Surnames: a New Source for the History of Social Mobility

Gregory Clark, Neil Cummins, Yu Hao, Dan Diaz Vidal[1]

This paper explains how surname distributions can be used as a way to measure rates of social mobility in contemporary and historical societies. This allows for estimates of social mobility rates for any population for which the distribution of surnames overall is known as well as the distribution of surnames among some elite or underclass. Such information exists, for example, for England back to 1300, and for Sweden back to 1700. However surname distributions reveal a different, more fundamental type of mobility than that conventionally estimated. Thus surname estimates also allow for measuring a different aspect of social mobility, but the aspect that matters for mobility of social groups, and for families in the long run.

KEYWORDS: Social Mobility, intergenerational correlation, status inheritance

Introduction: Social Mobility Concepts

We assume social status can be measured by a cardinal number y which measures some aspect of social status such as income, wealth, occupational status, longevity or height. Conventionally social mobility rates have been estimated by economists from the estimated value of β in the equation

$$y_t = \alpha + \beta y_{t-1} + u_t \qquad (1)$$

where y is the measure of social status, t indexes the generation, and u_t is a random shock. β will typically lie between 0 and 1, with lower values of β implying more social mobility. β is thus the persistence rate for status, and 1- β the social mobility rate. Also if the variance of status on this measure is constant across generations then β is also the intergenerational correlation of status. And in this case β also

[1]Clark, Department of Economics, University of California, Davis, CA 95616. gclark@ucdavis.edu. Cummins, Department of Economic History, LSE. Hao, ----. Diaz Vidal, Department of Economics, University of California, Davis, CA 95616,

estimates the share of the variance of status in each generation that is explicable from inheritance. This share then will be β^2. The reason for this is that if σ^2 measures the variance of the status measure *y,* and σ_u^2 measures the variance of the random component in status, then, from equation (1)

$$var(y_t) = \beta^2 var(y_{t-1}) + var(u_t)$$

$$\sigma^2 = \beta^2\sigma^2 + \sigma_u^2$$

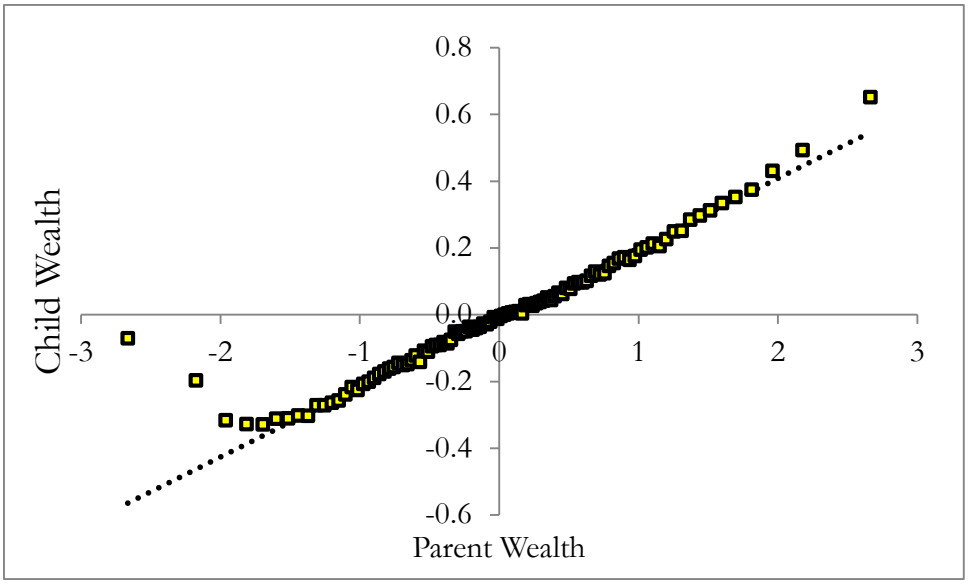
If equation (1) is the correct description of the inheritance of social status in any society, then in steady state any measure of status such as the logarithm of income or wealth will show a normal distribution.

Equation (1) involves a number of strong simplifying assumptions. It assumes, for example, that social mobility rates are the same across the whole of the status distribution, from top to bottom. But we shall see that the empirical evidence is that this assumption is not too far from reality.

For example, a recent study of intergenerational wealth mobility in Sweden assembled data from Danish tax records that allows a comparison of the wealth of 1.2 million children with that of their parents.[2] The huge size of the Danish wealth data set means that the authors can divide the parents into percentiles and look at the average wealth of children for each parent percentile, measured again as a percentile of the child wealth distribution. Other than the top and bottom 3 or 4 percent of parental wealth, the picture has the linear character equation (1) assumes. One persistence rate, 0.20, describes inheritance across the middle 92 percent of the distribution (figure 1). The greatest deviation appears in the bottom 4 percent of parental wealth, where the children are much richer than we would expect. But the parents at the bottom of the distribution have negative wealth. This suggests not chronic, grinding poverty (the truly poor do not get to borrow much), but more likely indebtedness to finance a business venture or training. The fact that this is not

[2] Boserup, Kopczuk, and Kreiner 2013.

Figure 1: Inheritance of Wealth in Denmark, 1997–2012.



truly the bottom of the wealth distribution explains the breakdown of the stable relationship. Children in the top 3 percent of the parental-wealth distribution also show slightly greater wealth inheritance. But though this effect is statistically significant, it represents only modest deviations from the single persistence rate in real terms: the persistence rates implied by the top three percentiles are 0.24, 0.23, and 0.22 respectively.

Estimating Mobility Rates from Surnames

Conventional estimates of mobility rates require knowledge of the social status of parents and their children. Such data is publicly available on a systematic basis only in a few societies. In the contemporary world this requires long duration survey panels such as the US NLSY, or population registry data that assigns unique family identifiers, as in the modern Nordic countries. In the nineteenth and early twentieth century it is possible to link families using successive censuses, as for England 1841-1911, and the USA 1850-1940. But the linkage of individual parents and children through censuses, where spelling of surnames and first names is highly idiosyncratic, is a difficult and time consuming process. And as we shall see below there are

reasons to question if the conventional estimates of social mobility reveal its true rate for more generalized measures of status.

For the reason above we have until recently had no idea of what social mobility rates were in pre-industrial societies. We have had no idea whether, for example, the Industrial Revolution in England was associated with a period of enhanced social mobility compared to what came before and what came after.[3]

However, in many societies people have surnames, and these surnames are inherited unchanged through the patriline. Men bearing the surname *Boscawen* born in England 1900-1930, for example, are descended from someone in the group of men bearing the surname *Boscawen* in 1870-1900. Thus using surnames to group people we can identify groups of sons who collectively descended from a group of fathers, without knowing the exact descent relationships. The fact that surnames can proxy for the transmission of the y chromosome between generations has long been of interest to geneticists.[4] However only recently have there been attempts to utilize surnames to estimate social mobility rates.[5] Here we describe two methods of estimating intergenerational social mobility from surnames, but have been other techniques recently developed, not all however suitable for historical data.[6]

Instead of estimating β from

$$y_t = \alpha + \beta y_{t-1} + u_t \quad (1)$$

we can use

$$\bar{y}_{kt} = \alpha + \beta \bar{y}_{kt-1} + \bar{u}_{kt} \quad (2)$$

[3] See Clark and Cummins, 2014, for a review of the evidence on this.
[4] See Lasker, 1985, King and Jobling, 2009, Garza-Chapa, Rojas-Alvarado, and Cerda-Flores, 2000.
[5] Weyl (1989) used surnames to identify social groups, and to measure their relative status in the modern US, but did not attempt to measure rates of regression to the mean.
[6]Güell, Mora, and Telmer (2007) use cross-section information content of surnames in current census records to estimate intergenerational mobility. But their approach involves assumptions about such things as surname mutation rates that are hard to verify. Collado, Ortuño-Ortín, and Romeu (2013) develop a technique more suitable for historical data, but primarily focus on estimating transition probabilities between status categories, probabilities that do not translate easily into intergenerational correlations of status.

where k indexes surname groups and $\bar{}$ indicates averages. We can, for example, compare the average status of everyone born with the surname *Boscawen* in 1800-1829 to those born with this surname 1830-1859, the 30 year interval between the time periods here representing the assumed average length of a generation.

This averaging across surnames would be expected to produce an attenuated estimate of the β lining fathers and sons for several reasons. First we have to take all those born with a class of surnames in a time interval (t, t+n) and compare them to those born in the time interval (t+30, t+n+30), the 30 years representing the average interval between generations. This introduces error in that some children of the generation born in the interval (t, t+n) will not be born in the interval (t+30, t+30+n). And some of those born in the interval (t+30, t+30+n) will have fathers not born in (t, t+n). Second the surname method counts those in (t, t+n) who have no children equally with those who have large numbers of children. Third the surname method includes wives of men bearing the surnames who adopted those surnames on marriage. Fourth there will potentially be some adopted children among the younger generation, as well as those who changed surnames from their birth surname. For all these reasons the surnames can only provide an imperfect estimate of the average of the actual parent child status linkages. This imperfection should bias the surname estimates towards zero.[7]

However, these surname estimates of β are always much greater than the β estimated from individual family linkages. Thus one surname study (Clark and Cummins, 2013) looked at the inheritance of wealth at death in England for those dying 1858-2012. For each person dying we can estimate their normalized wealth at death, which is their log wealth minus average log wealth. Table 1, for example, shows the estimated $\hat{\beta}$s for children born in each of four periods in England, 1888-1917, 1918-59, 1960-87, and 1999-2012 estimated from equation (1) for individual family linkages.[8] Also shown are the estimated $\hat{\beta}$s from equation (2) based on the rare surnames of those who on average died wealthy in 1858-1887. These estimates are, surprisingly, consistently higher than those from the direct family linkages. This

[7] The bias caused by adoption in Japan will be towards greater status persistence, because in this case there is mainly adult adoption of sons-in-law, and the sons so adopted are typically selected based on their ability to carry on family businesses. See Clark et al., 2014, chapter 10.

[8] For the years 1988-1998 there are no usable estimates of wealth at death, since in these years wealth for the probated is reported within just a few broad bands.

Table 1: Individual versus Surname estimates of Wealth Inheritance

Period of child death	$\hat{\beta}$ Individual Links	$\hat{\beta}$ Rich Rare Surnames
1888-1917	0.48	0.71
	(0.03)	(0.03)
1918-59	0.41	0.69
	(0.02)	(0.03)
1960-87	0.41	0.73
	(0.02)	(0.03)
1999-2012	0.46	0.83
	(0.06)	(0.07)
All	0.43	0.74
	(0.01)	(0.02)

Note: Robust standard errors in parentheses.
Source: Clark and Cummins, 2013.

turns out to be a general feature of social mobility rates using surnames. The intergenerational persistence of status estimated in this way is typically in the order of 0.7-0.8.

Clark et al. 2014 explains this seeming discrepancy in the following way. The proposal is that we must distinguish between measures of a family's surface or apparent social status and their deeper social competence, which is never observed directly.[9] What is observed for families is their attainment on various partial indicators of social status: earnings, wealth, occupation, education, residence, health, longevity. Each of these derives from underlying status, but with a random component. Thus the proposed model of status transmission is

[9] In psychometric terms, underlying status is a latent variable.

$$y_t = x_t + u_t \quad (3)$$

$$x_t = bx_{t-1} + e_t \quad (4)$$

where x_t is the family's underlying social competence, and u_t is the random component, and b is the persistence rate of underlying status.

The random component of aspects of social status exists for two reasons. First, there is an element of luck in the status attained by individuals. Second, people sometimes sacrifice aspects of status such as income and wealth for other aspects such as education or occupational prestige. University professors are a classic example of this tradeoff.

The above implies that the conventional studies of social mobility, based on estimating the intergenerational correlation β in the relationship

$$y_{t+1} = \beta y_t + v_t \quad (5)$$

for various partial measures of status—earnings, wealth, education, occupation and so on —underestimates the true intergenerational correlation *b* that links underlying social status across generations. In particular, the expected value of conventional estimates β is not the underlying *b* but instead θ*b,* where $\theta = \frac{\sigma_u^2}{\sigma_x^2 + \sigma_u^2}$ is less than one. Further, the greater the random components of any measured aspect of status, the smaller will θ be.

Since we have these two measures - *b* for underlying social mobility, and β for partial measures of status - why is it that the underlying *b* is the true rate of social mobility? The reason is that if we were to measure families' status by an average of the various observed aspects of status, $\bar{y}_t$, then

$$\bar{y}_t = x_t + \bar{u}_t \quad (6)$$

where $\bar{}$ indicates an average of the various random components. But as we average status across many aspects—earnings, wealth, residence, education, occupation, health, longevity—the average error component shrinks toward zero. Thus the

intergenerational persistence of average measured social status lies somewhere between b and β. The underlying b gives us potentially a better measure of the persistence of status on average for families, as opposed to the persistence of any particular aspects of status. Also if we want to predict the correlation of any measure of status over n generations it will be θb^n.

But when we consider the social mobility of large groups of people identified by race, religion, national origin, or even surnames (where whether a surname belongs to a high or low status catergory has been identified in some earlier generation), the measure b will unambiguously be the one that reveals their rate of social mobility. For now at the group level

$$\bar{y} = \bar{x}.$$

Now the $\bar{y}$ accurately tracks $\bar{x}$ without the intrusion of the errors, and we can correctly estimate underlying social mobility. When we look at such groups of individuals, the underlying, slow rate of social mobility becomes apparent even when we can observe only the usual partial indicators of underlying social competence. This is why the surname groups provide a measure of underlying rates of social mobility. But any grouping that is independent of the current random elements determining a partial measure of status will do the same. That is why it will always seem that racial, ethnic and other minorities within societies experience slower than expected social mobility.

Implementing Surname Measures of b – Direct Estimates

Where we have direct measures of social status by surname, implementing the estimation of b is straightforward. We need only identify groups of surnames that are preselected as having high or low status, and then examine what happens to the average status of these surnames over time. We need to make an assumption about what generation lengths are to get the intergenerational b. But with surname averages it is possible to also estimate bs using periods shorter than a generation length such as a decade.

The surname estimates of wealth persistence in England reported above come from such a procedure. To identify surnames of high or low status, rare surnames

were used, since common surnames in England, having been established by 1300, tend to differ little in social status. The average wealth of the surname was established for those dying 1858-1887. Then surnames were grouped as rich, prosperous or poor, and the average wealth of each of these groups computed for the subsequent generations. Clark and Cummins 2013 describes these results in detail.

Another example of such an estimation is that done by Daniel Diaz Vidal for occupational status in Chile. The underlying source is the Electoral Register of 2004, which records for six million voters their name, age, location, and occupation. This allows people to be assigned a measured status in two ways.[10] The first is based on the average earnings of their occupation. The second on the average earnings of people living in their municipality. Since people only have an occupation on completing schooling, we look only at people born before 1980, who will be aged 25 by the time of the register. If we assume an average generation interval of 30 years we can then compare average occupational or locational status for those born 1920-49 compared to those born 1950-79.

To get elite and underclass groups of surnames Diaz Vidal can use two procedures. First surnames in an immigrant society like Chile can be classified by ethnic and national origin. Thus there is a class of surnames associated with the Mapuche, the main surviving indigenous population of Chile. There are also surnames associated with immigrant groups of Basque, German, French and Italian origin. Basque settlers, for example, were an early elite in colonial Chile. But further Diaz Vidal can identify, as in the case of England, rare surnames associated with earlier wealth in Chile in the nineteenth and early twentieth century. An annual agricultural yield report was compiled, for example, in 1853 to determine agricultural taxes.[11] From this list, Diaz Vidal selected those last names that appeared between 3 and 30 times a contemporary Chilean population census. The average yield value of a parcel of land in the 1853 report was 379 pesos. Diaz Vidal takes large holders as holding parcels of yield greater than 1,500 pesos. There is a second list of large landholders in 1920, from which again Diaz Vidal selects those with rare surnames.

[10] For details on the sources for Chile see Clark et al., 2014, chapter 11, 199-211.
[11] Estado que manifiesta la renta agrícola de los fundos rústicos que comprende el impuesto anual establecido en la sustitución del diezmo por la ley de 25 de Octubre de 1853, Imprenta del Diario, Calle de la aduana #40, Valparaiso, in Octuber of 1855

Table 2: Estimated Chilean Social Mobility Rates, births 1920-1979

Surname Group	N 1920-49	N 1950-79	Ratio N	Ave Occupational Earnings, 1920-49	Ave Occupational Earnings, 1950-79	Implied b
Mapuche	7,036	17,389	2.47	-0.304	-0.239	0.79
Basque	8,755	17,841	2.04	0.225	0.169	0.75
Large Landowners, 1853	2,731	5,201	1.90	0.396	0.371	0.94
Large Landowners, 1920	1,680	3,069	1.83	0.450	0.415	0.92
All	895,145	2,059,057	2.30	0.000	0.000	-

Note: The numbers reported in each period are those who the electoral register lists with an occupation.

Table 2 shows the numbers of people from each of four such surname groups found with an occupation in the 2004 electoral register born 1920-49 and 1950-79. For the country as a whole there are 2.3 times as many people recorded with an occupation in 1950-79 as earlier. But interestingly for the low status group, the Mapuche the ratio later is greater at 2.47, while for the high status groups it is lower.

The table also shows the average log occupational earnings of each group, relative to the average for all electors. Thus columns five and six show for birth cohort

$$\frac{1}{N_{ik}}\sum_i \ln w_{ik} \; - \; \frac{1}{N}\sum_i \ln w_i \qquad (7)$$

Where $\ln w_i$ is the log occupational earnings for each elector, N is the total number of electors with occupations, N_{ik} is the number of electors with occupations in

surname group k, and $\ln w_{ik}$ is the log occupational earnings of each member of group k. For those with Mapuche surnames born 1920-49 the value of -0.304 implies that their average occupational earnings are only 74 percent of the overall average for this birth cohort. For those with the rare surnames of large landowners in 1920 the value of 0.450 for the 1920-49 birth cohort implies that their average occupational earnings are 57 percent higher than the overall average for this birth cohort.

The b estimate in the final column comes just from the equation

$$\overline{\ln w_{k1}} = b\,\overline{\ln w_{k0}} \qquad (8)$$

where the subscript 1 indicates the generation born 1950-79, and the subscript 0 the generation born 1920-49. As can be seen, these estimates suggest strong persistence of occupational status for both the high status and low status groups.

Using historical census records from other countries, such as England 1851-1911, it will be possible to construct similar measures of social mobility by surname groupings using the occupations listed in the census, and some translation of these into earnings equivalents.

Implementing Surname Measures of b – Indirect Estimates

The data from historical sources can often take another form, which is that where the indicator of the social status of surnames is their frequency among elites and underclasses compared to their frequency in the general population. For England, for example, there is information on the population shares of surnames from 1538 and the beginning of parish registers of baptisms, marriages and burials to the present. In this same interval there is also information on the shares of various surnames at Oxford and Cambridge, the only universities in England before 1836, and even after this the most prestigious. Thus for each period after 1500 we can estimate for each surname its relative status, the measure being

$$\frac{Share\ of\ surname\ z\ at\ Oxbridge}{Share\ of\ surbame\ z\ in\ Oxbridge\ age\ cohort} = RR_z$$

By definition for the average surname in England in any period this number will be 1. But for high status surnames the number will exceed 1, and for low status surnames it will fall below 1.

To find elite surnames associated with Oxford and Cambridge (Oxbridge) 1800-29 we do the following. We use the 1881 census, the English census that was most carefully digitized, to identify surnames held by less than 500 people in England in that year. We classify as the rare surnames of Oxbridge 1800-29 any surnames that are not held by 500 or more people in 1881. This generates 3,312 individual surnames held by Oxbridge students in these years. These surnames were held by 421,024 people in 1881, and by 972,314 in 2002. To estimate the population share with these rare surnames in each student cohort we use records of marriages in England 1837-1915, and records of births 1916-1995. The share of the population with this sample of rare surnames in each generation of students, taking here a generation as 30 years, is shown in the second column of table 3. This share varies from 1.5 percent to 1.7 percent, increasing over time.

Table 3 shows the numbers of students with these surnames in each thirty year period starting in 1800 at Oxbridge, as well as the total numbers of students in each period. Column five shows the share of these surnames as a share of all Oxbridge students. As can be seen, in the earlier period these surnames represent more than 30 percent of students despite being held by an estimated 1.5 percent of the population. The last column shows the relative representation of these surnames at Oxbridge 1800-2013 by period. As can be seen there is a steady decline in that relative representation across generations, though it is still nearly 1.5 in 2010-13.[12] Figure 2 shows the relative representation by generation in logarithms.

What is the persistence rate of educational status implied by the last column of table 3. To measure this we make the following three assumptions.

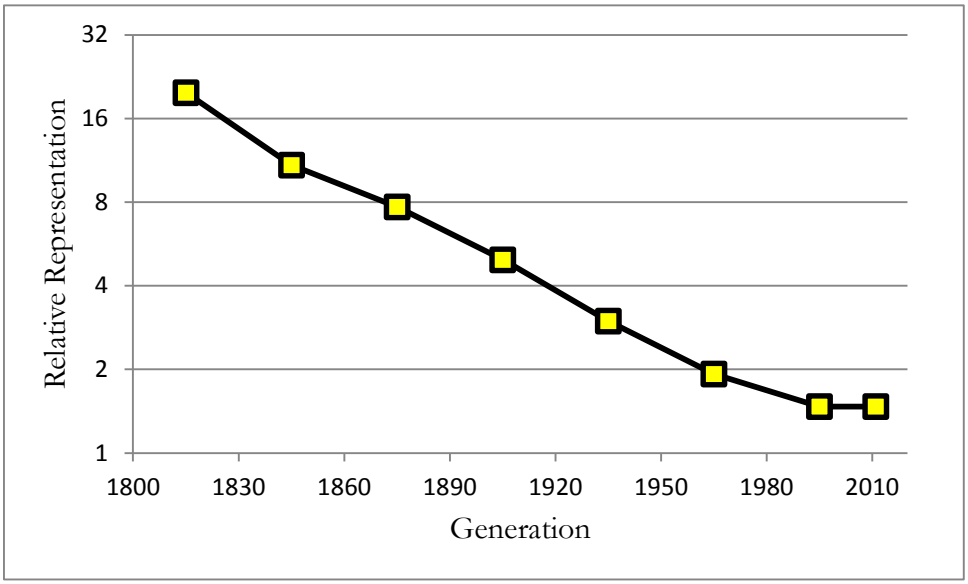
(a) Oxford and Cambridge represent the top x% of the educational status distribution, where x is measured in each cohort by the share of males in England and Wales who attend these universities.

[12] To calculate the relative representation in the periods after 1829 an allowance has to be made for the increasing share of foreign students at Oxbridge. The England and Wales share of students is calculated from 1830 on by period as 0.99, 0.97, 0.95, 0.92, 0.90, 0.82, and 0.62.

Table 3: Rare Surnames at Oxbridge, 1800-29

Generation	Share Population Rare Oxbridge Surnames %	Rare Surnames 1800-29 at Oxbridge	All Oxbridge attendees	Share Rare 1800-29 Surnames Oxbridge %	Relative Representation
1800-29	1.61	5,675	18,650	30.4	19.06
1830-59	1.59	4,063	24,415	16.6	10.57
1860-89	1.57	4,477	38,678	11.6	7.59
1890-1919	1.61	2,239	30,961	7.23	4.73
1920-49	1.65	2,974	67,927	4.38	2.89
1950-79	1.70	4,545	156,645	2.90	1.90
1980-2009	1.74	4,633	222,063	2.09	1.47
2010-3	1.74	872	49,243	1.77	1.47

Figure 2: Relative Representation, Rare Elite Surnames, Oxbridge, 1800-2013



(a) Educational status is normally distributed with constant variance.
(b) The elite surname group from 1800-29 has the same variance of educational status as the population as a whole among its members.

We consider below the plausibility of each of these assumptions. But a consequence of these three assumptions is that with them we can fix for each generation what the mean status of the rare elite surnames of 1800-29 is, measured as standard deviation units above the mean educational status. Table 4 shows the relative representation of our 1800-29 elite surnames by generation, as well as the estimated Oxbridge population share among males. Employing assumptions (2) and (3) we can then fix the implied mean educational status of this surname group by generation, as is shown in column 4 of the table. Here the mean status of the rare surnames is shown in terms of standard deviations above the social mean.

Figure 3 shows diagrammatically how this happens. The relative representation of the rare surnames in each period, combined with the cutoff level for the elite population, determines what share of the elite surname population lies in the elite group, and hence where the mean of the elite surnames lies relative to the social mean. Thus in 1800-29 the rare elite surnames had a relative representation of 19.84. The Oxbridge elite represented 0.64 percent of males. Thus the implied share of males with the elite surnames attending Oxbridge was 12.2 percent (19.06×0.64). This in turn implies that the mean value for educational status for the elite surnames was 1.57 standard deviations above the social mean in 1800-29.

Once we know the implied mean of status for the 1800-29 elite rare surname group, we can then calculate for each period the implied correlation of status b with the previous generation. From equations (4) and (5), and assuming with averaging that $\bar{y}_t = \bar{x}_t$, that is that the average measured educational status of the surnames is the average actual status

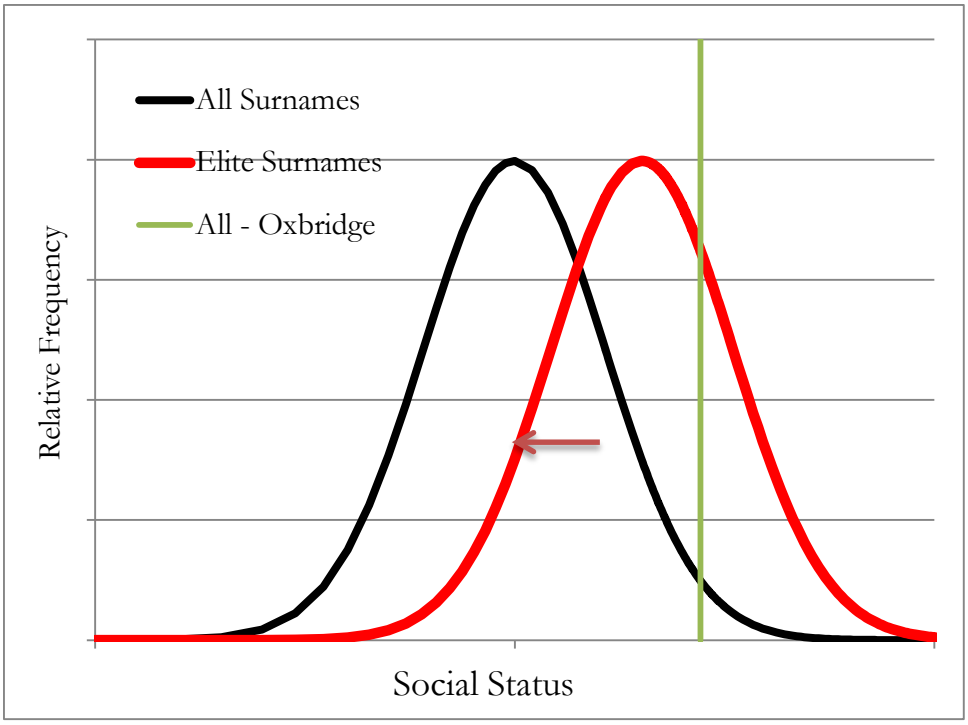
$$\bar{y}_{t+1} = b\bar{y}_t + \epsilon_{t+1} \qquad (9)$$

where ϵ_{t+1} is an error term corresponding to various mis-measurements. These are errors in measuring of the share of the surname population in each cohort, the share of these names at Oxbridge (in some periods we have just a sample of Oxbridge students, not the population), the share of the domestic population among Oxbridge

Table 4: Implied persistence rates for 1800-29 elite rare surnames

Generation	Relative Representation	Oxbridge elite share %	Implied Mean status (standard deviation units)	Implied b
1800-29	19.06	0.64	1.32	-
1830-59	10.57	0.62	0.99	0.75
1860-89	7.59	0.53	0.81	0.81
1890-1919	4.73	0.48	0.59	0.73
1920-49	2.89	0.70	0.41	0.70
1950-79	1.90	1.16	0.26	0.63
1980-2009	1.47	1.27	0.15	0.60
2010-3	1.47	1.24	0.15	0.99

Figure 3: Regression to the Mean of Elite Surnames



students, and the degree of eliteness that Oxbridge attendance implies. The unbiased estimated value of b for each period is then just

$$\frac{\bar{y}_{t+1}}{\bar{y}_t} \tag{10}$$

These estimates are shown in the final column of table 4. The average is 0.74, though the individual b estimates range from 0.60 to 0.99.

Suppose we assume, however, that this variation is just the product of the aforementioned measurement errors, and fit one b value to the whole of the data. To do this note that equation (9) implies

$$\bar{y}_{t+n} = b^n \bar{y}_t + \epsilon^*_{t+n} \tag{11}$$

or
$$\ln \bar{y}_{t+n} = \ln \bar{y}_t + \ln(b) \, . \, n + \ln \epsilon^*_{t+n} \tag{12}$$

So just by estimating the coefficient h in the OLS best fitting relationship

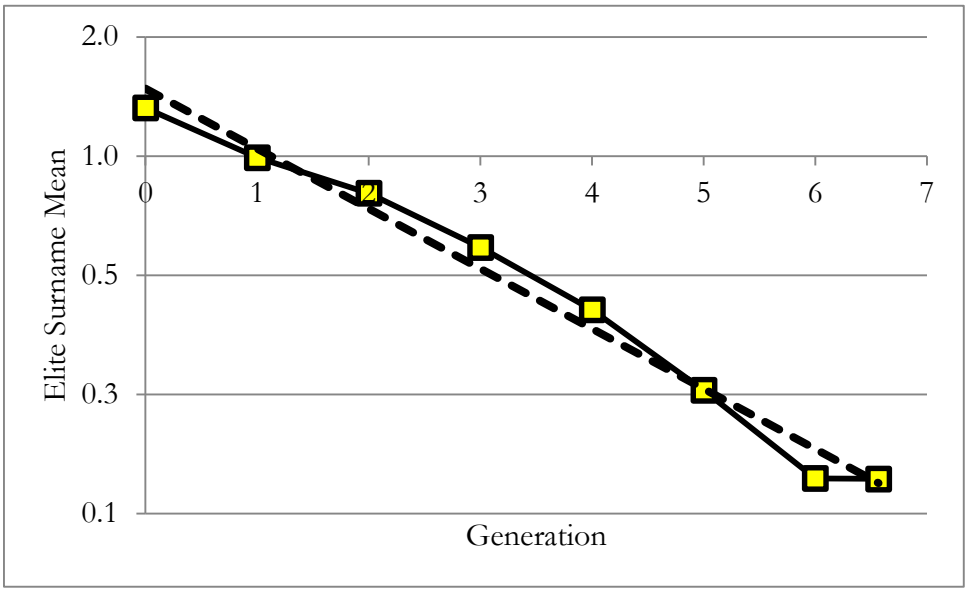
$$\ln \bar{y}_{t+n} = g + h.n$$

we can estimate the best fitting b for the whole set of observations, assuming this has a constant value. The b estimated in this way is 0.650, with 5 percent confidence bounds of (0.614, 0.687). As figure 4 shows the R^2 of this fit is good, being 0.983.

This raw sample of rare surnames appearing at Oxford and Cambridge 1800-29 has deficiencies, however. Included in the sample are names that were rare in England, because they were of Scottish, Irish, or foreign origin. Some of these surnames became much more frequent in England by 2002 because of migration of Scots and Irish to England. One example is the name Adair, held by 397 people in England in 1881, but by 2,043 in 2002, because of immigration from Ireland.

Thus the 1800-29 rare Oxbridge surname sample saw a ten percent increase in its population share between 1881 and 2002, while the average English surname saw a 15 percent decline in surname share. That means that many of the people bearing these surnames in 2002 are not descended from those holding the surnames in England in 1800-29. This should bias the estimated intergenerational correlation

Figure 4: Best Fitting constant b estimate



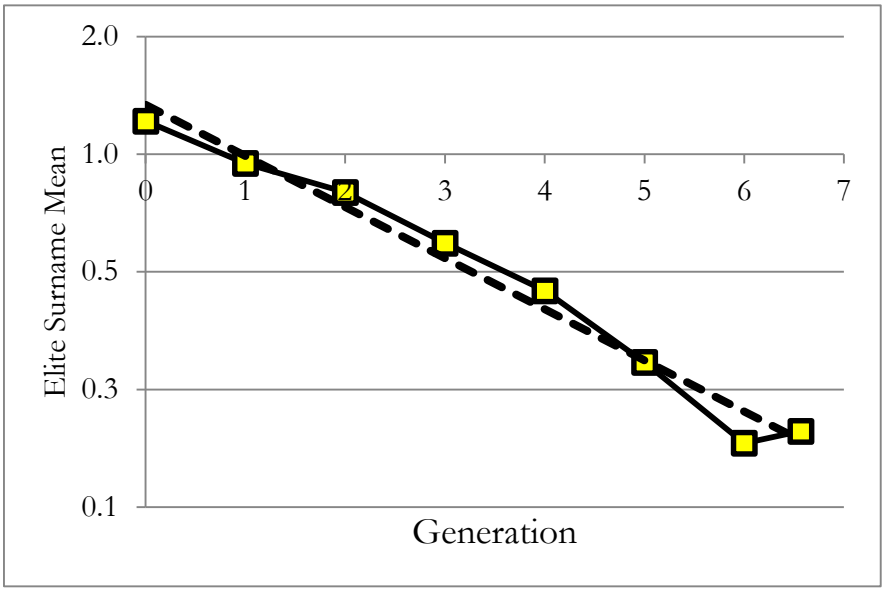
downwards. A second deficiency of the sample is that English elites in the later nineteenth and twentieth centuries have tended to form new surnames by combining the surnames of both parents. Any hyphenated surname is now much more likely to appear at Oxbridge than the average surname. These hyphenated descendants of the original surnames are not included in the sample. Thus again there will be a downwards bias in the estimated intergenerational correlation.

To eliminate some of these potential biases a rare surname sample that excluded names whose population concentrations lay outside England was constructed. Thus all names beginning "Mc" or "Mac" or "O'" were removed if of Scottish or Irish origin. Also any surname with more than 40 occurrences in the 1881 census was removed if its frequency in 2002 was more than 2.5 times the earlier frequency (the expected frequency would be 1.85). To this sample was added any hyphenated surname formed from one of the so reduced 1800-29 rare surname sample. This produces a surname sample with a population frequency about half the size of the raw sample by 2002. Table 5 shows the relative representation of surnames by generation in this amended rare surname sample, the implied mean status of the surname elite in each generation, and also the implied persistence rate of educational status. Persistence is, as expected, higher with this sample. The mean value is 0.79.

Table 5: Rare Surnames at Oxbridge, 1800-29, restricted sample

Generation	Rare Surnames 1800-29 at Oxbridge	Relative Representation	Implied Mean Status	Implied b
1800-29	4,007	14.68	1.35	-
1830-59	2,866	9.12	1.04	0.77
1860-89	3,078	7.42	0.89	0.86
1890-1919	1,538	4.74	0.65	0.73
1920-49	2,013	3.16	0.49	0.76
1950-79	2,765	2.07	0.33	0.66
1980-2009	2,483	1.58	0.18	0.62
2010-2	457	1.63	0.20	1.13

Figure 4: Best Fitting constant b estimate, restricted 1800-29 sample



If we instead estimate an average value by OLS regression that estimate is 0.70, again higher than with the OLS estimate from the raw 1800-29 sample. As figure 5 shows the fit from assuming a constant value of b over time is reasonable.

The estimates above are based on the relative representation of elite surnames relative to the average surname in England. But by the time we get to the students entering Oxbridge 2010-13, there had been substantial additions to the original English population through immigration from Scotland, Ireland, other parts of Europe, South Asia, Africa, and East Asia. Evidence from surname frequencies suggests that at least 24 percent of the 2010 English and Wales born cohort entering college have ancestors in 1800 who were not English then. This can affect estimates of modern rates of social mobility, depending on the character of the new populations entering England. If these populations are low status relative to the domestic population, their entry will make social mobility rates for the elites of the established populations seem lower. If they are of high status, it will make social mobility seem faster. Another measure of social mobility is just the movement of average social status for surnames compared to the average English surname.

Surnames ending in the letters ..bury, ..berry, ..dge, ..don, ..ham, ..land, ..ton, and ..tone, for example, are mainly English place names with endings unusual outside England. Though originally high status in the middle ages, by 1800 these names had declined to close to average status, as witnessed by their relative representation at Oxbridge being only 1.10 in 1800-29, 1.13 in 1830-59, 1.02 in 1860-89, and 1.04 in 1890-1919. We can thus take the "Locative 8" as a standard to measure the relative representation among English surnames of the rare elite Oxbridge surnames of 1800-29. Reassuringly using this measure produces an estimate of social mobility rates that are very similar to those measuring surname status against the population as a whole. This implies that immigrants to England in the years 1830-2013 have tended to have the same socioeconomic status as the "native" English population by the time we reach 2013.

In estimating b we are relying on assumptions (a)-(c) above. How reasonable are these assumptions?

Let us consider first assumption (c) that elite groups have the same variance of status as the population as a whole. If equations (3) and (4) are indeed descriptive of

the mobility process, then any elite group, no matter its initial variance of status, would soon converge close to the social variance of status over a modest number of generations. Based on equations (3) and (4) the long run variance of observed status will be

$$\sigma^2 = \sigma_y^2 = \frac{\sigma_e^2}{1-b^2} + \sigma_u^2 \quad (13)$$

If in the initial period there is no variance in y, then the variance in the subsequent periods will be

$$\sigma_1^2 = \sigma_e^2 + \sigma_u^2$$

$$\sigma_2^2 = (1 + b^2)\sigma_e^2 + \sigma_u^2$$

$$\sigma_3^2 = (1 + b^2 + b^4)\sigma_e^2 + \sigma_u^2 \ \ldots$$

For the wealth estimates for England discussed above, $b^2 \approx 0.5$, and $\sigma_u^2 \approx 2\sigma_e^2$. In this case an elite that started all at the same social level with no variance would have a variance three quarters of the social variance after one period, seven eighths after the second period, fifteen sixteenths after the third period and so on. Thus any elite that has been in existence for more than a couple of generations should have a variance nearly as great as that of the whole population.

Similarly if the elite were to start with greater variance than the general population, then even if all that extra variance comes from a greater dispersion of underlying status, which is the persistent element, there will be quick convergence on the variance of the general population. Thus if the initial variance is

$$\sigma_0^2 = \sigma^2 + \sigma_A^2$$

where σ^2 is the variance of the population, and σ_A^2 the additional variance of underlying social status, then in generation n the variance of status of the elite will be

$$\sigma_n^2 = \sigma^2 + b^{2n}\sigma_A^2$$

which, given that $b^2 \approx 0.5$, implies that within four generations less than ten percent of the excess variance in status will remain.

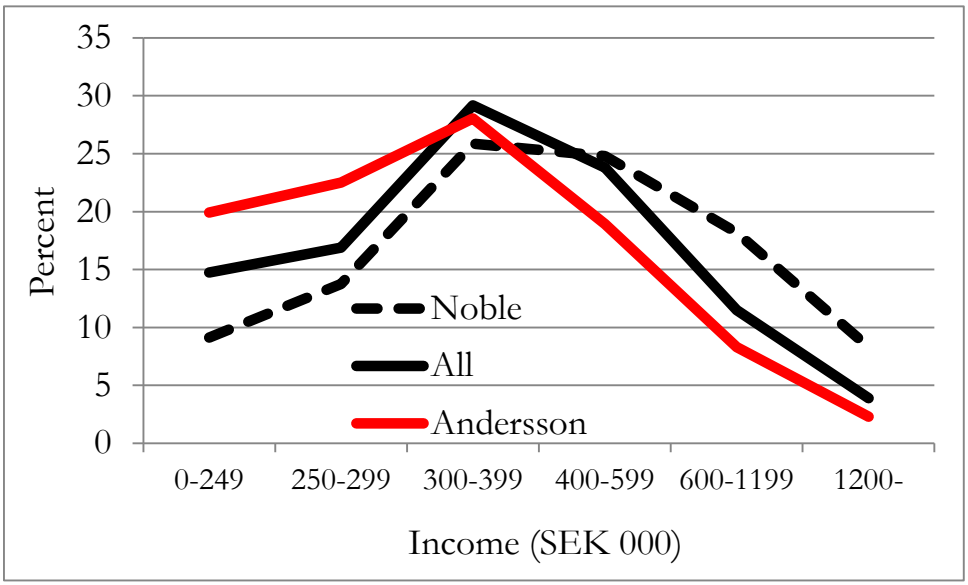
We can test this prediction of the theory of mobility, that elite groups that have existed for some generations should have a variance of status similar to that of the general population, using data from modern Sweden. In Sweden public tax records mean we can get the distribution of income for the population as a whole, for elite surname groups such as those bearing the names associated with noble families established before 1800, and for underclass surname groups such as those bearing the surname *Andersson*. Patronyms in Sweden were typically adopted by lower class families. Since the aristocratic surnames were mainly formed more than 200 years ago, seven generations before the present, they should have a variance now equal to the population mean if mobility follows the dynamic specified in equations (3) and (4). *Andersson* as a hereditable surname is more recent, but is at least three generations old. So it again should show the population variance.[13]

Figure 5 shows the distribution of each of the classes of surnames in Stockholm and five suburban towns in 2008, dividing income into the categories 0-249, 250-299, 300-399, 400-599, 600-1,199, and 1200- thousand Swedish Kroner. As can be seen the Noble and *Andersson* surnames both look like they have as wide a dispersion as all surnames, but with the mean either higher or lower than the average. Since the distribution of income in the overall surname sample is skewed with almost no incomes below 200,000 SEK, but some incomes as high as 22 m. SEK, we take logs to generate a distribution closer to normality. The standard deviation of log income for the sample of the whole surname population is 0.5. For the Noble surnames it is 0.63, and for the *Andersson* surnames 0.45. This does not accord exactly with the predictions of equations (3) and (4) above. But there is clear sign in the data of truncation of incomes at around 200,000 SEK for tax filing. The difference in standard deviations between the three groups may well just reflect this truncation.

In any case the Swedish data supports the idea that even elite groups will typically display as much variance in social status as the population as a whole. That means that the extraordinary persistence seen with surname representation among elite populations such as Oxbridge students cannot be interpreted just as reflecting persistence among a small concentrated elite.

[13] For details on the history of surnames in Sweden see Clark et al., 2014, chapter 2, 19-44.

Figure 5: Income Distribution for All Surnames, *Andersson,* and Noble Surnames, Sweden, 2008.

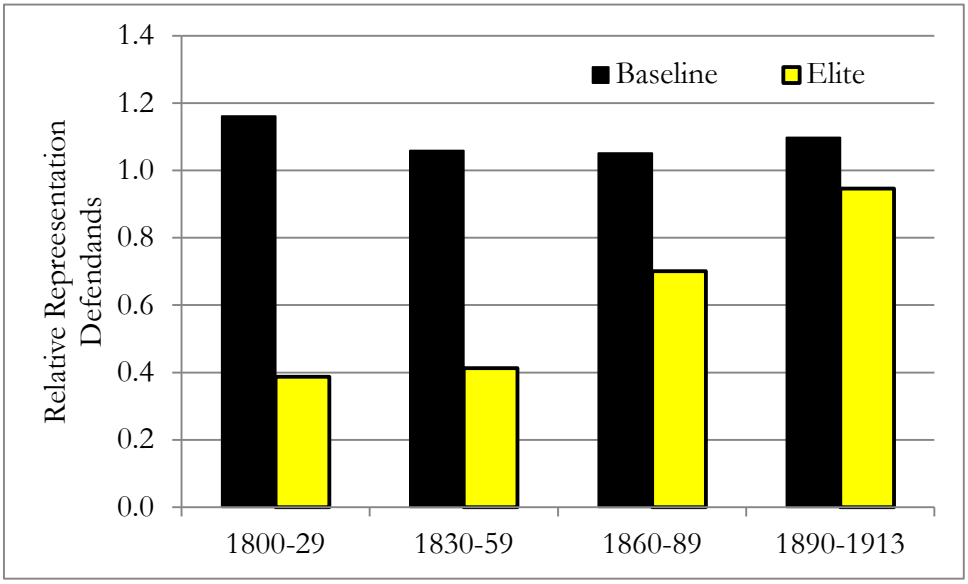


Notes: The distribution of taxable income overall was estimated from a random sample of all reported returns in these *kommuns*.
Source: 2008 tax returns for the *kommuns* of Botkyrka, Huddinge, Haninge, Nacka, Stockholm, and Täby (Kalenderförlaget 2008a,b,c).

Further evidence that elite surnames typically have as much variance in status as the population as a whole comes from England. One source we have for lower status surnames are those that appear with unusual frequency in criminal courts. Those indicted in such courts tend to be disproportionately of lower status. For England a convenient source is the *Proceedings of the Old Bailey*, published 1674-1913, which have been completely digitized. The Old Bailey was in these years the principle criminal court of Middlesex, and thus its area of jurisdiction covered most of London north of the Thames. These records encompass 253,382 defendants and 203,502 victims. We can thus measure the relative status of surnames in Middlesex 1674 to 1913 by their relative frequency among defendants and victims.

Figure 5 shows for a set of locative surname know to be of average status by the nineteenth century their frequency per defendant compared to their frequency per victim for the periods 1800-29, 1830-59, 1860-89, and 1890-1913. This number

Figure 5: Relative Representation of Surnames among London Criminal Defendants



Source: The Proceedings of the Old Bailey, 1674-1913 (http://www.oldbaileyonline.org/).

Table 6: Old Bailey Records and Implied Surname Status

Period	Relative Representation Defendants	Standard Error of relative representation	Implied Mean Status (2% cutoff)	Implied Mean Status (10% cutoff)	Implied Mean Status (Oxbridge data)
1800-29	0.333	0.025	0.42	0.55	1.22
1830-59	0.390	0.027	0.36	0.48	0.95
1860-89	0.667	0.072	0.16	0.22	0.80
1890-1913	0.862	0.102	0.06	0.08	0.59

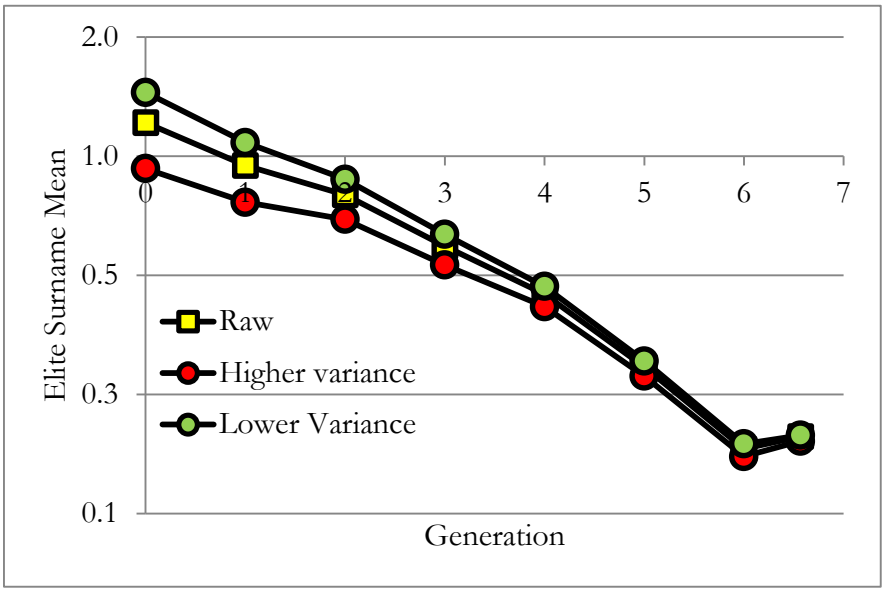
Source: The Proceedings of the Old Bailey, 1674-1913 (http://www.oldbaileyonline.org/).

averages slightly about 1.1, so these average status surnames are slightly overrepresented among defendants in London compared to victims. In comparison is shown the relative frequency by period of the rare elite Oxbridge surname sample from table 5. In the period 1800-29 these surnames are only 0.39 times as frequent among defendants as victims. Taking the locative surnames as a standard, their relative representation among criminal defendants is only 0.33 in 1800-29. This comports with the situation of the elite portrayed in figure 3, where we expect them to be underrepresented in underclass groups.

However, the information from the criminal defendants suggests that the variance of status among the surname elite must be greater than in the population as a whole in 1800-29. Table 6 shows the calculated relative representation of Oxbridge elite surnames among the Old Bailey defendants by generation from 1800, and the standard error of that relative representation. If that variance was the same as for the population then there would be even fewer of these elite surnames in the lower tail of the distribution than is observed in figure 5 in all periods. We do not know exactly what bottom share of the status distribution criminal defendants were drawn from. Table 6, however, calculates the implied mean status, in standard deviation units, of the elite surnames if defendants were either the bottom 2% or the bottom 10% of the status distribution. In either case the implied mean status from elite surname shares in this lower tail, if the variance of elite surnames was the same as for the population as a whole, is lower than their implied mean status from their shares at Oxbridge. This implies the variance of status among these surnames must be greater than for the population as a whole. An initial variance 1.5 times that of the population would fit the observed over- and under-concentration of these names at the status extremes.

Assuming that the extra initial variance was all contributed from greater variance of underlying status, figure 6 shows the implied path of mean status for the restricted Oxbridge 1800-29 rare name sample. Initial mean status will be lower in this case, but the later average status the same as before. So the implied overall persistence rate will increase in this case. The estimated average persistence rate thus rises from 0.707 to 0.744. Suppose we assumed to the contrary that the initial variance of status in the elite group was only two thirds that of the general population. Figure 6 also shows this path. Now initial mean status is higher than previously estimated, so the persistence rate is lower. But it is now 0.684 compared to the baseline estimate of

Figure 6: Implied Path of Status with different Initial Variance Assumptions

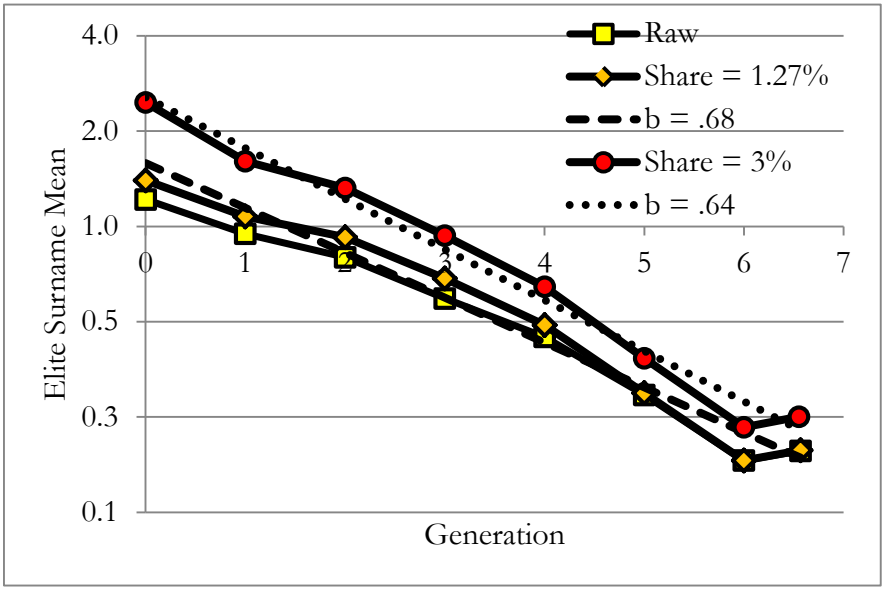


0.707. So the assumption about the initial variance level of the elite group makes little difference to the long run persistence estimate.

Assumption (a) above assumes that we know the status cutoff for Oxbridge, which is estimated through the share of each the male population cohort attending Oxbridge. This implies in table 4 that Oxbridge has become a less elite club over time, now encompassing about twice as many of each cohort. But in reality there were other avenues available to ambitious young men (and later women) aside from Oxbridge in the nineteenth century: the armed forces, the legal profession, banking and finance, commerce. The Oxbridge elite of 1800-29 might thus be a much less exclusive club than the 0.64% of the cohort estimated from the share of each cohort of males attending. Similarly now there are other excellent universities that people can attend that offer alternatives to Oxford and Cambridge, so while only 1.2% of the current generation attends Oxbridge, it may represent a much less exclusive elite than this.

Figure 7 shows the implied mean status of the restricted group of rare Oxbridge surnames of 1800-29 under the assumption that the Oxbridge share did not increase over time, and always represented the top 1.27%, or that it did not increase over

Figure 7: Paths of Implied Mean Status, Different Assumption on Size of Oxbridge Elite



time, but was always a more encompassing 3% of each cohort. These alternative assumptions do not significantly change the conclusions above that mobility rates revealed by surnames are slow, and that they show no sign of increase in the modern era. Thus if we assume that Oxbridge represented the upper 1.27% of the educational status distribution throughout, then the implied persistence parameter declines from 0.707 to 0.683. If we assume further that Oxbridge always actually represented the top 3% of the educational status distribution throughout, this implies an even lower persistence parameter of 0.640, but the same steady pattern of persistence across generations is observed.

Thus while these estimates of social mobility rates rely on assumptions (a)-(c) above, we can see that the resulting estimates of b are not highly sensitive to the precise size of the elite share assumed in each period, or to the assumed initial variance of status within the elite group compared to the population as a whole.

Social Mobility and Geography

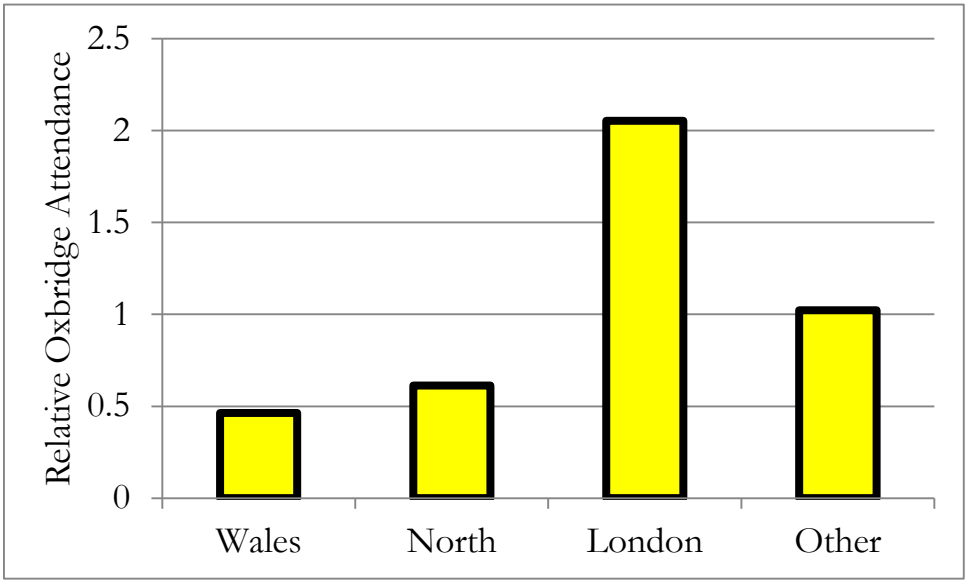
There are significant regional differences in income, educational attainment, and health in England as in many countries. For example, a recent study shows that the chance of someone attending Oxford or Cambridge depends on the region of their secondary school. Per person aged 16-17 in 2013, there were twice as many admitted to Oxbridge from Greater London than in England and Wales as a whole, half as many admitted from Wales, and only 60% as many admitted from the North. Figure 8 shows this pattern.

If geographic location affects life chances then this will be a factor influencing measured rates of social mobility. This effect can be quite important. For example, in China, despite their being only now 4,000 surnames in use by the Han population, we can detect some surnames that retain elite status even now, despite the long establishment of surnames in China. These surnames were identified as those that occurred at unusually high rates among the *jinshi* of the Qing dynesty 1820-1905. The *jinshi* were those who attained the highest rank on the exam system of the imperial eras, and their surnames have been recorded for posterity. Selecting surnames that were at least four times as frequent among jinshi as the three most common Chinese surnames we find 13 elite Qing surnames.[14] These surnames now constitute just 0.055% of the modern population, though this is 800,000 people.

Figure 9 shows the relative representation of these surnames among Qing jinshi 1820-1905, and in comparison their relative representation among later elites under the Nationalists, and now under Communism.[15] The modern elites employed are high officials in the Nationalist government in China from 1912 to 1949; professors at the ten most prestigious Chinese universities in 2012; chairs of the boards of companies listed in 2006 has having assets of US$1.5 million and above; and members of the central government administration in 2010.

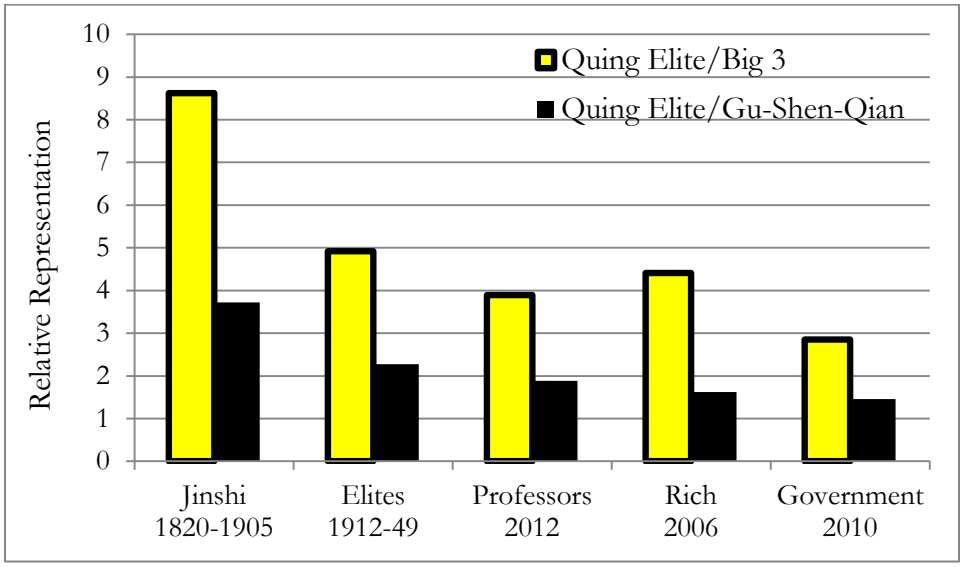
[14] Hao and Clark, 2012. The three most common Chinese surnames used for comparison, which we label the "Big 3", are *Wang* (王), *Li* (李), and *Zhang* (張), are now held now by more than 270 million people (21 percent of the population).
[15] Assuming that the population share of surnames has been stable over time.

Figure 8: Relative Oxbridge Admission Rate by Region



Source: Adams and Nye, 2013.

Figure 9: Quing Elite Surnames in later China



Source: Hao and Clark, 2012.

The relative representation in each case is measured as the fraction of these names among elites compared to the fraction of the three most common surnames. These numbers imply a very strong persistence of status even within the Communist era. The implied persistence rate for the Qing elite surnames per generation is 0.96 between the Imperial and Republican era, and still 0.92 between the Republican and Communist eras.[16]

However, the populations bearing the thirteen Qing elite surnames are all concentrated in the lower Yangzi River valley. Lower Yangzi surnames are heavily overrepresented both among exam passers in the imperial era and in modern Chinese elites. Any surname overrepresented in Jiangsu appears much more among current Chinese elites than surnames centered in such inland provinces as Sichuan. This means that in measuring social mobility we have to decide whether we want to incorporate the persistence that comes from the different economic fortunes of regions in measures of overall persistence rates.

We can control for the geographic influence in measures of social mobility, however, by instead measuring the relative representation of the thirteen Qing elite surnames among modern elites compared to three equally regionally favored surnames, *Gu* (顾), *Shen* (沈), and *Qian* (钱), the "regional three," that have only average status within the lower Yangzi. These three surnames are held by more than ten million people now, so they offer a large and stable comparison group.

Figure 9 also shows the relative frequency of these surnames in various modern Chinese elites with respect to the "regional three" surnames. Relative to the regional three surnames (*Gu, Shen,* and *Qian*), the thirteen Qing elite surnames are less overrepresented but still distinctive in the Qing and modern eras. Their relative representation was 2.28 among high Nationalist officials, 1.88 among professors at elite universities in 2006, 1.62 among chairs of company boards, and 1.46 among central government officials. Controlling in this way for geographic effects on mobility rates the estimated persistence in status between generations is still remarkably high. For the Republican Era it is 0.89, and for the Communist Era 0.82.

[16] This is assuming that the jinshi represented to top 0.05% of the Imperial population, that Republican Elites were the top 0.5% of that population, and that modern elites under Communism are again the top 0.5% of the population.

But this still leaves the question: is geography itself an important determinant of social status, or are the populations concentrated in the more successful regions of societies more capable? That is, should we control for geography in estimating social mobility rates, or is geography itself completely endogenous?

We see above in figure 8 that geography in England is seemingly highly predictive of educational success, in terms of admission to Oxbridge. But are regional effects actually playing any role as opposed to a generalized sorting of the population by region in terms of inherent capabilities? Is the North of England disadvantaged by its location, or have selective population movements resulted in a pool of inhabitants of the North of England of lower average ability?

We can test whether geography played any independent role in social mobility in England again using surnames. By the nineteenth century in England common surnames typically had close to average social status. Some such as those based on location – *Sutton, Preston, Ramsey* - were high status surnames of the medieval period, but the slow but inexorable force of regression to the mean meant they had declined to average status by the nineteenth century. We would thus expect all common surnames to have the same average status by 1800 and later.

However, such common surnames can be strongly regionally concentrated. The North of England, for example, is only about one fifth of the total population of England and Wales. But for some common surnames more than 60% of those born with the surname in England 1980-2000 were in the North: *Greenhalgh, Haworth, Heaton* and *Sutcliffe*, for example.[17] For other common names the share in the north fell well below the population share of 20%: *Church, Holloway, Oakley, Webb, Weston*.

If region of birth actually matters for socio-economic success, rather than being just correlated with the socio-economic status of populations, then we will find that now the regional distribution of such surnames will predict the average socio-economic status of the surname. But given the long history of these surnames, and the expectation that average status for common surnames will regress to the mean,

[17] The regional concentration of surnames was estimated from births recorded in the General Register Office, *England and Wales Civil Registration Indexes*, London as recorded on Ancestry.com. The north was taken as Cheshire, Lancashire, Yorkshire and all counties north of these.

then if geography has no independent effect then the locational concentration of these surnames will not matter to their occurrence rates at Oxbridge.

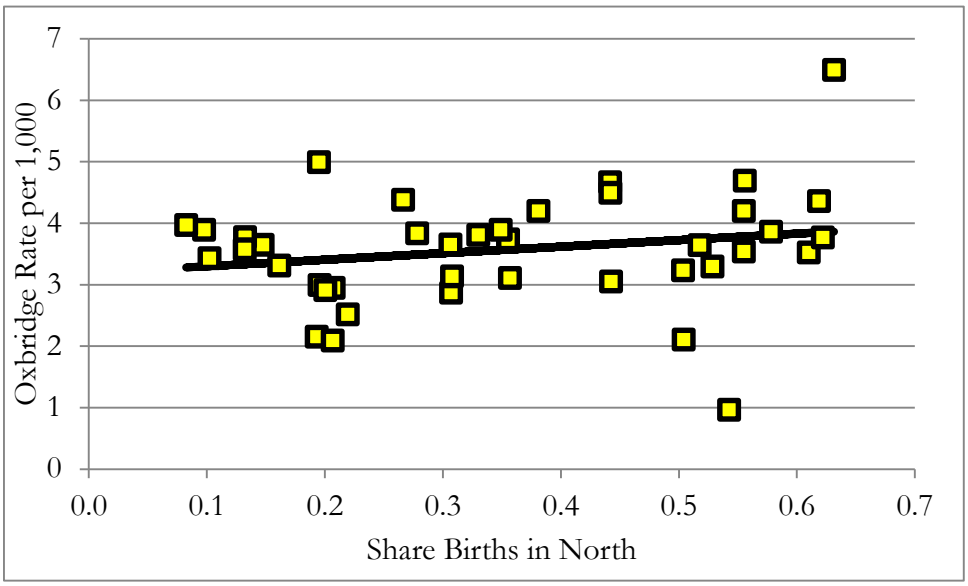
We test this by constructing a sample of 40 common surnames, with at least 7,000 births 1960-2000, which had varying degrees of regional concentration in those births 1980-2000. Figure 10 shows these surnames by the percent of births in the north, compared to the numbers of students recorded with these surnames per 1,000 births at Oxbridge 1980-2013. There is no sign that the surnames located in the North are disadvantaged in Oxbridge admissions. The expected gradient in figure 10 on the share of the surname in the North if actual regional effects explain differences in enrollment rates at Oxbridge by region should be -1.9. The actual estimated gradient is +1.1, as is shown in figure 10. Since the standard error on this estimate is 0.85, the hypothesis that there were no regional effects on Oxbridge admissions cannot be rejected. Thus the surname evidence in England supports the idea that here region of birth was actually, whatever the appearance in figure 8, not playing a significant role in determining life chances as measured by Oxbridge attendance.

We can conduct another regression test of the effect of the regional concentration of these surnames on Oxbridge attendance where we use the admission rate by region to construct for each surname an expected admission rate based on its regional concentration.[18] Thus for *Sutcliffe*, where 62% of the births 1980-2000 were in the North, the expected attendance rate is 0.78 of the average. For *Church*, where only 10% of births were in the North the expected attendance rate is 1.04. When we regress attendance rates 1980-2013 against this expected attendance rate the estimated coefficient, however, is -1.9, negative rather than positive, though the standard error is 1.58, so this coefficient is not significantly different from 0.

What can explain the absence of any regional effect on educational outcomes based on common surname locations despite the substantial regional differences in Oxbridge enrollment? It has to be that the regional differences in attainment are the product of selective migration across regions, and of selective immigration into different regions of England. A surname concentrated in the North must have seen selective migration of more elite holders to regions in the South with higher Oxbridge admissions rates, so that its overall admission rate is not affected by its

[18] The four regions used were the North, Wales, Greater London, and all other counties.

Figure 10: Oxbridge Attendance Rates and Regional Concentration of Surnames



regional concentration. And some of the regional differences in admission rates may stem not from differences in the performance of the population of English descent, but from the characteristics of immigrants from Scotland, Ireland and other countries. London, for example, has double the rate of Oxbridge enrollment than the country as a whole. But these common English surnames are all underrepresented in London. There they typically occur at only 60% of their overall share in the population. So the advantage of London in Oxbridge admissions may stem from attracting elite immigrants from outside England.

Thus while in a country as large and diverse as China geography may play an important role in determining overall social mobility rates. But the surname evidence suggests that in England there is no significant role for geography itself now.

First Names

Though this article concentrates on surnames, another potentially interesting source for historical study is first names. First names carry much more information typically about family status at the time of birth than do surnames. This is because the surname links someone to the status of some distant ancestor, while the first name gives information about the status of parents at the time of birth. As long as there are class differences in first name preferences, first names will carry status information. And class differences in name preferences seem to be a surprisingly common feature of societies.[19]

Olivetti and Paserman (2013) suggests using this status information in first names as a way of estimating the intergenerational correlation of occupational incomes from census records. Suppose in the USA the average man called Peter has an implied occupational income in the 1880 census which is 50% above average. If we take sons in 1910 who had fathers called Peter, we can then estimate the rate of regression of occupational income to the mean by looking at the average occupational income of these sons. Compared to the individual father-son links this introduces error, since the Peter's of the first generation get weighted equally no matter how many sons they have in the second generation. But using the first names does all Olivetti and Paserman to also look at the correlation of daughter's husband's occupational income with their fathers in law. Thus it allows estimation of the rate of regression to the mean in the maternal line as well as the paternal. The rates of social mobility in the nineteenth century USA seem to be the same along both lines.

The Olivetti and Paserman estimates using first names in general show high degrees of social mobility in the USA 1870-1930. Indeed, if the estimates are done at the regional level to remove the effects of regional differences in economic performance discussed above the average estimated β is 0.26, showing implied rapid rates of social mobility.[20] Here the attenuation expected from the errors introduced by this method of aggregation does seem to operate.

---

[19] Since such class differentiated first names allow educational institutions and potential employers to infer the class and even racial background of applicants, economists have been puzzled by their prevalence in such ethnically fractured societies as the USA. See Fryer and Levitt, 2004.

[20] Olivetti and Paserman, 2013, table 13.

Why when we aggregate people by surnames in England or Chile do they show more persistence than when we aggregate by first names in the USA? This is an interesting question. In terms of equation (3) above

$$y_t = x_t + u_t \qquad (3)$$

we postulate that observed social status is the combination of a persistent component x, and a transient component, u. If we aggregate across people based just on the observed status y, then the relative size of the transient component does not change, and estimated social mobility rates will not be any lower using the aggregates than in the case of individuals. Seemingly aggregating based on first names, as in Olivetti and Paserman, does not shrink the share of the transient component, and so shows as much social mobility as conventional estimates. But aggregating based on rare surnames succeeds because those bearing the surnames are related, and so have correlated values of x.

So while first names contain significant information on the social status of families, it is not clear that first names provide a basis for estimating long run social mobility rates. Also the first name method cannot be extended beyond two generations.

The differences in the current status of first names can be illustrated with the Oxbridge data. For Oxford 2008-13 we have the first names of 14,449 students with surnames of English and Welsh origin. From this we can derived the percentage distribution of first names for those matriculating 2008-13 for each gender. This we can compare with the distribution of first names for each gender for a sample of 14,813 births in England and Wales 1991-5 with surnames of English and Welsh origin, who would be 18 in 2009-2013. If we divide the Oxford share by the population share, then for each first name we get a relative representation of the name at Oxford, which represents it probability relative to an average first name of achieving entry to Oxford. This thus supplies a status ranking of first names. Table 6 shows the ten surnames with the most and least chance of appearing at Oxford for names held by at least 0.3% of each gender in the sample.

Table 6: Revealed First Name Status, 2008-13

Name	Numbers Oxford 2008-13	Share Oxford	Count Births Sample	Births %	Relative Chance of Attending Oxford
Shane	0	0.00	31	0.42	0.00
Shannon	1	0.01	47	0.63	0.02
Paige	1	0.01	46	0.62	0.02
Jade	3	0.04	127	1.71	0.03
Kayleigh	1	0.01	28	0.38	0.04
Danny	1	0.01	24	0.32	0.04
Reece	2	0.03	32	0.43	0.06
Bradley	3	0.04	39	0.53	0.07
Connor	5	0.07	61	0.82	0.08
Stacey	2	0.03	27	0.36	0.08
Stephen	83	1.21	32	0.43	2.81
John	175	2.29	60	0.81	2.82
Catherine	61	0.89	23	0.31	2.87
Richard	156	2.04	52	0.70	2.90
Elizabeth	117	1.71	42	0.57	3.01
Katherine	102	1.49	36	0.49	3.07
Anna	90	1.31	31	0.42	3.14
Simon	93	1.22	27	0.36	3.33
Peter	128	1.67	35	0.47	3.54
Eleanor	116	1.69	34	0.46	3.69

There are some common names – Shane, Shannon and Paige - which are estimated to imply the person less than one fortieth of the average chance of attending Oxford. There are other names – Eleanor, Peter, Simon, Anna, Katherine, and Elizabeth – which are estimated to make the person more than three types as

likely as the average child to attend Oxford. This means that the chance of an Eleanor born 1991-95 attending Oxford was more than 100 times as great as a Jade.

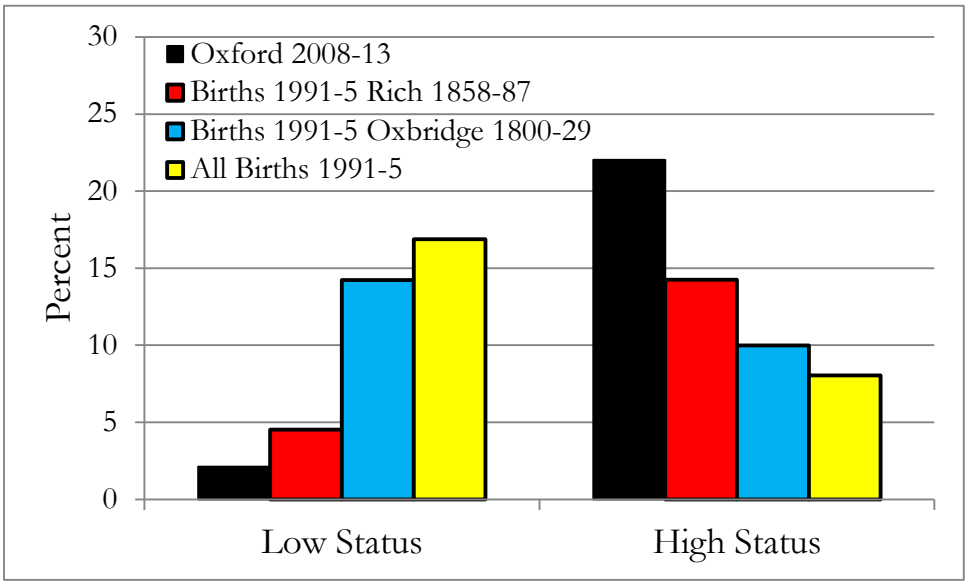
The simple act of naming reveals enormous amounts about a child's prospects, revealing again how much is determined for the child at birth. Taking names with at least 12 occurrences in the sample from the general population, those with a chance of appearing at Oxford 2008-13 less than a quarter of the average constitute 17% of the general births sample. Yet these surnames represent only 2% of students attending Oxford. In contrast surnames with more than double the frequency of Oxford than in births 1991-5 constitute only 8 percent of the population, but 22 percent of Oxford students.

Figure 11 shows these contrasting sets of surnames, labelled as "low status" and "high status" and their relative frequencies among the birth cohort, and at Oxford. Also shown in the figure is the share of births 1991-5 with these first names for a sample of rare surnames where the average holder died wealthy 1858-87.[21] The distribution of first names of those bearing these rare surnames still differ significantly from those of the general population. They have many fewer than expected of the low status first names, and many more than expected of the high status first names. This reflects the continued elite status of the descendants of this group of the nineteenth century rich.

Also shown in the figure is the share of births 1991-5 with these first names for a subset of surnames from the surname sample derived from rare surnames appearing at Oxbridge 1800-29.[22] Though they are not as distinct as the descendants of the rich, the distribution of first names of those bearing these rare surnames still differs significantly from those of the general population. Thus for births 1991-5 low status surnames are 16.9% of the general population, but only 14.2% of these higher status surnames from 200 years earlier. High status surnames are 8% of the general population, but 10% of these earlier higher status surnames. Allowing for sampling errors, both these differences are significant at the 1% level.[23]

[21] These surnames all occurred 40 times or less in the 1881 census. There were 309 children born 1991-5 with these surnames.
[22] The sample was all surnames beginning with letters a, and ba-beag.
[23] The t-statistics for the difference were respectively 3.61 and 3.16.

Figure 11: First Name Distributions Births 1991-5, Oxford 2008-13.



Though first names indicate the social status of families strongly in modern England, and thus can also indicate the average social status of surnames, they are not likely to be useful in measuring how well surnames are retaining social status over time. This is because which first names indicate high or low status changed over time in England, and also the degree to which social status was reflected in first names was lower in the past than in current England.

Thus table 7 shows for students matriculating at Oxbridge 1800-29 with English origin surnames the distribution of first names. This is compared to the the distribution of first names of men with English surnames marrying 1807-1837, for all first names held by at least 0.5% of grooms.[24] Again first names are predictive of the chance of attending Oxbridge. But the gradient in first names is notably less steep than for births 1991-5. Aside from Abraham, no common first name is associated with a less than one in four chance compared to the average of attending Oxbridge, while for 1991-5 births there are many surnames associated with a less than one in ten chance compared to the average of Oxbridge attendance. So interestingly the social distance between elites and underclasses, as revealed in first

[24] The marriages were those from parish registers and other sources recorded at https://familysearch.org/search.

Table 7: Revealed First Name Status, 1800-29

Name	Numbers Oxbridge 1800-29	Share Oxford	Count Marriage Sample 1810-1830	Marriages %	Relative Chance of Attending Oxbridge
Abraham	7	0.0	73	0.7	0.07
Jonathan	21	0.1	58	0.6	0.26
Isaac	36	0.3	73	0.7	0.35
James	628	4.4	1,164	11.5	0.38
Joseph	295	2.1	529	5.2	0.40
Stephen	43	0.3	71	0.7	0.43
Benjamin	81	0.6	116	1.1	0.50
Samuel	220	1.5	273	2.7	0.57
Peter	56	0.4	67	0.7	0.59
Daniel	54	0.4	59	0.6	0.65
David	81	0.6	86	0.8	0.67
Thomas	1,249	8.8	1,076	10.6	0.82
John	2,184	15.3	1,861	18.4	0.83
William	1,883	13.2	1,546	15.3	0.86
Robert	616	4.3	477	4.7	0.92
Richard	537	3.8	354	3.5	1.08
George	967	6.8	500	4.9	1.37
Henry	941	6.6	300	3.0	2.22
Edward	676	4.7	206	2.0	2.33
Francis	263	1.8	65	0.6	2.87
Charles	889	6.2	197	1.9	3.20

name choices, seems to have widened in England between 1800 and 1990, despite the rise of public education and common access to a variety of broadcast media. So ironically in modern England first names can serve much better as indicators of the social status of different groups than they serve in the past.

Conclusions

The status information content of surnames, and how this changes over generations, is shown above to be a useful measure in many early societies of the rate of social mobility. Clark et al. (2014) shows that surname evidence generally suggests slow rate of social mobility, slower than those estimated by more conventional methods.

In the main society discussed above, England, surnames were established long ago by 1300, and the forces of regression pulled them all common surnames towards average status. Also it is expected that the variety of surnames in any such society will decline over time because of the random elements in fertility, as was demonstrated in a famous result by Watson and Galton (1875). However, the existence of a large number of rare surnames in 1300, created in part by the vagaries of English spelling, the creation of new names by hyphenation and by imports of foreign surnames such as those of the Huguenots, means that there is still a very large number of rare surnames in modern England. In England in 2002 there were an estimated 255,000 surnames held by between 5 and 500 people. These rare surnames by random chance vary greatly in average social status, and so provide plenty of opportunity to observe social mobility rates.

In immigrant societies such as in all of the societies of the Americas the variety of national origins of the population creates again significant status differences across surnames. Even in societies such as China and Korea where there are very few surnames - about 4,000 for the Han population in China, and a mere 250 for the entire Korean population – other naming practices allow for measuring social mobility through surnames.

Thus in China up until the Communist era of 1949 and later it was common to denote people by both name and place of origin, where the place of origin was the ancestral home of the family. Though Fan is a common surname of average status, the Fan family of Ningbo was an elite descent group in the Imperial Era. By tracing the relative status over time of such surname groupings as the Fan of Ningbo we can measure social mobility rates in China through surnames (Hao and Clark, 2012). Similarly in Korea Christopher Paik points out that while surnames themselves are very common and uninformative of status, until recently people also identified

themselves by both their surname and their clan or *bon-guan* (Paik, 2012). Membership in these clans is patrilineal. There are claims that although clan membership is supposed to descend strictly through the male line, in the nineteenth century many arrivistes from lower-status groups affiliated themselves fraudulently with clans of distinguished lineage. Even if that is correct, by 1898, under the Japanese Family Registration Law all family names became fixed in Korea, so the modern surname-*bon-guan* combinations should indicate with high fidelity relationships to people born more than a hundred years ago. In total, these surname–place of origin combinations provide 3,783 distinctive family names by 2000. This is still not a large number, but these surname-place or origin combinations differ enough in status to measure social mobility rates even in Korea.

Thus the status information content of surnames, and its change across generations, seems to provide a window into a fundamental and important type of social mobility across a wide variety of societies and epochs.

References

Adams, Richard and Philip Nye. 2013. "Cambridge and Oxford places still dominated by south-east applicants." The Guardian, Sunday 9 June. Data.

Boserup, Simon Halphen, Wojciech Kopczuk, and Claus Thustrup Kreiner. 2013. "Intergenerational Wealth Mobility: Evidence from Danish Wealth Records of Three Generations." Working Paper, University of Copenhagen.

Brasenose College. 1909. *Brasenose College Register, 1509-1909*. Oxford, Basil Blackwell.

Cambridge University. 1954. *Annual Register of the University of Cambridge, 1954-5*. Cambridge: Cambridge University Press.

Cambridge University. 1976. *The Cambridge University List of Members, 1976*. Cambridge: Cambridge University Press.

Cambridge University. 1998. *The Cambridge University List of Members, 1998*. Cambridge: Cambridge University Press.

Cambridge University. 1999-2010. *Cambridge University Reporter*. Cambridge: Cambridge University Press.

Clark, Gregory and Neil Cummins. 2013. "Malthus to Modernity: England's First Demographic Transition, 1760-1800." Forthcoming, *Journal of Population Economics*.

Clark, Gregory and Neil Cummins. 2014. "Inequality and social mobility in the Industrial Revolution Era." Forthcoming in Roderick Floud, Jane Humphries, and Paul Johnson (eds.), *The Cambridge Economic History of Modern Britain*. Cambridge: Cambridge University Press.

Clark, Gregory, with Neil Cummins, Daniel Diaz Vidal, Yu Hao, Tatsuya Ishii, Zach Landes, Daniel Marcin, Kuk Mo Jung, Ariel M. Marek, Kevin M. Williams 2014. *The Son Also Rises: 1,000 Years of Social Mobility*. Book Manuscript in preparation for Princeton University Press.

Collado, M. Dolores, Ignacio Ortuño Ortín, and Anrés Romeu, 2008, "Surnames and social status in Spain," *Investigaciones Económicas*, vol. 32, no. 3, pp. 259-287.

Collado, M. Dolores, Ignacio Ortuño-Ortín, Andrés Romeu. 2013. "Long-run intergenerational social mobility and the distribution of surnames." Working Paper.

Elliott, Ivo (ed.). 1934. *Balliol College Register, 2nd edition, 1833-1933*. Oxford: John Johnson.

Emden, Alfred B. 1957-9. *A Biographical Register of the University of Oxford to AD 1500* (3 vols.). Oxford: Clarendon Press.

Foster, Joseph. 1887. *Alumni Oxonienses: the Members of the University of Oxford 1715-1886: their parentage, birthplace and year of birth, with a record of their degrees*: being the Matriculation Register of the University. 4 Vols. Oxford: Parker and Company.

Foster, Joseph. 1893. *Oxford Men and Their Colleges, 1880-1892, 2 Volumes*. Oxford: Parker and Co.

Fryer, Roland G. and Steven Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." Quarterly Journal of Economics, 119(3): 767-805.

Garza-Chapa, Raúl, María de los Angeles Rojas-Alvarado, and Ricardo M. Cerda-Flores, 2000, "Prevalence of NIDDM in Mexicans with paraphyletic and polyphyletic surnames", *American Journal of Human Biology*, vol. 12, no. 6, pp. 721-728.

Güell, Maia, Rodrigue Mora, Jose Vicente and Chris Telmer, 2007. "Intergenerational mobility and the informative content of surnames." CEPR Discussion paper No 6316.

Hao, Yu, and Gregory Clark. 2012. "Social Mobility in China, 1645–2012: A Surname Study."Working Paper. University of California, Davis.

Kalenderförlaget. 2008a. *Taxerings- och förmögenhetskalender för Stockholms kommun 2008*. Stockholm.

———. 2008b. *Taxerings- och förmögenhetskalender för Stockholms län Norra 2008*. Stockholm.

———. 2008c. *Taxerings- och förmögenhetskalender för Stockholms län Södra 2008*. Stockholm.

King, Turi E. and Mark A. Jobling. 2009. "What's in a name? Y chromosomes, surnames and the genetic genealogy revolution," *Trends in Genetics*, 25(8): 351-360.

Lasker, Gabriel W. 1985. *Surnames and Genetic Structure*. Cambridge University Press.

Olivetti, Claudia, and M. Daniele Paserman. 2013. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1930." NBER Working Paper no. 18822.

Oxford University. 1924, 1972, 1981, 1996, 2000, 2004-8, 2010. *The Oxford University Calendar*. Oxford: Clarendon Press.

Paik, Christopher. 2013. "Does Lineage Matter? A Study of Ancestral Infl uence on Educational Attainment in Korea." Working Paper, New York University Abu Dhabi.

Peña, Pablo A., 2013, "Surname Frequency and Lifespan," manuscript.

Peña, Pablo A., 2013, "Surname Frequency and Human Capital," manuscript.

Weyl, Nathaniel. 1989. *The Geography of American Achievement.* Washington, DC: Scott-Townsend.

Venn, John and Venn John. A. 1940-54. *Alumni Cantabrigienses, a biographical list of all known students, graduates and holders of office at the University of Cambridge, 1752-1900*, 6 vols. Cambridge: Cambridge University Press.

Watson, Henry William, and Francis Galton. 1875. "On the Probability of the Extinction of Families." *Journal of the Anthropological Institute of Great Britain* 4: 138–44.

Appendix

The Oxbridge Surnames Database

The printed sources for this database were Brasenose College (1909), Cambridge University (1954, 1976, 1998, 1999-2010), Elliott (1934), Emden (1957-9), Foster (1887, 1893), Oxford University (1924, 1972, 1981, 1996, 2000, 2004-8, 2010), Venn and Venn (1940-54). To get student surnames for the years 2008 and later the e-mail directories of Cambridge and Oxford were used: Cambridge: http://jackdaw.cam.ac.uk/mailsearch/, Oxford: http://www.ox.ac.uk/applications/contact_search/. The Oxford e-mail directory does not specify even implicitly the status of the people listed, which includes faculty and staff. Students were probabilistically identified as names linked only to a college, and that did not appear in all the years 2010, 2011, 2012 and 2013. Surnames of women who took courses at Cambridge 1860-1900 were obtained from http://venn.lib.cam.ac.uk/acad/search.html.

The incompleteness and informality of records at Oxford and Cambridge in earlier years, and the imperfect sources in later years such as exam results lists and e-mail directories, means that the database is necessarily always just a sample of those attending the universities.

Table A1 shows the total stock of people identified as attending Oxbridge in each generation, assumed to be 30 years. In earlier years this is just a sample of those attending the universities. From 1800 to 1892 this is a nearly complete list of all matriculating students. 1892-2009 the data is once more just a sample of all attendees. The third column shows the estimated total numbers of students in each generation. The fourth column gives the population of those surviving to age 16 in each generation from which the student population was drawn from. Before 1870 this population is assumed to be males only. Thereafter an increasing number of females attended the university, until it is assumed that by 1990 the all males and females aged 16 are potential Oxbridge attendees.

Table A1: Surnames at Oxbridge

Generation	Oxbridge Students observed	Estimated Total Oxbridge Students	Assumed Domestic Share	Population students drawn from	Oxbridge cohort share (%)
1800-29	18,649	18,649	0.99	2,246,609	0.64
1830-59	24,415	24,415	0.99	3,245,746	0.62
1860-89	38,678	38,678	0.96	7,085,936	0.53
1890-1919	30,962	47,526	0.93	9,265,992	0.48
1920-49	67,927	92,854	0.88	11,589,095	0.70
1950-79	156,645	192,254	0.86	14,209,853	1.16
1980-2009	221,196	314,956	0.76	18,838,670	1.27
2010-13	49,243	52,200	0.69	2,610,768	1.24

In later generations increasing numbers of Oxbridge students have been drawn from outside England and Wales. For 1980-2012 the Oxford University Gazette summarizes the fraction of students drawn from outside England and Wales (http://www.admin.ox.ac.uk/ac-div/statistics/student/, http://www.ox.ac.uk/gazette/statisticalinformation/studentnumberssupplements/). Cambridge has similar statistics for 2000-10. (http://www.admin.cam.ac.uk/offices/planning/sso/reporter/index.html).

Thus in 2012 only 62.3% of Oxford students were domiciled in England and Wales. In 2010 the equivalent numbers for Cambridge are 61.9%. However, many students from outside England and Wales were drawn from populations that contained substantial numbers of immigrants from England and Wales: Scotland, Northern and Southern Ireland, the USA, Canada, Australia, New Zealand, South Africa. These students constituted 14.4% of the Oxford student population in 2012. The equivalent numbers for Cambridge in 2010 were 10.5%.

We thus took the "English" surname share at Oxbridge as 69% in 2010-3, and 76% in 1980-2009. We project these foreign surname shares backwards by

measuring the share of typically German, Swedish, Dutch, Spanish, Italian, Chinese and Indian surnames at Oxbridge 1800-1979.

The final column of table A1 shows the implied share of the eligible population attending Oxbridge. From 1800 to 2013 this has varied. At its peak in 1980-2009 it was 1.27%, at its minimum in 1890-1919 it was 0.5%.

A generation is taken to be 30 years. Some studies have assumed a generation as short as 20 years for pre-industrial society. But in England from 1538 onwards the average women gave birth to her first child at age 25 or later, and the average man at 27 or later, so that the average interval for a generation would be around 30 years. If the generation length is actually shorter than this then true social mobility rates will be slower.

Population Shares

In the period 1830-2013 population shares of surnames groups for the rare surnames of 1800-29 were estimated for 4 benchmark periods, 1837-57, 1877-97, 1965-85, and 1985-95. The 1837-57 and 1877-97 benchmarks were estimated from the national register of marriages for these years, since child mortality was still significant in these years and differed by social class. The 1965-85 and 1985-95 benchmarks came from the birth register. The population share for 1830-59 for Oxbridge was taken as the 1837-57 benchmark, and that 1860-1919 from the 1887-1897 benchmark. The population share 1980-2009 came from the 1965-85 benchmark, and for 2010-2 from the 1985-95 benchmark. Population shares 1920-1979 were linearly interpolated from the shares 1877-97 and 1965-85.

For the earlier surname elites population shares 1560-89, 1680-1719 and 1770-99 were estimated from parish marriage records as recorded in Ancestry.com. For 1881 the share was estimated from the census, again as recorded on Ancestry.com. For 2002 the share was derived from the Office of National Statistics database of surname frequencies in England and Wales, as listed at http://www.taliesin-arlein.net/names/search.php. Population shares were linearly interpolated between these dates.