The Big Sort: Selective Migration and the Decline of Northern England, 1780-2018

Gregory Clark and Neil Cummins*

June 28, 2018

Abstract

The North of England is now poorer and less educated than the South. Using complete population data at the surname level 1837-2006, and a large sample of individuals born 1780-1929, this paper shows two things. First an important element in the decline of the North was selective outmigration of those with education and talent. This migration is evident even for the generation born 1780-1809, and continued to those born 1900-1929. There was also selective migration to the South of those with education and talent coming from outside England - Irish, Scottish, Pakistanis and others. However the migration of talent to the South created no significant external benefits to workers in the South, as would be predicted by the doctrines of the New Economic Geography. Surnames concentrated in the North do not show any national disadvantage in education, occupation or wealth. Also for workers of a given education or social background there is at most a very modest locational disadvantage associated with being born in the North. Thus there will be no efficiency gain from facilitating further migration south from the North, or from further efforts to bolster the economy of the North through government aid.

1 Introduction

Despite significant regional government aid, the North of England, and also Wales, lag the south in output per person, educational attainment, and life expectancy, as is indicated in Table 1 below.[1] The more extensive government aid to the North and Wales is indicated by a larger share of output being generated through public sector employment. The disadvantage of the North in terms of output per person is actually of long standing. Estimates of regional outputs per capita in England and Wales for 1871-2001 by Nick Crafts (2005) suggest the North and Wales were disadvantaged relative to the South East already by 1871 even when the staple industries of the Industrial Revolution were still booming (see figure 1). That disadvantage increased in the years since 1990, but was of much longer origin.

*Gregory Clark; UC Davis and LSE. Neil Cummins; LSE. Thanks to participants at seminars at UC Davis, LSE, Stanford, Oxford, World Cliometrics Conference, Strasbourg, Economic History Association Meetings, San Jose, University of Southern Denmark, UCL School of Slavonic Studies, European Historical Economics Society, Tübingen, University of Valencia, and the Bank of England.

[1]The North here is taken as the traditional counties of Cheshire, Cumberland, Durham, Lancashire, Northumberland, Westmorland, and Yorkshire.

Table 1: North and Wales versus South, 2012-2016

Region	Gross Value Added per person, 2015	Life Expectancy at Birth, males 2012-14	Oxbridge Offers per 1,000 aged 16-17	Average House Value, 2015	Share Public Sector Employment,
North	20,821	78.2	2.6	134,981	18.6
Wales	18,002	78.5	1.9	145,293	20.8
South	28,207	80.3	4.7	247,697	15.5

Note: North defined as the traditional counties of Cheshire, Cumberland, Durham, Lancashire, Northumberland, Yorkshire, Westmorland.

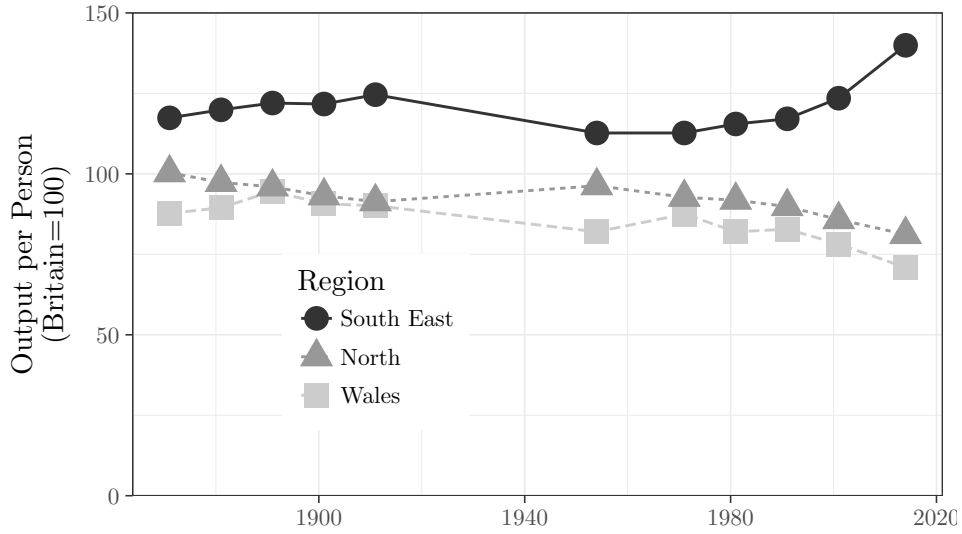

Figure 1: Output per person, North, Wales and South-East, 1871-2014

Sources: Crafts (2005), table 4, p. 59 (1871, 1881, 1891, 1901, 1911), table 6, p. 61 (1954, 1971, 1981, 1991, 2001). ONS, 2014. Geary and Stark (2015)and 2016 present alternative estimates. But these seem implausible based on the 2014 ONS benchmark.

This paper, using both population scale data at the surname level, and a large sample of individual level data, shows two things. The first is that an important element in the decline of the North lay in selective migration of those with education and talent out of the North. This selective migration is evident as early as the generation born 1780-1809, and continued at least to the generation born 1900-29. There is also evidence through surname analysis of selective migration of those coming from outside England to the South compared to the North. Those with economic abilities among the Irish, Scottish and Pakistanis, for example, were more likely to locate in the South of England.

The second thing shown in the paper, however, is that despite this selective migration of education and talent to the South, little, if any, social or economic disadvantage burdens persons residing in the North as opposed to the South. Despite the tenets of the New Economic Geography, there is no sign of significant external benefits from the concentration of education and talent in the South, or from the larger size of the London urban area than of cities in the North. Persons of a given level of education or ability did not gain any advantage from relocating from North to South. Either there is an absense of such external benefits from education or agglomeration, or these benefits are weak enough that they have been completely counterbalanced by regional aid policies.

Why there has been long standing migration of the educated out of the North is unclear. One hypothesis is that the staple industries of the North – textiles, coal, iron, and shipbuilding – were characterized by high demands for relatively unskilled labor. Thus the North retained its unskilled population, and attracted unskilled migrants, leading to a decline in the average skill and education level of the Northern population. Lower levels of output and attainment in the North may thus reflect mainly lower economic and social abilities among the resident Northern population. We do certainly see that in the interval 1870-1914, when the North already had lower output per person than the South, it was still able to attract significant numbers of migrants from the South, and from Ireland and Scotland. Figure 2, for example, shows that the North's proportion of births in England and Wales reached its maximum only in 1920. There is the possibility thus of a Roy type model of migration where the unskilled moved North, and the skilled to the South, in a period where the North had locational advantages for unskilled workers ((Roy, 1951; Borjas, 1987)).

Here we test the two competing explanations of the decline of the North. In the first there is now a locational disadvantage from living in the North. This disadvantage manifests itself for a person of given innate abilities as lesser educational attainment, poorer social outcomes, and lower economic productivity for those located in the North compared to the South. The source of this disadvantage might be in part or in all positive externalities that come from more educated people living in the South, as is posited in the doctrines of the New Economic Geography. But whatever the source the implication would be that if a person of given innate talent was moved from North to South their social and economic outcomes would be improved. This is the "bad location" hypothesis. It is such a hypothesis that lead Leunig and Swaffield (2008) to argue that ending restrictive land use policies in the South that drive up housing costs, could benefit people in the North by facilitating their relocation to the more prosperous South.

The alternative explanation of the decline of the North is that it is entirely created by selective migration of people with greater social and economic abilities out of the North and towards the South. If a person of given innate talent was moved from North to South their social and economic outcomes would not change. That is, the higher rates of unemployment, lower life spans, and lower educational attainment may just reflect the lower inherent socio-economic status of the remaining northern population. The poorer outcomes in the North are purely a matter of selection. This, rather starkly, could be labelled the "bad people" hypothesis for poorer performance in the North.

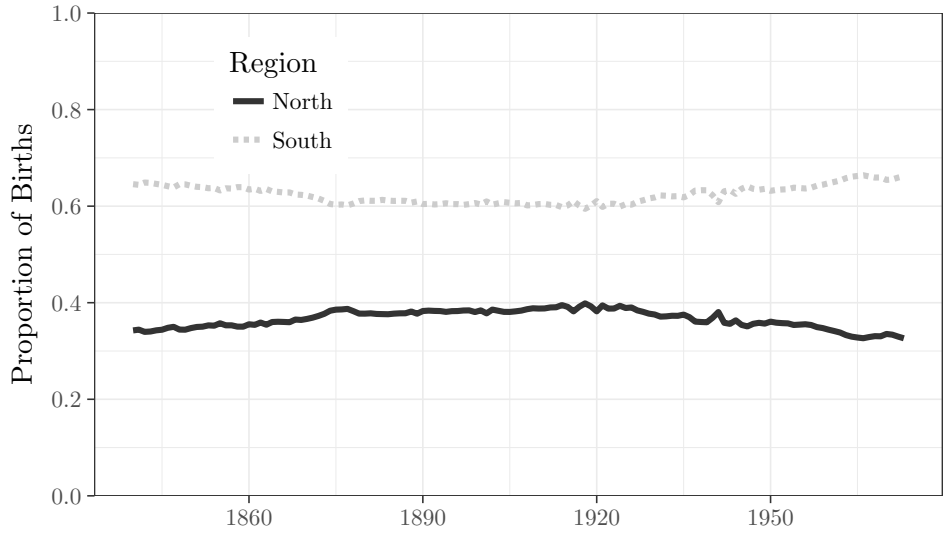

Figure 2: Proportion of Births North and South, 1838-1973
Source: 100% Sample of English Births, Marriages and Deaths, 1838-1973.

We could, of course, find that the poorer performance of the North was some combination of these two elements.

We test those competing hypotheses using the fact that many surnames in England were regionally located in the 1837 when general registration of births, deaths and marriages began. There are many "Northern" and many "Southern" surnames, which are still more concentrated in their home regions. If the first hypothesis for the decline of the north, that the location itself is now unfavorable for social outcomes or earnings, is correct we will see a general decline in the status of these Northern surnames accompanying the decline of the North. If the second hypotheses is correct, however, the decline of the North will not be associated with any general decline in the status of Northern surnames. Below we find the evidence is strongly in favor of the second hypothesis. The North declined mainly as a product of selective outmigration of talent from the region. However, interestingly, we find evidence that this outmigration of talent began even in the early nineteenth century. The seeds of later decline were planted at least from the beginning of the nineteenth century, and perhaps even earlier.

The outcomes of interest we observe at the surname level are educational performance, political representation, house values 1999, and wealth at death. Using these we can show both that ancestral Northern surnames show no disadvantage, but also that those with Northern surnames in the South are indeed a social elite, while those with Southern surnames in the North appear to be an underclass. We can also show that migrants to the North from other countries were typically of lower status than natives of the North, while migrants to the South were of higher status than the native population there.

We also have assembled a large lineage database of 277,000 individuals born in England 1750-2018 which allows us to examine the patterns of mobility for individuals born 1780-1929. This database shows clearly the selective migration from the North to the South, but also that this

selective migration was occurring even for those born 1780-1809, right at the start of the Industrial Revolution. The clear conclusion is that the decline of the north was mainly caused by the outmigration of talent from the region, and not by blight attached to the place itself.

2 Testing the Two Hypotheses

We identify for the 1840s two sets of names from the records of all deaths in England 1837-2006 as revealed in the General Registry Office indices of deaths which give name and registry district. First those where 80% or more of people dying with the name in the 1840s had deaths registered in the north of England, defined as Cheshire, Cumberland, Durham, Lancashire, Northumberland, Westmorland, and Yorkshire, or in Wales.[2] Thus circa 1840 near all Ainscoughs, Birtwistles and Calderbanks lived in the North.

We define second a set of surnames where 10% or fewer of people holding the name were dying in the North or in Wales in the 1840s. Nearly all Northcotts and Vanstones, for example, lived in the South. Figure 3 shows the proportion of people born with these surnames in the North of England 1840-1973. As late as 1973 still 62% of the ancestral northern surname holders were born in the north, and only 13% of ancestral southern surname holders were born there. So the surnames retain a distinct regional presence, even though there is convergence over time.

On the first "bad location" hypothesis for the decline of the North the ancestral northern surnames should have lower status than the ancestral southern in recent years because more holders are located in the North,and thus exposed to the unfavorable environment. On the pure selection hypothesis, there should be no difference in status between the ancestral northern and southern surnames in recent years. Also holders of ancestral northern surnames living in the south should be higher status than the other inhabitants of the South, and in particular higher status than the ancestral southern name holders in the south.

Formally if y is an outcome measure such as education, occupational statue, wealth or health, and we assume that average innate ability did not differ across the founding populations of English regions, we estimate the coefficients in the following equations:

$$y_i = \alpha + \beta_N N_i + \varepsilon_i \tag{1}$$

where y is an outcome measured at the individual level, i and N is a categorical variable which is 1 for those residing in the North of England or in Wales . We consistently observe that $\beta_N < 0$. Thus it would appear that the North suffers from a negative "Northern effect". If bad geography now in the North is the source of these poorer outcomes then when we estimate

$$y_i = \alpha + \beta_{N^A} N_i^A + \varepsilon_i \tag{2}$$

where N^A denotes is a categorical variable which is 1 where the ancestral origin of an individual's surname is the North or Wales, the coefficient $\beta_{N^A} < 0$ also, since those with ancestral Northern surnames are more exposed to the bad geogaphy of the North. Ancestral northern surnames will have lower educational attainment, and poorer occupation outcomes nationally since more of them are exposed to the adverse conditions of the North. If, however, the poorer outcomes in the North are solely the product of sorting of talent, with no consequent external effects from location, then

[2]Below we generally classify Wales as part of the North, since Wales and the North show the same economic disadvantages in the modern era.

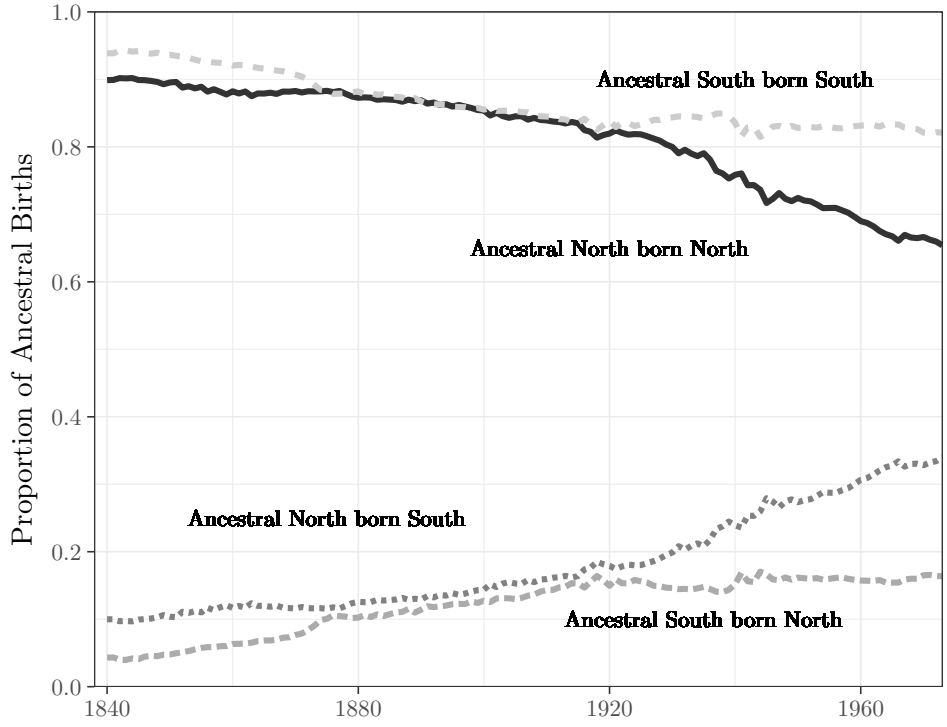

Figure 3: Inhabited Region by Ancestral Origin

Source: 100% Sample of English Births, Marriages and Deaths, 1838-1973.

Table 2: Geography versus Sorting: Hypothesised Relationships

Parameter	Geography	Sorting
β_N	$Neg.$	$Neg.$
β_{N^A}	$Neg.$	0
$\beta_{N^A_N}$	$Neg.$	$Neg.$
$\beta_{N^A_S}$	0	$Pos.$
$\beta_{S^A_N}$	$Neg.$	$Neg.$

we will find $\beta_{N^A} = 0$. Those bearing northern surnames will not show any disadvantage at the national level.

Finally we can also estimate

$$y_i = \alpha + \beta_{N^A_N} N_i N_i^A + \beta_{N^A_S} S_i N_i^A + \beta_{S^A_N} N_i S_i^A + \varepsilon_i \tag{3}$$

where N and S are a categorical variables which are 1 for those residing in the North and in the South, and where $N^A$and S^A are categorical variables which are 1 where the ancestral origin of an individual's surname is the North or in the South. If bad geography explains all the poorer outcomes of the North then we will find that $\beta_{N^A_N} = \beta_{S^A_N} < 0, \beta_{N^A_S} = 0$. If sorting without externalities explains all the effect then $\beta_{N^A_N} < 0, \beta_{N^A_S} > 0, \beta_{S^A_N} < 0$. Table 2 shows these predictions.

The key element is that if location plays any significant role in creating poorer social and economic outcomes in the North, then $\beta_{N^A} < 0$.

3 Data

The data utilized consist of two main sources. First there are 100% samples of births and marriages in England and Wales 1837-1973, of deaths 1837-2006, of probate records 1892-1992, and the electoral register 1999. There are 103,864,223 individual birth records 1837-1973, 87,215,127 individual death records 1837-2006, 14,948,900 Probate records with wealth at death 1892-1992, and 31,551,398 voter records 1999. The sources, construction and descriptive detail of this data are contained in a stand alone appendix available online at `http://neilcummins.com/BigDataAppendix.pdf`. We match these with a large sample of people attending Oxford and Cambridge 1800-2016, as well as records of registered doctors in Britain 1856-2017, and of all MPs for England and Wales 1800-2017, and property values by postal code in 2017. This allows us to look at outcomes at the surname level overall, and at the regional level for lifespan, wealth and house value in 1999/2017.

As noted above, ancestral Region for surnames is defined North where 80% or more of deaths 1840-50 for a surname were in the North or Wales. Ancestral Region is defined as South where 90% or more of deaths for the years 1840-50 were in the South. The more severe cutoff for Southern names is employed because the south represented in the 1840s 65% of all deaths, so a surname with no regional concentration would have 65% of deaths in the South compared to 35% in the North and Wales. This results in 60,158 Ancestral Southern Surnames, 27,086 Ancestral Northern Surnames.

The second main data source is the Families of England database. This database consists of 277,000 individuals born 1750-2018 who belong to one of 445 rare surname lineages, and are linked to their parents and children, and where we generally also have the place of birth and of death.

For many individuals we also have their higher educational attainment, occupation at age 40, and wealth at death. We can thus examine for people born 1780-1929 the characteristics of migration between the North and Wales, and the South of England

4 Surname Results

For the first test of the relative status nationally of Northern and Southern surnames we have a variety of outcomes we can use. These include attendance rates at Oxford and Cambridge, medical doctors per 1,000 holders of the surname, Members of Parliament per 1,000 holders of the surname, and average probated wealth per adult death by surname type.

Figure **4** shows the relative rate of enrollment of people holding these surnames to Oxford and Cambridge by decade 1850-2009. As can be seen, despite the overall lower rate of admittance of people living in the North, shown in table 1, the rates of enrollment for ancestral Northern rates actually rose over time in terms of relative representation at Oxbridge from the 1850s to the 1930s. From the 1930s to 2000s the ancestral Northern surnames have very close to the same chance as the ancestral Southern to enroll at Oxbridge. From 1950-2009 there is no statistically significant difference in enrollment rates at Oxford and Cambridge for names of Northern versus Southern origin. Those with ancestral Northern names who migrated outside the region must thus have a higher rate of enrollment at Oxbridge than average to counter the lower rate of admission of those who remained in the North revealed in table 1. It is interesting also that the period supposedly associated with the decline of the North is one where the descendants of northerners achieve full equality in Oxbridge admissions.

Figure **5** shows the number of doctors per 1,000 people holding Northern versus Southern surnames from the medical register 1859, 1883, 1911 and 1931. Also shown are the numbers of doctors currently registered by year of first registration for the periods 1940-59, 1960-79, 1980-99, and 2000-17 relative to the numbers of holders of Northern and Southern surnames in 2002. This shows a very similar pattern to that for Oxbridge. The numbers of doctors per person with each type of surname was in favor of Southern surnames in the nineteenth century, when the North was economically vibrant, but the balance shifted modestly towards the Northern surnames from 1931 on. From those registered from the 1960s onwards Northern ancestry surnames actually show a modest but statistically signifiifcant advantage in their likelihood of being registered doctors.

Figure 6 shows the relative representation of northern versus southern surnames in Parliament by decade 1800-2017, averaged across 20 year periods. Each MP is counted only on their first admission to Parliament. Here, despite the economic decline of the North, political achievement for ancestral Northern surnames was greater than for ancestral Southern surnames for all of the twentieth century. The Northern Parliamentary constituencies, however, have tended to have less population than the Southern constituencies because of faster population growth in the South between periodic revision of the constituency boundaries. This may account for some of the political success of Northern origin surnames.

We link the individual addresses in the electoral roll of 1999 to house price data by postcode in 2017 (from the land registry). There are 1,758,312 postcodes in the UK so this is a highly specific estimate of house values. Table 3 reports the average house price for regions by Ancestral surname status. As predicted on the sorting hypothesis those bearing ancestral Northern surnames living in the South have higher house values than those living in the South with ancestral Southern surnames. Also in the North those with ancestral Southern surnames have house values no greater than their ancestral Northern surname neighbors, even though the Northerners by implication are

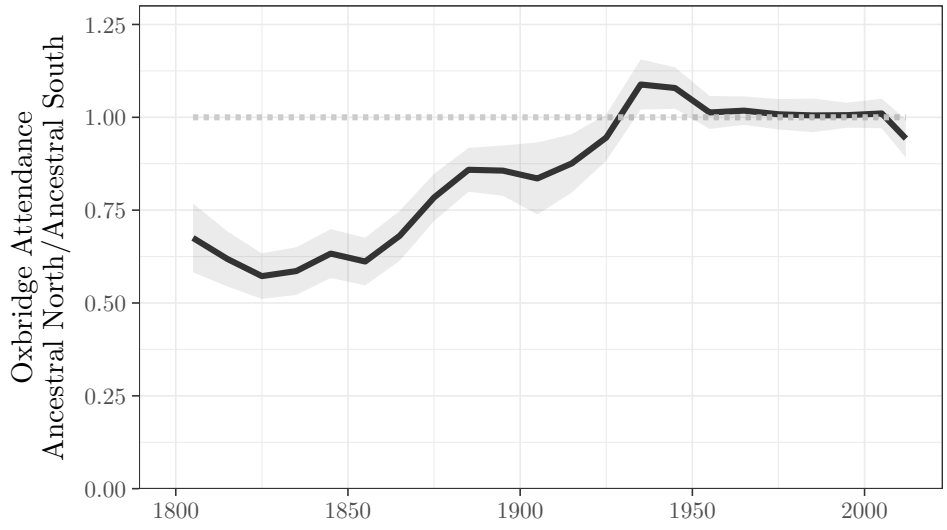

Figure 4: Relative Oxbridge Admission Rate, Northern versus Southern Surnames

Source: 100% Sample of English Births, Marriages and Deaths, 1838-1973, and Families of England Data - See appendix for details. The confidence inetrvals are constructed from the standard error of the ratio as $Var(\frac{x}{y}) = \frac{\bar{x}^2}{\bar{y}^2}\left[\frac{var(x)}{\bar{x}^2} + \frac{var(y)}{\bar{y}^2} - 2\frac{cov(x,y)}{\bar{x}\bar{y}}\right]$ where x is the ratio of Oxbridge attendees with northern surnames to holders of northern surnames, y is the ratio for Southern surnames. We assume that here $cov(x, y) = 0$, so $Var(\frac{x}{y}) = \frac{\bar{x}^2}{\bar{y}^2}\left[\frac{var(x)}{\bar{x}^2} + \frac{var(y)}{\bar{y}^2}\right]$.

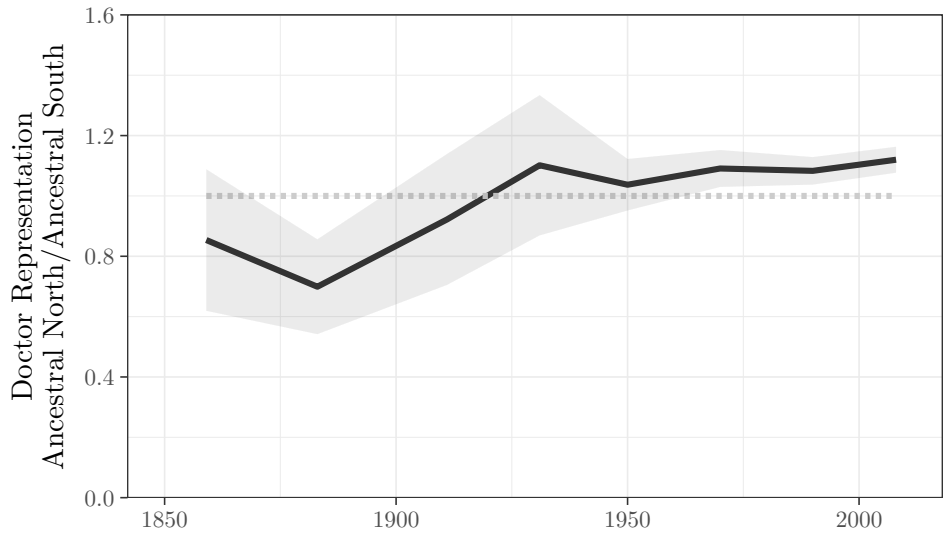

Figure 5: Relative Representation among Medical Doctors, Northern versus Southern Surnames

Source: 100% Sample of English Births, Marriages and Deaths, 1838-1973, and Families of England Data - See appendix for details. The confidence inetrvals are constructed from the standard error of the ratio as $Var(\frac{x}{y}) = \frac{\bar{x}^2}{\bar{y}^2}\left[\frac{var(x)}{\bar{x}^2} + \frac{var(y)}{\bar{y}^2} - 2\frac{cov(x,y)}{\bar{x}\bar{y}}\right]$ where x is the ratio of doctors with northern surnames to holders of northern surnames, y is the ratio for Southern surnames. We assume that here $cov(x, y) = 0$, so $Var(\frac{x}{y}) = \frac{\bar{x}^2}{\bar{y}^2}\left[\frac{var(x)}{\bar{x}^2} + \frac{var(y)}{\bar{y}^2}\right]$.

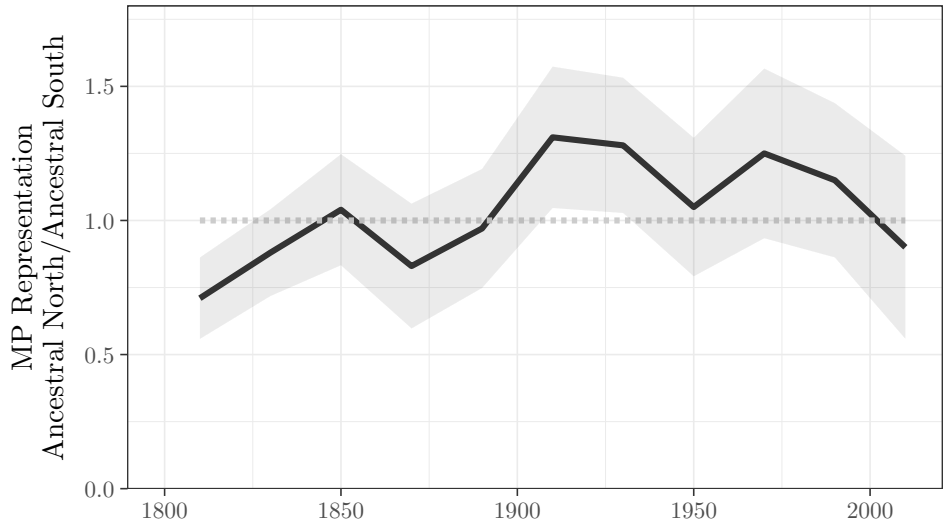

Figure 6: Relative Representation in Parliament, Northern versus Southern Surnames

Source: 100% Sample of English Births, Marriages and Deaths, 1838-1973, and Families of England Data - See appendix for details. The confidence inetrvals are constructed from the standard error of the ratio as $Var(\frac{x}{y}) = \frac{\bar{x}^2}{\bar{y}^2}\left[\frac{var(x)}{\bar{x}^2} + \frac{var(y)}{\bar{y}^2} - 2\frac{cov(x,y)}{\bar{x}\bar{y}}\right]$ where x is the ratio of MPs with northern surnames to holders of northern surnames, y is the ratio for Southern surnames. We assume that here $cov(x, y) = 0$, so $Var(\frac{x}{y}) = \frac{\bar{x}^2}{\bar{y}^2}\left[\frac{var(x)}{\bar{x}^2} + \frac{var(y)}{\bar{y}^2}\right]$.

a negatively selected group from the original northern population. Thus the Southern migration to the North, again by implication, must have been also negatively selected.

Table 3: Average Postal Code House Prices in 2017 by Ancestry, £

Surname	E&W	North	South
All	287,217	164,969	347,835
English	279,406	165,229	334,922
Ancestral North	227,147	163,767	363,333
Ancestral South	312,008	168,970	339,091
Irish	275,762	155,376	363,528
Scottish	295,928	165,829	380,549
Pakistani	306,757	146,479	387,392

With the electoral register address data we can also look at the selection of migrants from outside England to the different regions of England. Here we identify surnames associated with three significant streams of immigrants: Scottish, Irish and Pakistani. All three groups are positively selected when found in Southern England. Their house values are higher than those in Southern England with ancestral Southern surnames. But those with ancestral Irish or Pakistani surnames in the North are negatively selected, even relative to the negatively selected domestic Northern population. Those with Scottish surnames are about equivalent in house values to the native Northern population, which implies again negative selection. Thus external migrants have magnified the sorting effects within England in producing regional differences in economic and social outcomes.

We get measures of wealth at death by surname in two ways. The first is to measure probate rates by surname. In England only those whose estate was above a certain wealth threshold were probated, with this rate varying over time, but being in the order of 40% nationally from 1950 onwards. Thus probate rates are a good measure of the median wealth status of different surname types and locations. We calculate the Probate Rate as $\frac{N_{Probated}}{N_{Dying}}$ where $N_{Probated}$ is calculated from the complete Probate Registry sample, 1892-1992. N_{Dying} is calculated form the complete death sample over the same interval, as the numbers of people dying aged 21 and above in that year. Since the probate and death records both give the location of death we can calculate probate rates both by region and surname ancestry.

Figure 7 shows the probate rate of all people dying 1892 to 1992 in the north of England, relative to that of those dying in the south[3]. Throughout these 100 years probate rates, and by implication median wealth, were typically much lower for those dying in the north than for those dying in the south, in line with table 1. On average 1892-1992 probate rates in the North of England were only 0.82 of those in the South. Also shown in figure 8 is the same measure for probate rate nationally of ancestral northern surnames compared to probated ancestral southern names. Here the probate rate for ancestral Northern surnames almost equal to that for ancestral Southern surnames, averaging 0.96. The implied median wealth of ancestral northern names is very close to that of southern names, despite the South having much higher house prices. Again a pure sorting model of regional wealth differences is supported.

We can also compare probate rates for movers and stayers in each region, as revealed by surname ancestral origins. Figure 9 shows probate rates 1892-1992 for anncetral North surnames, dying in the south and the north, and for ancestral South surnames dying in the north, all as a ratio to probate

[3]Median wealth is used because the wealth distribution is highly skewed.

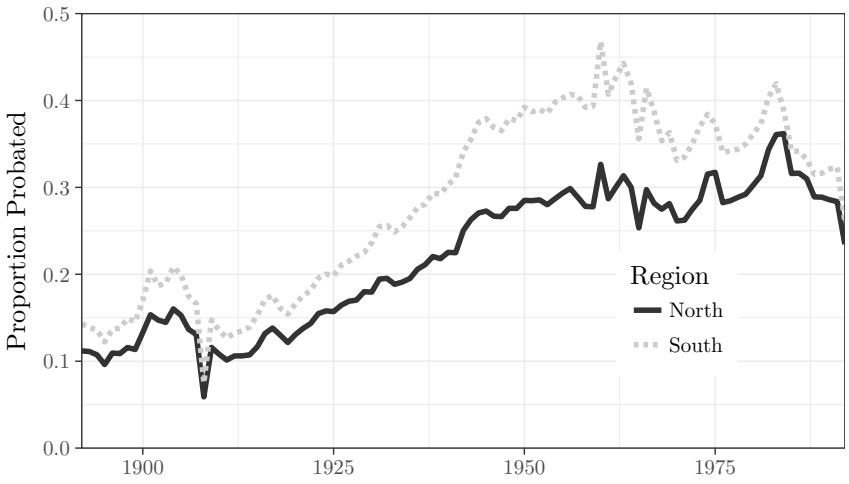

Figure 7: Probate Rate, by Region of Death
Source: 100% Sample of English Probate Calendar, 1892-1992 and 100% sample of Births, Marriages and Deaths, 1838-1973.

rates for ancestral South surnames dying in the South. Those with ancestral northern surnames were probated at higher rates in the South than their Southern counterparts, until around 1975, after which the probate rates are similar. Also Northern ancestry surnames dying in the North were consistently probated at higher rates than Southern ancestry surnames dying in the North, despite the evidence that the residual Northern population in the North was negatively selected. So again we get evidence of positive selection in the migration south, and negative selection in the migration north.

The second measure we can derive from the probate records is an estimate of mean wealth at death by region and surname type. In this we count the wealth of all those probated, and an amount equal to half the maximum wealth at which probate is no longer required. Thus

$$MeanWealth_I = ProportionProbated_i * MeanProbatedWealth_i + ProportionNotProbated_i * MinWealth_i * .5$$

This mean wealth estimate by surname ancestry and region of death is reported in figure 10. This mean wealth measure is again mostly consistent with the pure sorting story. There is clear evidence of selective migration from the North. The ancestral Northern names in the South are wealthier than their native Southern counterparts for most of the interval 1892-1980. But in the North the ancestral Northern names have wealth lower than the ancestral Southern names in the South. However, migration from South to North seems to have been negatively selective. Ancestral Southern names in the North have lower wealth than their counterparts dying in the South. In particular their wealth is no higher than the negatively selected Northern surname population in the North. So these Southern migrants to the North must also be negatively selected from the Southern population.

A further measure of social outcomes is the infant mortality rate, calculated from the complete birth and death registers for England and Wales 1866-1973 which give for deaths age at death. We calculate the Infant Mortality Rate as $\frac{N_{DyingUnder1}}{N_{Births}}$. $N_{DyingUnder1}$ is derived from the complete

Figure 8: Probate Rate, by Ancestral Region

Source: 100% Sample of English Probate Calendar, 1892-1992 and 100% sample of Births, Marriages and Deaths, 1838-1973.

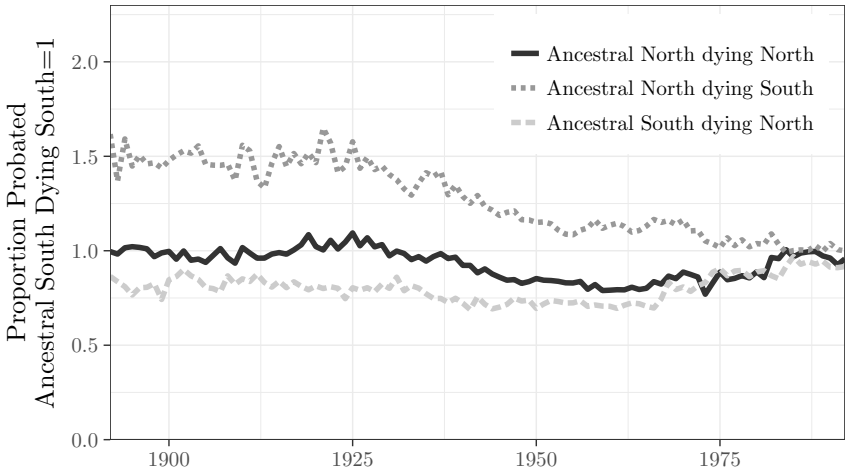

Figure 9: Probate Rate, by Ancestral Region and Region of Death

Source: 100% Sample of English Probate Calendar, 1892-1992 and 100% sample of Births, Marriages and Deaths, 1838-1973.

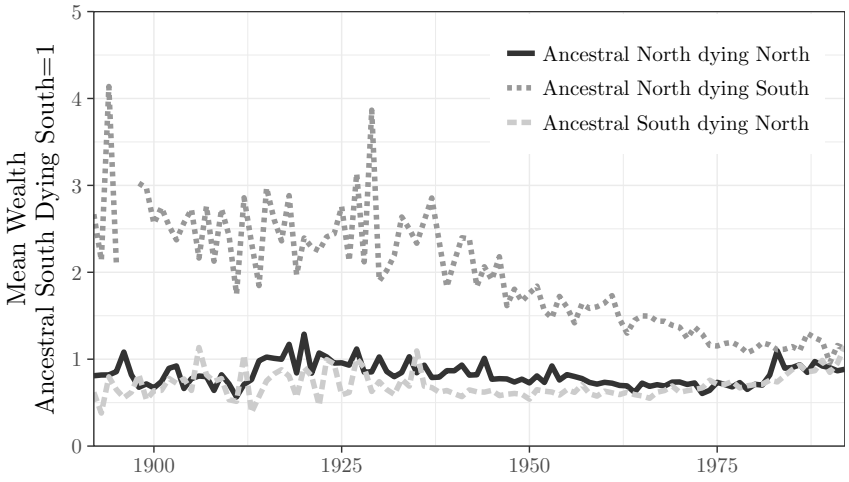

Figure 10: Mean Wealth by Region and Ancestry,1892-1973

death records, N_{Births} from the complete birth records. Here, however, the results are more consistent with geographic disadvantages in the North,and inconsistent with pure sorting by social class. All through these years infant mortality rates were higher in the Northern and in Wales than in the South, as is shown in table 4, which gives the infant mortality rates per 1,000. Infant mortality rates in the North were consistently about 20% higher than in the South 1866-1973. If we instead focus on ancestral Northern versus Southern surnames, as in column 4 of the table, then we do still find that nationally infant mortality rates are lower for ancestral Southern surnames throughout the years 1866-1973, though the differences are modesly narrower than for the pure geographic differences. So there is sign here that there were regional infant mortality rate differences, not just regional differences created by differences in average social status. This conclusion is reinforced by the final three columns of table 4 which show the infant mortality rates relative to those of ancestral Southern names in the South of ancestral Northern names in the North, ancestral Northern names in the South, and ancestral Southern names in the North. Northerners in the South have closs to the same infant mortality rates as Southerners in the South, while Southerners in the North have the same rates as Northerners in the North. So here location effects rather than sorting dominates the patterns.

We can also look at average adult lifespan by surname type, as in table 5. This shows average age at death for those dying 21 and older. Here, however, the picture is extremely confused in terms of the debate about whether poorer social outcomes in the North are the product of location or purely of sorting. The North and Wales as regions show consistently lower lifespans 1866-2006, though the gap has been narrowing. This much is consistent with the poorer social outcomes generally in the North. However, as with infant mortality there is almost as much of a Southern premium in adult longevity if we look just at ancestral Northern names, many of whom now reside in the South. Interestingly in this case when we turn to ancestral surnames from the North now located in the South we find there is still a lower lifespan associated with these, all the way through to 2006. These longevity estimates are thus inconsistent with any model of the original of differences

Table 4: Relative Infant Mortality Rates by Location and Ancestry, 1866-1973

Period	North	South	North/ South	Ancestral North/ South	Ancestral North in North	Ancestral North in South	Ancestral South in North
					Ratio to South in South		
1866-74	231	196	1.18	1.22	1.25	1.04	1.23
1875-99	222	194	1.14	1.15	1.18	1.02	1.20
1900-24	168	135	1.24	1.21	1.28	1.04	1.26
1925-49	75	57	1.32	1.22	1.34	1.02	1.30
1950-73	25	21	1.22	1.13	1.23	1.02	1.25

Source: 100% Sample of Births and Deaths, 1866-1973
Rates are Infant Deaths per 1,000 Births

Table 5: Adult Age at Death by Location and Ancestry, 1866-2006

Period	North	South	North/ South	Ancestral North/ South	Ancestral North in North	Ancestral North in South	Ancestral South in North
					Ratio to South in South		
1866-74	53.78	56.95	0.94	0.95	0.94	0.96	0.88
1875-99	55.24	58.55	0.94	0.96	0.94	0.96	0.89
1900-24	57.84	60.64	0.95	0.97	0.96	0.97	0.92
1925-49	63.74	65.66	0.97	0.98	0.97	0.97	0.95
1950-74	70.13	71.44	0.98	0.98	0.98	0.98	0.97
1975-99	74.25	75.21	0.99	0.99	0.98	0.98	0.98
2000-06	76.82	77.83	0.99	0.98	0.98	0.98	0.98

Source: 100% Sample of Deaths, 1866-2006

in social outcomes between North and South.

However, in balance the surname evidence above strongly supports the proposition that the difference in social outcomes between North and South is the productive of many years of selective migration of higher status individuals from North to South, and lower status individuals from South to North, but a migration that did not generate consequent locational effects through externalities associated with education or ability.

5 Individual Results

We have created a genealogical database for England and Wales of 277,000 individuals linked into rare surname family lineages for people born 1750-2018. This database was originally designed to estimate rates of social mobility, so it oversamples the upper and lower status groups: 14% of the people are from rich lineages, 77% average, and 9% poor. The status of each surname lineage was determined by the average wealth at death of people in the lineage dying 1858-1887. Most of these people were born in England. For those born and dying in England and Wales we typically have

Table 6: Typical Occupation Scores

Occupation	Score
Judge	100
MP	92.4
MD	87.8
Colonel Army	67.8
Dentist	54.8
Civil Servant	51.8
Teacher	46.9
Bank Clerk	35.8
Police Sergeant	28.0
Cabinet Maker	23.0
Plumber	16.3
Laborer	11.5
Coal Porter	5.0
Refuse Collector	4.0
Convict	0.4

Note: Occupational status score for 242 occupations on an index of 0-100. The index is a weighted average of average ln wealth at death by occupation, average frequency of higher educational qualifications by occupation, and average frequency of schooling/training at ages 11-20 by occupation.

both their place of birth, marriage and death. We can thus observe in this database, back to those born 1750, the migration patterns of people within England, and from England abroad.

Occupations are given in the censuses of 1841, 1851, 1861, 1871, 1881, 1891, 1901, and 1911 as well as the population register of 1939. There are also occupation statements in some marriage registers for both grooms and the fathers of the marriage parties, for fathers in birth registers, for the deceased in death registers, and also in some years for the deceased or for executors in probate records. We measured occupational status as the occupation held by the person closest to age 40. We translated these various occupational statements into 242 occupational categories – carpenter, laborer, solicitor, dealer, stockbroker etc. We gave these occupations a social status score between 0 and 100. That score was created as an equally weighted factor of three elements: average normalized ln wealth at death by occupation, average fraction of people in each occupation with a university degree or equivalent, and average fraction of males in each occupation who were in school or in training when observed ages 11-20 in the censuses of 1851-1911, and the population register of 1939[4]. Illustrative occupational scores are shown in table 6.

What are the movements of people as a function of the occupational status of their fathers? The

[4]Each factor was standardized to the same standard deviation so that they played equal weight in the final occupational score.

correlation of father occupational status with that of sons in this database is 0.65, so father's status is a good proxy for that of the son, independent of how moving might affect sons' occupational status. Also by using father status we can count also the movement of daughters, who in these years did not typically have occupations if they were married. We measure this by dividing people into cohorts born 1780-1809, 1810-39, 1840-69, 1870-99, and 1900-29. We consider only individuals where the place of birth and death is known. We can then compare, for people living to at least 21 years, the numbers born in north and south by father's occupational status compared with the numbers dying in each region. Table 7 shows the results where we divide fathers into those with low status occupations (status score 0-20), and those of high status occupations (score 30-100). Because of overseas migration (about 5%) and missing death records, there is a general shortfall of about 10% in deaths compared to births. But a consistent pattern appears, even from the earliest generation of births, 1780-1809, of those with high father occupational status showing a smaller ratio of deaths to births within the north, and a larger ratio of deaths to births in the south. In the south there is a modest gain in numbers by time of death, while in the north there is a substantial loss. For the low status occupations the reverse is true. There are gains in the north, but declines in the south. Only once we get to births 1900-29 do we see a decline in deaths compared to births in the North for the low status group. But even for 1900-29 the movement southward of the more elite occupational group is dramatically stronger than for the low status group.

Table 7: Lineage Data, Migration Trajectories, births 1870-1929

Period of Birth	Occupational Rank of Father	Born in North	Die in North	Born in South	Die in South	North Deaths/ Births	South Deaths/ Births
1780-1809	High	57	56	394	426	0.98	1.08
	Low	43	56	322	290	1.30	0.90
1810-1839	High	188	177	1,088	1,098	0.94	1.01
	Low	173	229	1,318	1,130	1.32	0.86
1840-1869	High	293	228	1,569	1,623	0.78	1.03
	Low	387	492	2,130	1,754	1.27	0.82
1870-1899	High	308	223	1,399	1,331	0.72	0.95
	Low	756	769	2,574	2,208	1.02	0.86
1900-1929	High	179	135	822	847	0.75	1.03
	Low	737	672	2,358	2,245	0.91	0.95
All	High	1,025	819	5,272	5,325	0.80	1.01
	Low	2,096	2,218	8,702	7,627	1.06	0.88

Note: High Status is an occupational score >30, low status is <20.

Table 8 summarizes the effects of a Northern birth on status outcomes of occupational rank, wealth and higher education. (The full regression tables are reported in Appendix section A.1.)

The other question we can tackle with the lineage data is whether there was a penalty from being born in the North or in Wales. For given family characteristics, were child outcomes less good if a child was born in the North or Wales compared to being born in the South? The other father characteristics we can control for here are:

Higher Education. This is an indicator variable equal to 1 (0 otherwise) if the person had a

Table 8: The Status Birth Scar of the North, births 1780-1929

Birth	Occupation Rank	ln(Wealth)	Higher Education
1780-1809	−1.21	−.004	.134
1810-1839	.34	.24	.000
1840-1869	.01	−.02	−.013
1870-1899	−.83*	−.11*	−.007
1900-1929	−2.01***	−.12*	−.015
1780-1929	−.91***	−.01	−.012**

Note:
Controlling for social status of father. See the appendix for the full details of the 15 OLS regressions
*p<0.1; **p<0.05; ***p<0.01

university education or ecclesiastical qualification, or was an attorney, doctor, chartered accountant, or member of an engineering society, or went to the military service academies such as Sandhurst.

Wealth at Death. Wealth at death comes from the Principle Probate Registry for all individuals for those dying 1858 and later. For those men dying 1825-1857 there is only a measure of wealth for those richer men probated in the Prerogatory Courts of the Archbishops of Canterbury and York. Since this censoring concerns just men of the first generation it will not create a problem for calculating the wealth correlations across generations. For the years 1858 there was a minimum wealth below which probate was not required. For these years we thus attribute to any individual not probated a wealth equal to half this minimum. This variable is normalized by estimated average wealth for all adults dying in each decade. Since the normalized variable is highly skewed we take the log of the resulting normalized measure.

For this sample of surnames we can also divide them into four groups based on average wealth at death 1858-1887: highest wealth (121 surnames), high wealth (84 surnames), average wealth (112 surnames) and lowest wealth (127 surnames).

In table 8 we estimate the average outcome for men born in the North, relative to those born in the South, in terms of occupational status, wealth and higher education attainment, controlling for their family lineage and the occupational status, wealth and higher education attainment of their fathers. We do this both for men born over the whole interval, 1780-1929, and by 30 year subperiods. If there was no disadvantage from being born in the North, which the surname evidence in section 4 suggests, then the coefficients on northern birth will be 0. In fact we do consistently find a modest negative northern effect. But this is of very small magnitude. Thus occupational status, controlling for the social status of fathers, is on average 0.9 points lower in the North than in the South on a status score than ranges 0-100, averages 26.4 and has a standard deviation of 21.0. Wealth is 1% less overall controlling for the social status of fathers, and here the difference is not statistically different. For attainment of a higher education or professional qualification the difference is 0.013%, compared to an average of 0.089%. Here the effect is more substantial quantitatively.

However the father control for family background may either undercontrol or overcontrol. If the North offers poorer opportunities, then this may be already incorporated in the parents' status. In

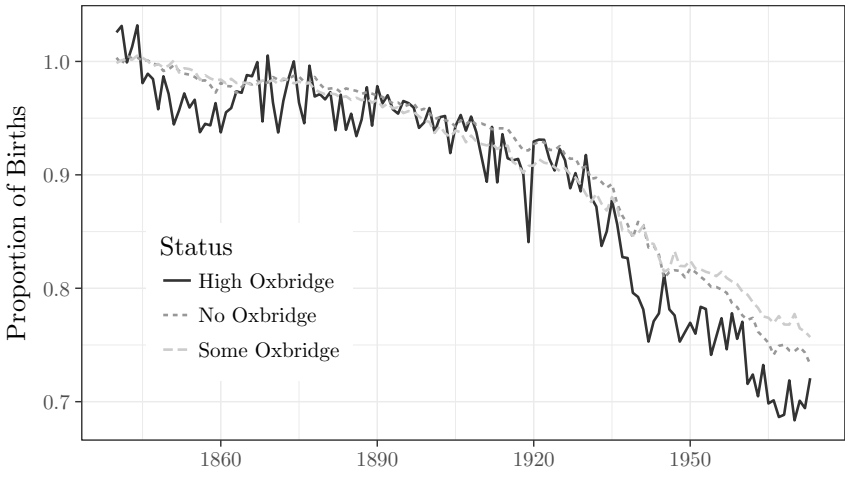

Figure 11: Proportion Births in the North 1840-1973, by surname status 1840-59
Source: 100% sample of Births, Marriages and Deaths, 1838-1973, and Families of England Data - See appendix for details.

that case we will not capture all the disadvantage of the north. If people of higher ability tend to move south, so that the unmeasured abilities of a person in a given occupation are greater in the south, then we will not have controlled fully for family background. The modest negative effects of northern birth may reflect just the imperfect proxies of occupation and education for underlying abilities. The individual data is, however, consistent with the possibility that northern birth did not create any significant disadvantage for sons in any interval from 1780 to 1929.

Confirmation of positive selection towards movement south by higher status families in the north can be found in the aggregate surname record. Figure 11 shows the proportion of births in the north by year from 1840 to 1973 for sets of ancestral northern surnames, where there were at least 100 births 1838-1859, relative to the proportion 1840-49. The first group are names where there were the most persons of this name at Oxford and Cambridge 1830-1859, by implication the most elite surnames. The second group was that where there was no-one with the surname at Oxford or Cambridge. The third group was all the intermediate surnames from the North. The higher status surnames diffuse more rapidly to the South of England than do the lower status surnames.

6 Conclusion

The poorer current economic and social outcomes in the north of England have two possible sources. The first is that the North has a locational disadvantage compared to the south that arises either from poorer access to markets, or from external economies associated with the population characteristics in the South or the production activities concentrated in the South. This creates for those located in the North disadvantaged in terms of productivity, earnings, employment opportunities, and education. The second source is the selective outmigration of those with greater economic and social abilities from the North to the South. In this paper we present strong evidence in favor of the second interpretation, both using both population level surname evidence and data on individual

families. In particular we find no evidence of any substantial locational disadvantage associated with residing in the North.

Holders of surnames concentrated in the North in the early nineteenth century were not disadvantaged in recent years in terms of education, occupation, political power, or wealth compared to the holders of surnames concentrated in the South. Since the holders of such ancestral Northern surnames are even now disproportionately located in the North, any geographic disadvantage of that area would have reduced the average social status of Northern surnames. Further holders of Northern surnames dying in the South were wealthier than holders of Southern surnames dying in the South. And there is sign that migration to the North was that of less advantaged Southerners. Holders of Southern surnames dying in the North were poorer that Northern surname holders dying in the North. These Northern surnames dying in the North were an adversely selected group, so the Southern migrants must also be adversely selected. When we look at surnames that ancestrally are of Scottish, Irish, or Pakistani origin we find again that holders of the surnames located in the North are an adversely selected group. Their house values in the North are at or below those of the adversely selected native English population there. But in the South their house values exceed those of the positively selected native population. Thus the pattern of external migration into England has also contributed to regional disparities.

The surname results are confirmed from a genealogical study of 277,000 people mostly born in England and Wales. Throughout the generations born 1780-1809, 1810-39, 1840-69, 1870-99, and 1900-29 there was net movement into the south by those with higher occupational background, and net movement into the North by those of lower occupational status. Controlling for lineage and parent status there is a very modest cost to being born in the North versus the South. For occupational status, for example, it is less than 1 point on a scale of 0-100. For wealth the cost is less than 1% relative to father's wealth. Thus we can posit a type of Roy model of locational sorting, where the skilled earned a premium in the south, and the unskilled a premium in the north.

What is not explained, however, is why there was this long-standing selective migration of the skilled out of the north. One cause may be the nature of Industrial Revolution production technologies. The staple industries of the Industrial Revolution in the north – textiles, coal mining, railways, iron and steel, shipbuilding – were characterized by heavy demands for relatively unskilled labor, and limited demands for skilled and educated labor. The industries of the south, particularly London, included insurance, banking, government, education, and trade, and the clergy, where more skilled positions were available. However, there is little evidence available on the relative wages of skilled and unskilled workers across the regions of England. And the average occupational status of workers in the 1841-1911 censuses does not seem to be significantly higher in the South as a whole than in the North. The South also had plenty of employment in unskilled service occupations.

Another possibility is that the migration of the skilled to the south was amenity based. The weather in the south is better. And there was a greater supply in the London area of many cultural amenities valued by middle and upper class families.

Whatever its source, the policy implications of the finding that the difference in outcomes between north and south are mainly explained by sorting, without substantial external benefits from this sorting, is that life chances for the current northern English population would not be significantly improved by moving more of the population to the south. The implication is also that there is no geographic disadvantage that the north faces. So further regional policies designed to compensate for a perceived northern disadvantage are misplaced. Whatever compensations the government is currently supplying have been sufficient to ensure that the ancestral northern population has no disadvantage in terms of educational and occupational attainment, and little

disability in terms of wealth.

The results here also cast doubt on a central tenet of the New Economic Geography, the existence of substantial positive externalities from the sorting of people of more education and ability into cities or regions (Marshall (1890), Jacobs (1969), Krugman (1991),Rauch (1993), Moretti (2004)). We find, despite substantial sorting by education and economic ability towards the South, no evidence of such substantial external benefits from the sorting when we compare the North and South of England. This is consistent with the literature that has empirically tested the externalities argument, such as Krashinsky (2011), where using sibling fixed effects he finds the wage premium with urban density is entirely a product of selection.

References

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth," Working Paper 24019, National Bureau of Economic Research November 2017.

Borjas, George J, "Self-Selection and the Earnings of Immigrant," *American Economic Review*, 1987, *77*, 531–553.

Crafts, Nicholas, "Regional GDP in Britain: Some Estimates," *Scottish Journal of Political Economy*, 2005, *52* (1), 54–64.

Geary, Frank and Tom Stark, "Regional GDP in the UK, 1861–1911: new estimates," *The Economic History Review*, 2015, *68* (1), 123–144.

_ **and** _ , "What happened to regional inequality in Britain in the twentieth century?," *The Economic History Review*, 2016, *69* (1), 215–228.

Jacobs, Jane, *The Economy of Cities*, New York: Vintage, 1969.

Krashinsky, Harry, "Urban agglomeration, wages and selection: Evidence from samples of siblings," *Labour Economics*, 2011, *18* (1), 79 – 92.

Krugman, Paul, "Increasing Returns and Economic Geography," *Journal of Political Economy*, 1991, *99* (3), 483–499.

Leunig, Tim and James Swaffield, "Cities Unlimited: Making Urban Regeneration Work," *Policy Exchange*, 2008.

Long, Jason and Joseph Ferrie, "Intergenerational Occupational Mobility in Great Britain and the United States since 1850," *American Economic Review*, June 2013, *103* (4), 1109–37.

Marshall, Alfred, *Principles of Economics*, London: Macmillan, 1890.

Moretti, Enrico, "Human capital externalities in cities," in J. V. Henderson and J. F. Thisse, eds., *Handbook of Regional and Urban Economics*, 1 ed., Vol. 4, Elsevier, 2004, chapter 51, pp. 2243–2291.

Rauch, James E, "Productivity gains from geographic concentration of human capital: evidence from the cities," *Journal of urban economics*, 1993, *34* (3), 380–400.

Roy, A. D., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 1951, *3*, 135–146.

Schurer, Kevin and Matthew Woollard, "1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version) [computer file]," 2000.

List of Tables

List of Figures

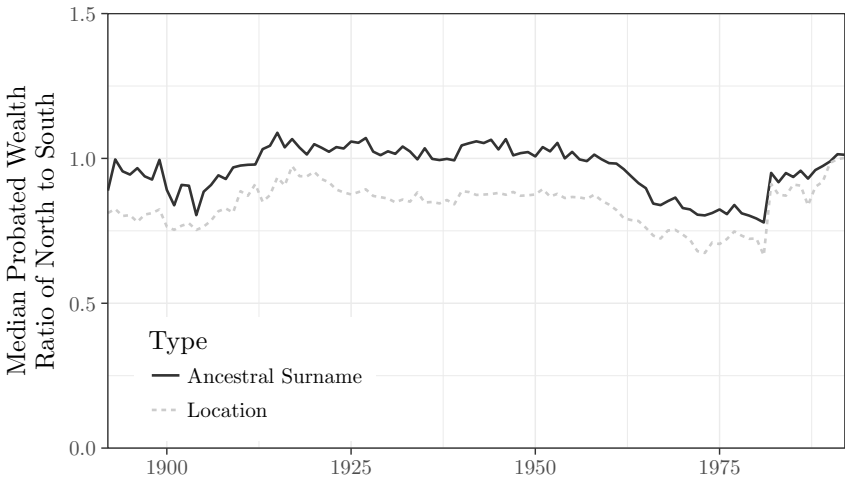

Figure 12: Probated Wealth by Region and Ancestry (dying in any region),1892-1992
Source: 100% Sample of English Probate Calendar, 1892-1992.

A Extra Results

A.1 Individual Results - Detailed Regression Tables of the effect on Status of a Northern Birth

Tables 10-14 report the details of the 15 regressions underlying table 8 in the main text.

A.2 Wealth Percentile Location

Figure 14 shows that the top 1 and top 1-10% left the North before the rest of the wealth distribiution. This offers some suggetsive support for the idea that the location of elites mattered for economic growth.

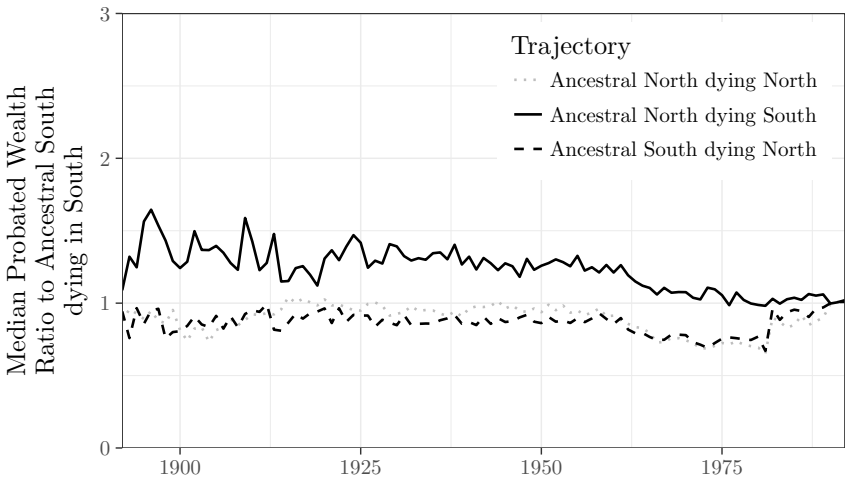

Figure 13: Median Probated Wealth by Ancestral Region and Region of Death,1892-1992

Table 9: The Penalty of the North

	Status Outcome		
	ln(Probated Wealth)		
Die in North	$-.199^{***}$		$-.209^{***}$
	(.001)		(.001)
Ancestral North		$-.031^{***}$	$.065^{***}$
		(.002)	(.002)
Ancestral South		$-.006^{**}$	$-.038^{***}$
		(.002)	(.002)
Year	$.041^{***}$	$.041^{***}$	$.041^{***}$
	(.00002)	(.00002)	(.00002)
Constant	-72.230^{***}	-72.559^{***}	-72.255^{***}
	(.038)	(.038)	(.038)
Observations	11,158,701	11,158,701	11,158,701
R^2	.287	.284	.287
Note:	*p<0.1; **p<0.05; ***p<0.01		
	Ordinary Least Squares Regression		

Table 10: The Correlation of Status and being Born in the North, 1780-1809

	Status Outcome		
	Occupational Rank	ln(Wealth)	Higher Education 1/0
Born in North	−1.208	−.004	.134
	(5.346)	(.556)	(.087)
Occupational Status, Father	.201**	.010	.003*
	(.095)	(.009)	(.001)
ln(Wealth), Father	1.058	.314***	.028*
	(.900)	(.099)	(.016)
Educated Father	1.733	−.074	.066
	(5.597)	(.529)	(.080)
Observations	136	217	259
R^2	.394	.259	.109
Note:	*p<0.1; **p<0.05; ***p<0.01		
	Ordinary Least Squares Regression		
	Controlling for lineage		

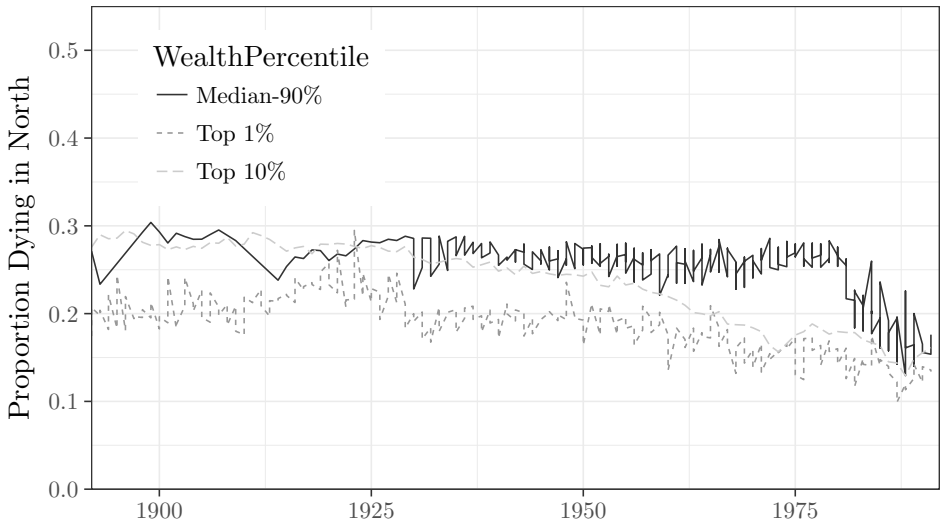

Figure 14: Proportion of Top Wealth Holders Dying in the North, 1892-1992
Source: 100% Sample of English Probate Calendar, 1892-1992 and 100% sample of Births, Marriages and Deaths, 1838-1973.

Table 11: The Correlation of Status and being Born in the North, 1810-1839

	Status Outcome		
	Occupational Rank	ln(Wealth)	Higher Education 1/0
Born in North	.339	.237	−.0002
	(1.598)	(.227)	(.028)
Occupational Status, Father	.528***	.031***	.003***
	(.039)	(.005)	(.001)
ln(Wealth), Father	1.645***	.360***	.009**
	(.222)	(.029)	(.004)
Educated Father	1.468	−.499**	.082**
	(2.210)	(.250)	(.032)
Observations	1,039	1,357	1,386
R^2	.642	.481	.160
Note:	*p<0.1; **p<0.05; ***p<0.01		
	Ordinary Least Squares Regression		
	Controlling for lineage		

Table 12: The Correlation of Status and being Born in the North, 1870-1899

	Status Outcome		
	Occupational Rank	ln(Wealth)	Higher Education 1/0
Born in North	$-.835^{*}$	$-.115^{*}$	$-.007$
	(.484)	(.068)	(.008)
Occupational Status, Father	$.402^{***}$	$.017^{***}$	$.002^{***}$
	(.017)	(.002)	(.0002)
ln(Wealth), Father	1.735^{***}	$.273^{***}$	$.013^{***}$
	(.092)	(.012)	(.001)
Educated Father	-1.089	$-.086$	$.064^{***}$
	(1.100)	(.127)	(.014)
Observations	4,298	5,000	5,179
R^2	.514	.339	.150

Note: *p<0.1; **p<0.05; ***p<0.01
Ordinary Least Squares Regression
Controlling for lineage

Table 13: The Correlation of Status and being Born in the North, 1900-1929

	Status Outcome		
	Occupational Rank	ln(Wealth)	Higher Education 1/0
Born in North	-2.011^{***}	$-.123^{*}$	$-.015$
	(.620)	(.070)	(.009)
Occupational Status, Father	$.375^{***}$	$.018^{***}$	$.002^{***}$
	(.024)	(.003)	(.0003)
ln(Wealth), Father	1.751^{***}	$.222^{***}$	$.017^{***}$
	(.145)	(.016)	(.002)
Educated Father	$-.601$	$-.155$	$.095^{***}$
	(1.561)	(.172)	(.019)
Observations	2,391	3,761	3,498
R^2	.431	.181	.158

Note: *p<0.1; **p<0.05; ***p<0.01
Ordinary Least Squares Regression
Controlling for lineage

Table 14: The Correlation of Status and being Born in the North, 1780-1929

	Status Outcome		
	Occupational Rank	ln(Wealth)	Higher Education 1/0
Born in North	$-.907^{***}$	$-.012$	$-.012^{**}$
	(.342)	(.045)	(.006)
Occupational Status, Father	$.420^{***}$	$.018^{***}$	$.002^{***}$
	(.011)	(.001)	(.0002)
ln(Wealth), Father	1.851^{***}	$.309^{***}$	$.012^{***}$
	(.061)	(.007)	(.001)
Educated Father	.202	$-.164^{**}$	$.065^{***}$
	(.692)	(.077)	(.009)
Observations	11,185	15,084	15,842
R^2	.564	.354	.147

Note: *p<0.1; **p<0.05; ***p<0.01
Ordinary Least Squares Regression
Controlling for lineage

B Families of England Database

The Families of England database aims to construct a complete genealogy of a representative set of English families from 1750 to 2018, a period of 8 generations, using public data sources. The database currently contains 277,429 individuals. The database is still very much a work under construction. The intergenerational linkages for these individuals are substantially complete for those born before 1930, but for those born later there is more work to be done on establishing these links. Currently there are 172,435 children linked with a father, 106,236 linked with a grandfather, 76,494 linked with a great-grandfather, 54,287 linked with a great-great-grandfather, 37,481 linked with a great-great-great-grandfather, 25,970 linked with a great-great-great-great-grandfather, and 18,084 linked with a great-great-great-great-great-grandfather (7 generations). However there is substantial ongoing work on establishing occupations, educational status, dwelling values, and wealth at death for each individual. Table 15 reports the currents sample sizes for various individual characteristics. We expect to add considerably more data on all the social outcome variables, and also on fertility.

To enable high linkage rate with the sources we have (described below) we adopt the strategy of following families with rare surnames, and follow descent along the male line. The vagaries of English spelling, and the varied ethnic background of the population in different parts of England, ensures that a substantial minority of the English population, even in 1800, held surnames that were shared with modest numbers of other individuals. To ensure that there is no bias in this procedure we will also link many of the daughters to their husbands, and wives to their fathers, to check that mobility and other characteristics along the female line have the same character as with the male line. Using such rare surnames we can achieve very high linkage rates between parents and children. Table 16, for example, shows the fraction of men bearing rare surnames who can be linked to their father.

For men born 1850-1949, and living to reproductive age, the linkage rate is greater than 90 percent. Typical linkage rates for historical intergenerational databases, using all surnames, at least in the US, are only around 20%.[5]. These linkages are also of high reliability in the years 1800-1930, since there are multiple sources in many cases identifying parents - censuses, birth records,

[5]Long and Ferrie, 2013, for example, link only 20% of adult sons to their fathers in England between 1851 and 1881.

Table 15: Current Status of the FOE Database

Outcome	Group	N
N	All	277,439
Death Location	All	106,121
Age at Death	All	111,164
Wealth at death	All	42,745
Higher Education	Men	32,211
Occupation	Men	27,115
School 11-20	All	15,305

Calculated from the Families of England Data

Table 16: Linkage Rates for Men born 1800-1999

Birth Period	Men 21+	Linked	Link Rate
1800-49	6,723	5,446	.81
1850-99	11,482	10,529	.92
1900-49	11,676	10,517	.90
1950-99	83,85	6,702	.80

Calculated from the Families of England Data

marriage records, passenger lists - and there are few alternative candidates who can get confused with the target individual. Thus for a sample of 7,626 recorded rare surname births 1860-1879, we identify a father or mother for 88%[6]. The reasons for failing to find at least one parent in the other 12% of cases are various. In some cases the name likely was misspelled in the birth record, and the person does not belong in the surname lineages used to form the sample. Of those not linked 60% show no further appearance in any record after their birth under the birth name. Likely in most of these cases the name is just misspelled on the birth register. In others the child dies before appearing in a census, or their father dies, or they are living with grandparents in the census, or the family emigrates[7]. Thus one third of those born not linked to a parent died before age 10. Again, in contrast, historical intergenerational databases in the US using the general population are claimed to mismatch one third of individuals to their parents (Bailey et al., 2017). A reflection of the likely high success rate in making linkages is the observed intergenerational correlation of occupational status. This is 0.67, which is much higher than that observed in other census based historical linked samples.

Though the numbers of recorded births for men and women is similar, and the match rate to fathers for the births is also similar by gender, the final dataset of family size by father is missing at least 12-14% of girls. This is because children in families can also be identified from the existence of a death record, or from their presence in a census or other record, where the birth was not recorded under the correct family surname. But adult women will only appear in a death or census record if they remain unmarried. Thus more sons are identified from such records, absent the birth record. Table 17 shows for men and women of the target rare surnames the numbers linked to fathers in total and by gender and type for births 1860-79, for all births and for those attaining age 21. Though an equivalent number of women are matched to fathers in the births sample, many more men are identified from ancillary records. This implies that at least 12% of girls are missing from the sample of births, and 14% from the sample of those attaining age 21.

To ensure a representative sample of people in each generation we have followed the strategy of including in the database all individuals bearing one of the target surnames whenever there is a birth, death or marriage record under that surname. We also try and follow the lineages of those who emigrate from England, typically to Canada, Australia, the USA, and New Zealand.

The genealogical linkages have been established in two ways. For a substantial subset of the data, 67,305 individuals we constructed the genealogical links ourselves. The other 193,690 individuals are from genealogies constructed by members of the Guild of One-Name Studies, a society devoted to studying the history and genealogy of rare surnames. The use of these Guild genealogies raises issues of selectivity, since it is more likely that a rare surname will be included in a Guild study if there is

[6]In some cases, where the child is illegitimate, only the mother is listed on birth records.

[7]We could identify the father by getting the birth certificate, but this is prohibitively costly.

Table 17: Share of Men and Women in Family Size Sample, 1860-79

	All	Men	Women
Births - all	6,205	3,218	2,987
Births - Birth record	5,826	2,877	2,949
Births - no Birth record	379	341	38
21+ - all	4,788	2,529	2,259
21+ - birth record	4,455	2,226	2,229
21+ - no birth record	333	303	30

Table 18: Missing Women by Family Size, pre-1880 marriages, children 21+

Family Size	All	All Children	Male	Female	% missing females
0	803	0	0	0	0
1	367	367	211	156	26.1
2	452	904	511	393	23.1
3	514	1,542	862	680	21.1
4 or 5	906	4,039	2168	1,871	13.7
6 or 7	554	3,560	1,876	1,684	10.2
8+	433	4,054	2,057	1,997	2.9
All	3,990	14,466	7,695	6,771	12

a current bearer of higher social status. But we can do extensive checks on the representativeness of these Guild contributed surnames, and find that at least for the 19th century they have average social status.

In both our reconstructions and those of the Guild genealogies the familial linkages - assigning fathers, mothers, and spouses - are established using a wide range of evidence. For England there are census records 1841, 1851, 1861, 1871, 1881, 1891, 1901, 1911. There is the Population Register of 1939. There is the register of births, deaths and marriages 1837-2005. The birth register 1912-2005 gives the surname of the mother. There are selective parish registers of births and marriages 1750-1930. There are probate records nationally, 1858-2018, and for the Canterbury and York Ecclesiastical courts 1750-1858. There are passenger lists for those leaving the UK 1890-1960, and for those entering the UK 1878-1960. There are Electoral Registers 1900-2012.

In recalcitrant cases in England we can, at cost, order the actual birth certificate which list the father and mother, or marriage certificate which lists marriage partners, their occupations and those of the fathers. We plan on doing this for a select sample of people marrying around 1990, so that we can get their occupational status, where they would typically be born circa 1960, as well as the occupational status of their fathers born circa 1930.

It is possible in many cases to check proposed familial linkages against genealogies uploaded by ancestry.com members. These genealogies are not always reliable. But the better ones cite source documents which can be inspected to see if the link is sound.

Ancestry.com records the age at death of many migrants from the England to Canada, Australia, NZ and USA. For Australia the voting rolls 1903-1983 give occupations. For the US the censuses 1850-1960 record occupations. Canada and New Zealand also have some occupational information from voting rolls. However, wealth at death is generally not available for migrants to these countries.

The social status indicators we have are age at death, wealth at death, schooling, occupation, location, and first names of children.

Wealth at Death: For England and Wales the Principle Probate Registry records whether someone was probated, and the value of their estate for all deaths in England 1858-2018. This information is the most comprehensive and unusual outcome result that we have for this database. The probate information is searchable at https://probatesearch.service.gov.uk/#wills. However, the estate values 1996-2018 are now obtainable only at cost of 10 pounds per person.

Schooling and Training: The censuses of 1851-1911, and population register of 1939, record whether anyone aged 10-19 is still attending a school, which gives us a measure of education for the earlier years. From the previous NSF project we have a database of all students who attended Oxford or Cambridge, 1750-2015. But this constitutes only 1-2% of each cohort. Complete records are available for attendees at the Royal Military Academy Woolwich (1790-1839) and Royal Military College Sandhurst (1800-1946). Complete records are available for Masters and Mates Certificates, 1850-1927, UK Medical Registers, 1859-2015, UK, Civil Engineer Lists, 1818-1930, UK, Electrical Engineer Lists, 1871-1930, UK, Mechanical Engineer Records, 1847-1930, UK, Articles of Clerkship, 1756-1874. From all these measures we can construct indices of educational attainment for people in the database born before 1900.

Occupation Status: The censuses of 1851-1911, and the Population Register of 1939 record occupations, so we can estimate adult occupations for the cohorts born 1920 and before. Passenger lists give occupations for international travellers up to 1960. Birth certificates record the occupation of father's, and from 1995 on that of mothers also. Marriage certificates record the occupations of husband and wife, and of fathers. So for a select sample we can estimate occupations for people born up to around 1980.

Dwelling Value: From the electoral census of 1999-2012 we have the address where adults were living in 1999-2012, from which we can infer using the Land Registry the property value in 2017. This gives an indirect measure of family income.

Children's First Names: Children's first names are a good proxy for family social status in modern generations. Using records of Oxbridge attendance and property values we can assign status measures to parents based on their child name choices.

After completing the genealogical links, and the status information, we will have potentially the following information for each person in the database

Date of birth, longevity, wealth at death, educational attainment, occupation, birth location, fertility, child mortality, death location, birth order, number of siblings, age at marriage.

C Data Sources

C.1 Births, Deaths, Marriages

General Register Office. England and Wales Civil Registration Indexes. London, England: General Register Office.

FreeBMD. England & Wales, FreeBMD Death Index: 1837-1915 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2006. Available at www.ancestry.com (last accessed: 01 Apr 2016).

FreeBMD. England & Wales, FreeBMD Birth Index, 1837-1915 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2006. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. England & Wales, Birth Index: 1916-2005 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2008. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. England & Wales, Death Index: 1916-2006 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2007. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. Public Member Trees [database on-line]. Provo, UT, USA: www.ancestry.com Operations. Inc., 2006. Original data: Family trees submitted by Ancestry members.

C.2 Social Outcomes

Censuses, 1841-911 and Population Register 1939

England and Wales, Censuses, 1841-1901. Available online at `http://www.nationalarchives.gov.uk/records/census-records.htm`.

England and Wales, Population Register, 1939. London, UK. 1939 Original Data: 1939 Register, 1939. Kew, Surrey, England: The National Archives of the UK (TNA). Available at `http://www.findmypast.com/1939register`.

The 1881 census of England and Wales was downloaded from `https://discover.ukdataservice.ac.uk/catalogue/?sn=4177&type=data%20catalogue` (Schurer and Woollard, 2000).

C.3 Wealth at Death

England and Wales, Index to Wills and Administrations, 1858-2012. Principal Probate registry, London (available online 1858-1966 at Ancestry.co.uk).

Gov.UK. Wills and Probate 1858-1996 [database on-line]. Original data: Principal Probate

Registry. Calendar of the Grants of Probate and Letters of Administration made in the Probate Registries of the High Court of Justice in England. London, England © Crown copyright. Available at https://probatesearch.service.gov.uk/Calendar#calendar (last accessed: 01 Apr 2016).

General Register Office. 1861. Annual Report of the Registrar General 1859. (pp. 173-181) Available online at http://www.histpop.org (last accessed: 01 Apr 2016).

Prerogative Court of Canterbury and Related Probate Jurisdictions: Probate Act Books. Volumes: 1850-57. Held at the National Archives, Kew. (Catalogue Reference: PROB 8/243-250.)

C.4 Educational Status

Ancestry.com. UK, Articles of Clerkship, 1756-1874 [database on-line]. Provo, UT, USA: www.ancestry.com Operations, Inc., 2012. Original data: Court of King's Bench: Plea Side: Affidavits of Due Execution of Articles of Clerkship, Series I, II, III (KB 105-107). The National Archives, Kew, Richmond, Surrey. Registers of Articles of Clerkship and Affidavits of Due Execution (CP 71). The National Archives, Kew, Richmond, Surrey. Available at www.ancestry.com(last accessed: 01 Apr 2016).

Ancestry.com. Cambridge University Alumni, 1261-1900 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 1999. Original data: Venn, J. A., comp.. Alumni Cantabrigienses. London, England: Cambridge University Press, 1922-1954. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. Oxford University Alumni, 1500-1886 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2007. Original data: Foster, Joseph. Alumni Oxonienses: The Members of the University of Oxford, 1715-1886 and Alumni Oxonienses: The Members of the University of Oxford, 1500-1714. Oxford: Parker and Co., 1888-1892. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. UK, Crockford's Clerical Directories [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2009. Original data: Crockford Clerical Directory. England: 1868-1932. See title page image of each directory for original source information. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. UK, Electrical Engineer Lists, 1871-1930 [database on-line]. Provo, UT, USA: Ancestry.com Operations, Inc., 2014. Original data: Society of Telegraph Engineers (later, Institution of Electrical Engineers) Membership Lists, 1887–1930. Institution of Engineering and Technology, Savoy Place, London, England. UK, Civil Engineer Lists, 1818-1930. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Ancestry.com. UK, Mechanical Engineer Records, 1847-1930 [database on-line]. Provo, UT, USA: Ancestry.com Operations, Inc., 2013. Original data: Mechanical Engineering Records, 1847-1930. London, UK: Institution of Mechanical Engineers. Available at v (last accessed: 01 Apr 2016).

Ancestry.com. UK Medical Registers, 1859-1959 [database on-line]. Provo, UT, USA: Ancestry.com Operations Inc, 2008. Original data: General Medical Council, comp. UK Medical Registers, 1859-1959. London: General Medical Council, 1859-1959. This data is provided in partnership with the General Medical Council. Available at www.ancestry.com (last accessed: 01 Apr 2016).

Royal Military Academy Sandhurst. The Sandhurst Collection [database on-line]. Available at http://archive.sandhurstcollection.co.uk/ (last accessed: 01 Apr 2016).

Senate House Library, University of London. 2016. University of London Students 1836-1934 [database on-line]. Available at http://www.senatehouselibrary.ac.uk/our-collections/special-collections/

archives-manuscripts/university-of-london-students-1836-1934 (last accessed: 01 Apr 2016).