0.64,2.05]	[-1.66,1.74]
Lat/Long			Yes			Yes
Year Effects		Yes	Yes		Yes	Yes
F State on Instrument	214.05	159.66	13.511	224.07	186.68	9.194
N:	1552	1552	1552	1552	1552	1552
T:	7	7	7	4	4	4

Table A4: Instrumental variables results controlling for estimated income per capita at the grid-square level. The first six columns use all data including those linearly interpolated, covering 1200-1400. The last six only use data as estimated from Maddison (2006). Ninety-five percent confidence intervals derived from standard errors clustered by unit are in brackets.

Table A5: The Relationship Between Urban Population and Political Fragmentation Controlling for the Number of Cities

Instrumental Variables Estimates Controlling For The Number of Cities

log(Urban Population)	0.44	0.71	0.71	1.17	0.86	0.72
	[0.37, 0.50]	[0.46, 0.97]	[0.46 , 0.96]	[0.44, 1.90]	[0.09, 1.63]	[0.13 1.31]
# of Cities	-0.24	-0.53	-0.51	-0.94	-0.68	-0.68
	[-0.36, -0.11]	[-0.85, -0.21]	[-0.83, -0.205]	[-1.65, -0.24]	[-1.30, -0.06]	[-1.29,-0.074]
Lat/Long		Y	Y			
Year Effects		Y				
Grid Effects				Y	Y	
Entrants_{t-1}					Y	
First Differencing						Y
N	1684	1684	1684	1684	1684	1684
T	7	7	7	7	6	6
F-Stat on Excluded Instrument	112.20	30.08	32.26	12.37	5.66	6.78

This table gives 2SLS estimates of the relationship between urban population on a given grid-square and the number of states that exist on it after controlling for the number of cities within the unit.

The Relationship Between Proto-Industry and The Number of States

Iron	0.42***	0.22***				
	(0.03)	(0.02)				
Textiles			0.45***	0.12**		
			(0.05)	(0.05)		
Iron+Textiles					0.35***	0.17***
					(0.02)	(0.02)
Geographic Controls	·	Y	·	Y	·	Y
AIC	4247.54	3902.88	4330.72	3964.48	4217.64	3910.69
θ	1.56	2.29	1.46	2.15	1.61	2.27

Table A6: The Relationship Between Proto-Industry and Political Fragmentation

Negative Binomial estimates of the relationship between the existence of proto-industrial centers before 1500 and the number of states existing on a given. Heteroskedasticity-robust standard errors are in parentheses. Geographic controls are latitude/longitude and the density of rivers on a given unit.

The Mediating Effect of Proto-Undustry

	OLS			**Poisson**		
	Iron	*Textiles*	*Total*	*Iron*	*Textiles*	*Total*
Mediation Effect (Proto-Industry)	0.09	0.05	0.14	0.06	0.04	0.14
	[0.05,0.12]	[-0.01,0.11]	[0.09,0.2]	[0.04,0.07]	[0.02,0.06]	[0.08,0.19]
Direct Effect (Urban Population)	0.27	0.3	0.21	0.33	0.38	0.21
	[0.16,0.38]	[0.17,0.42]	[0.09,0.34]	[0.27,0.39,]	[0.32,0.45]	[0.09,0.34]
Proportion Mediated	0.24	0.15	0.4	0.14	0.09	0.4
	[0.15,0.4]	[-0.02,0.36]	[0.23,0.66]	[0.11,0.18]	[0.05,0.14]	[0.23,0.65]

Table A7: The Mediating Effect of Proto-Industry on Political Fragmentation

This table gives estimates of the effect of urban population in the year 1200 on the number of states forming on a given 10,000 km^2 geographic unit in the year 1500 as mediated by the presence of proto-industry before 1500. Ninety-five percent confidence intervals derived from quasi-bayesian simulation are in brackets.