Prepared with SEVISLIDES

Intergenerational Mobility and the Informative Content of Surnames

Maia Güell (University of Edinburgh)
José V. Rodríguez Mora (University of Edinburgh)
Chris Telmer (CMU)

SIRE Conference, 19/11/2007

Summary *(1/2)*

Summary *(2/2)*

Motivation *(1/2)*

- It is difficult to measure mobility.
- Traditionally. Panel data correlation of parents and children.
  - Difficulty gathering panel data
    - impossible for LDC... you have to start now...
  - Difficult to get a direct measure of inheritance...
- But even if you have a panel (Solon 92,02,...):
  - Upward mobility bias because of noise.
  - Sample selection & Attrition.
- Difficult to **compare** across countries and across time.
  - and comparisons are the only meaningful thing.
- We do not know the *economic meaning* of mobility:
  - How does it correlate with Income, Growth, Inequality,...

Motivation *(2/2)*

- **Methodological point**: surnames allow to study many longitudinal questions without panel data (nor family linkages between individuals)

- Build a model that explain relationship between Informational Content of Surnames and inheritance (mobility).
  - It allows to build an alternative methodology to the study of inheritance.

- We show the the methodology works: Use Spanish Census Data, which allows to do tests that go beyond our methodology.
  - We show a fall in the amount of intergenerational mobility.
  - We obtain the same result when looking at the evolution of sibling's correlations.
  - We can explain it by an increase in the degree of assortative mating.

Intuition 1 (Inheritance)

- Surnames are a link between the present and the past, as they are inherited from parents.
- Surnames do not affect your income

... but are inherited along variables that affect your income.

- You do not observe these variables (how much economically meaningful inheritance did the individual get)
- ... but you observe the surname.
- The more information the surname has... it is because the more importance the OTHER variables that are inherited have in order to determine socioeconomic position.
- Thus, more informative content of surnames, less mobility.

Intuition 2 (frequency,birth,death)

- It works because surnames have a very skewed distribution:
  - a few surnames are very common,
  - but the huge majority of surnames are very unfrequent,
  - and most people have quite unfrequent surnames.
- For many people in society, those sharing their surname are substantially more likely to be family related that people than does not share the surname.
- Surnames define a partition of society that is informative about **family links** not *only* about ethnicity or migration origin.
- This is a consequence of the process of birth and death of surnames. It applies to most western naming conventions.

Usefulness

- This is a **very general point**, applicable to any problem of inheritance whenever:
  - you want to know the impact of the background (the parents) on a variable of the children.
  - but you do not observe the amount inherited, only the outcome on the children.
  - ... and you have available extensive information of the distribution of outcomes per surnames... which is quite common.
  - **Health, . . .**
- **Drawback:** We can say only how much inheritance matter, not discern whether is of one type or another (nature vs. nurture).

A Model of Surnames and Income

- Non-Overlapping generation model of Income and Surnames transmission and generation.
- In each period generates a "census":
  - pair $(surname, income)$ for each individual.
- **Income transmission process: Inheritance**
- **Surnames** are (almost always) inherited.
  - Death and Birth of lineages.
- **Informative content of surnames**. R^2. Fake Surnames.
- **Objective**: See whether ICS increases with ρ.
- **Procedure**
- **Results**

Income

- Income generation process: $Y_t = \rho Y_{t-1} + u_t$
- ρ is the measure of inheritance.
- Siblings have smaller conditional variance than population.

Surnames


- Model of only males (reproduction, but no females).
  - Surname inheritance via male line.

- Surnames (lineages) are born and die:
  - DEATH: The last male with the surname has no male children.
    - Probability of having (male) children: Q
      - Surname dies with prob $1-Q$ if you are only one with it.
    - Conditional on having children, the number of male children is fixed and equal to M
      - The expected number of male children: $E = QM$
  - BIRTH: "mutation".
    - You have surname of your father, unless you have a mutation.
    - Exogenous, constant arrival rate, μ.

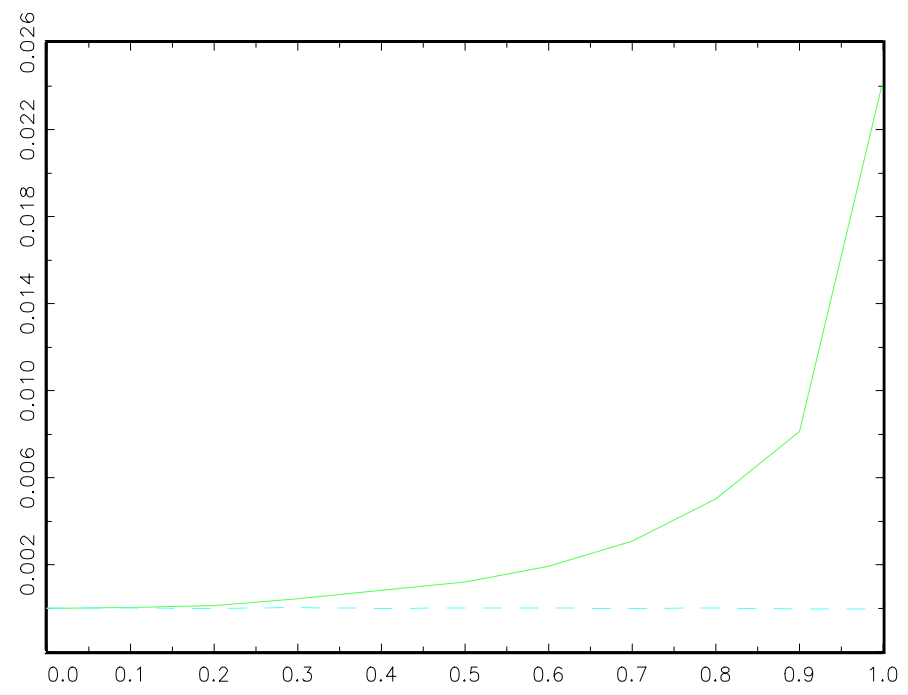



Informative content of surnames

- $R^2_{surname}$ of regression with $income$ in LHS, $surname$ in RHS.
- "Fake Surname": Non family-related partition of the economy with the same distribution than surnames.
  - "Fake surname" taken from a distribution that is identical to the one of the simulated economy (skewed).
- R^2_{fake} of regression with $income$ in LHS, $fake\_surname$ in RHS.
- Define Informational Content of Surnames as $R^2_{surname} - R^2_{fake}$
  - If only one surname: ICS=0
  - If one individual per surname: ICS=0
  - If the information is because partition, not because family: ICS=0
- **Reason:** Asymptotic properties are tricky (infinite sample, infinite surnames) and skewed distribution.

Objective

- We want to see whether it is true that the informative content of surnames is larger the larger the degree of inheritance.



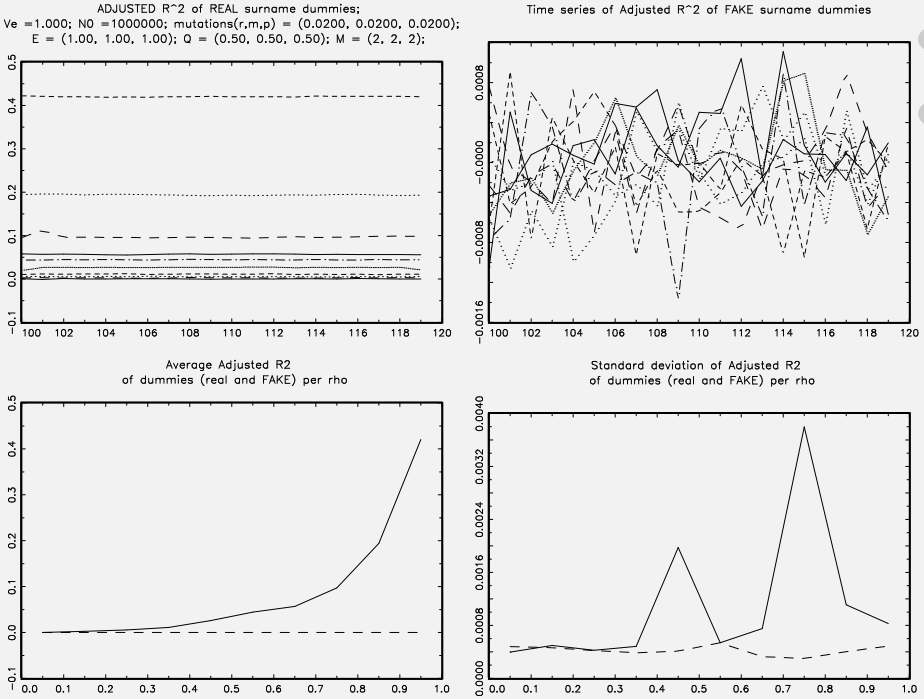
- $R^2_{surnames}$ is larger for larger ρ.
- R^2_{fake} is not larger for larger ρ

Procedure

- Assume exogenous parameters (ρ, V_u, Q, M, μ)
- Start with uninformative surnames (random). Both, skewed and uniform initial surname distribution.
- Converge to a joint distribution of surnames and income. Skewed.
- Measure Informative Content of Surnames.
- Do it many times
- Do it for many values of ρ

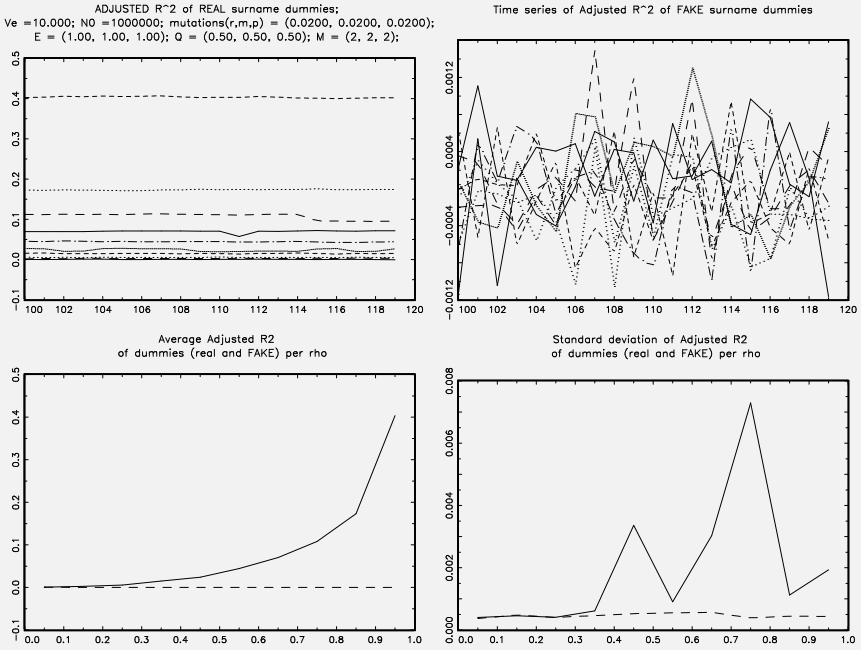
Results

- Surnames are informative, and their informational content increases with the degree of inheritance that there is in society.
  - $Q = \frac{1}{2}$, $M = 2$, $\mu = 0.02\%$, $V_e = 1$
  - Irrespectively of
    - **Conditional variance**.
    - **mutation rate**.
    - **or family size**

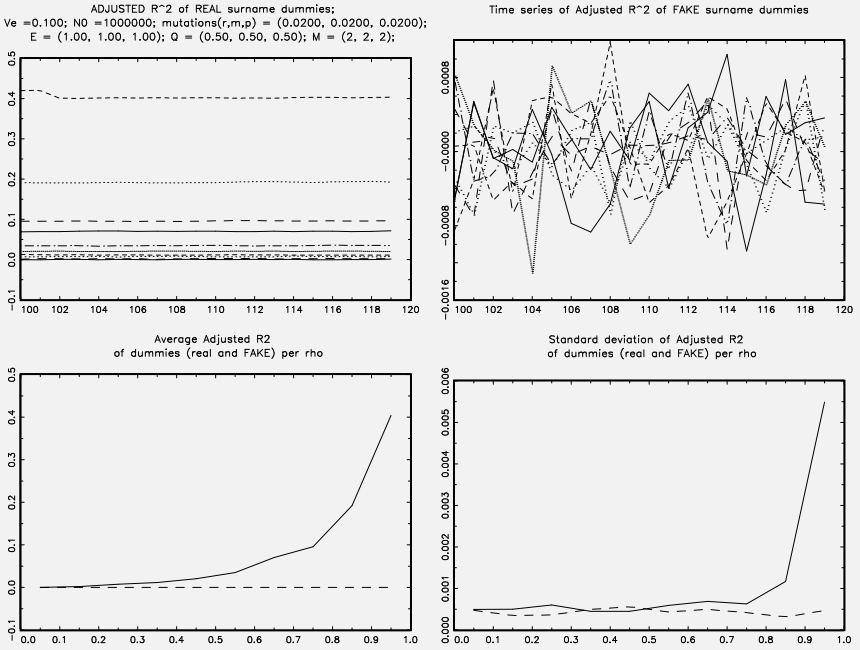


Conditional variance

- $V_e = 10$

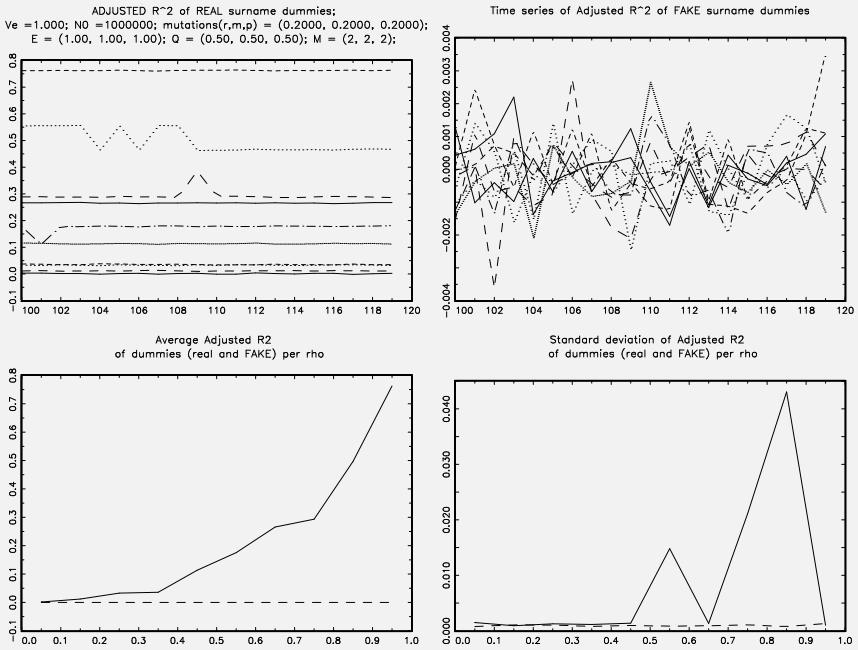


- $V_e = 0.1$

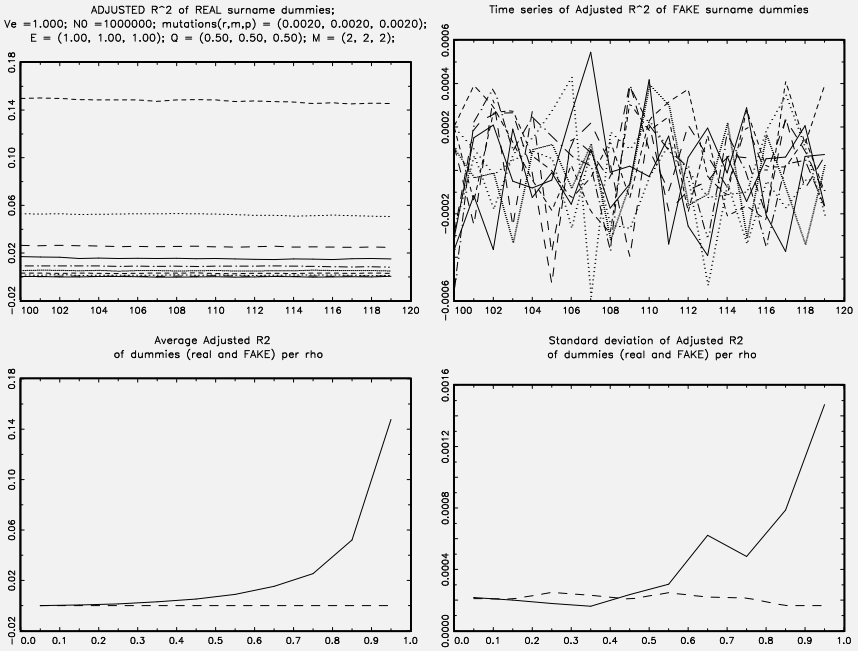


Mutation rate

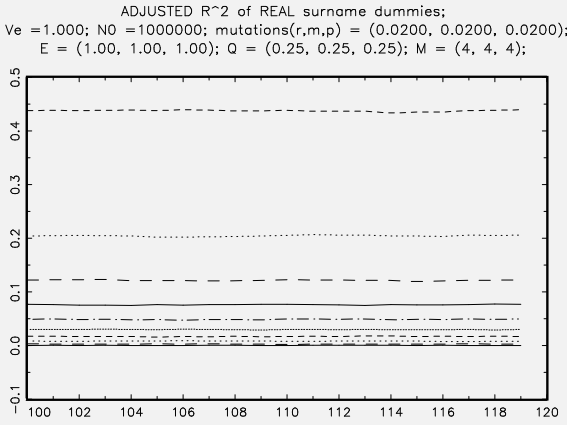
- $\mu = 0.2$

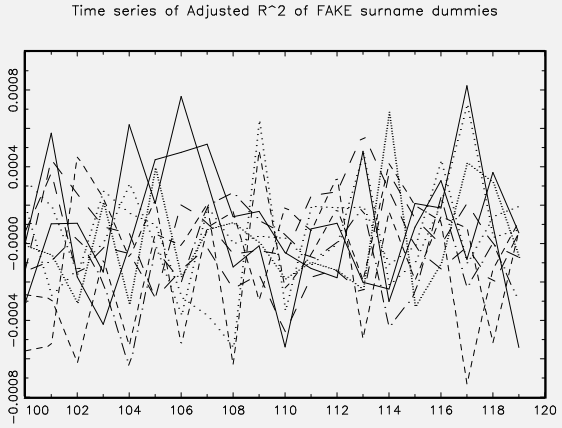


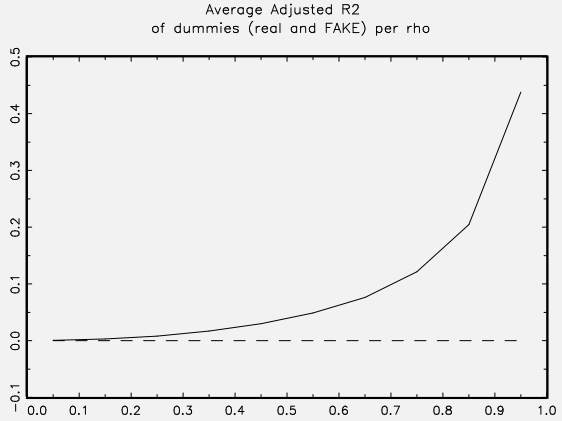
- $\mu = 0.002$

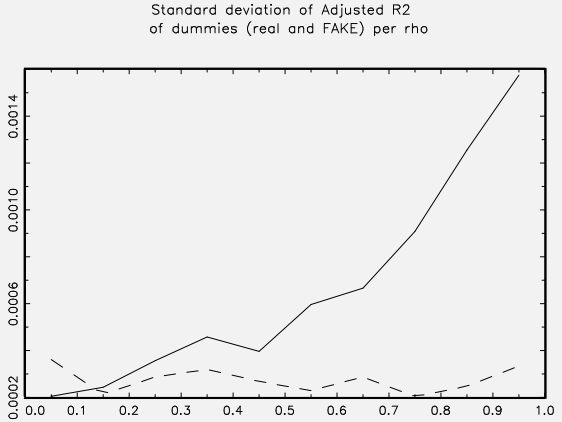


High Variance of Family Size









Extensions of the model


- Assortative Mating in Income or Education
- Ethnicity
- Inmigrants
- Frequency of Surnames



Assortative Mating in Income or Education

- Extension of the model. Inheritance from both father and mother.
- Assortative Mating: rich people gets to marry rich people.
- The more AM...
  - the more info the income of the husband has on the income of the wife.
  - the more info the income of the husband has on the inheritance than the wife passes
  - the more info the income of the husband has on the income of the kids
- More AM is identical to an increase in ρ in our model.

Ethnicity

- Imagine that belonging to a certain ethnic group may change the income process:

  $y_t = \alpha_{eth} + \rho y_{t-1}$

- If the surname contain information on the ethnicity, the surname contain info on the income.

- Surname contain info on ethnicity only in the measure
  - that ethnic characteristics are inherited through the male line
  - that there exists assortative mating according to ethnicity (judaism)

Inmigrants

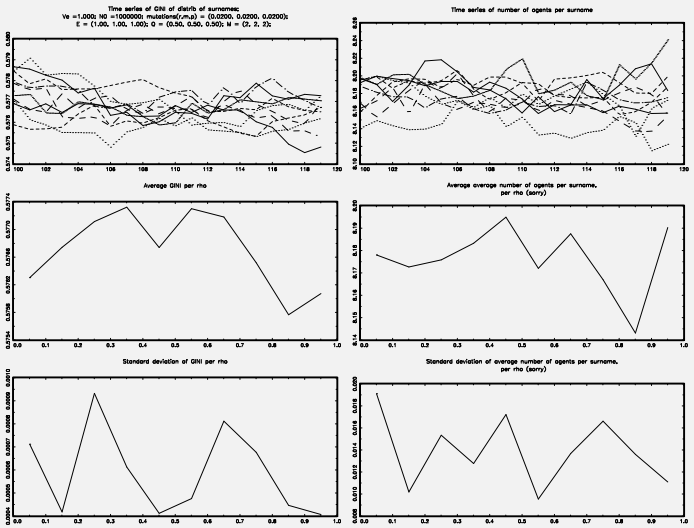
- In the measure that immigrants have different surnames that population.

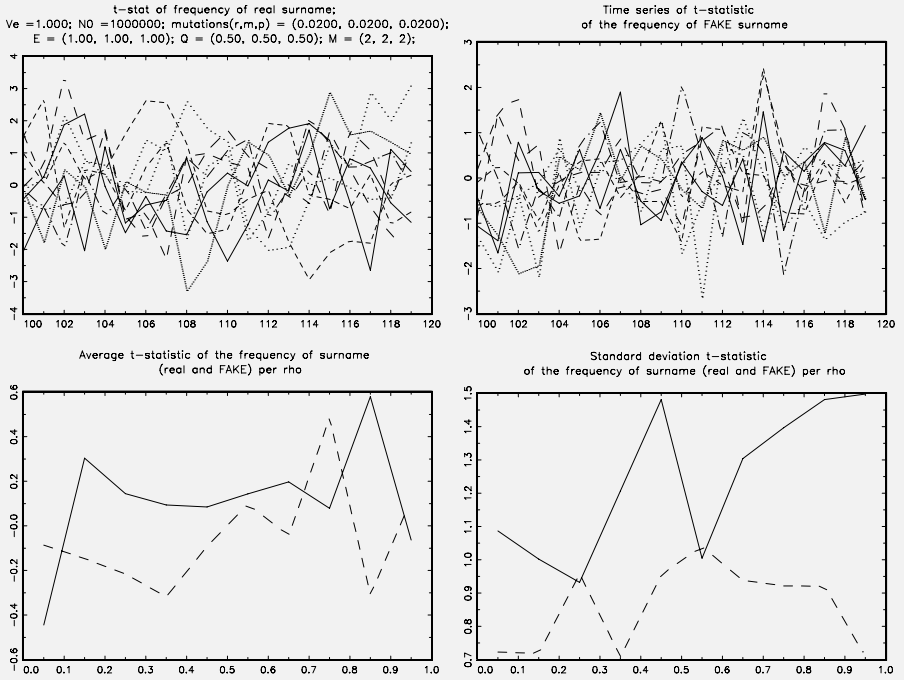
- If by the surname we can characterize the origin of the migrant, we can determine the speed at which migrants integrate in the recipient society.

Frequency of Surnames

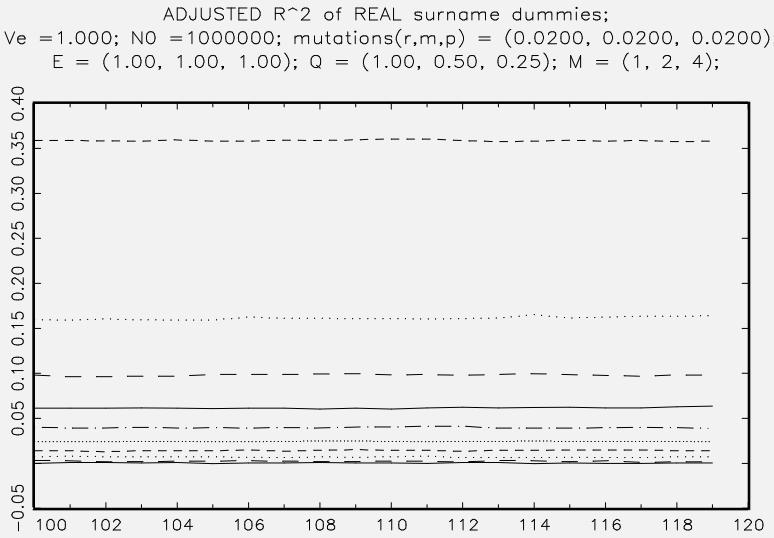
- In the previous model the specific surnames had information,
  - **but their frequency did not**.
- Frequency has information if probability of death/birth of lineages differs across income groups.
- **"Hereu" effect**: The rich want a male, $Q_{rich} = 1$, but $E = 1$ for all.
- **Average Fertility Differences**: The rich (male) have more children in average.
- Differences in mutation rates. "Gentryfication" versus immigration.
- **Summary:**
  - Even if frequency matters it is much better to use ICS to study mobility. ICS always works, frequency is impossible to interpret.
  - This has nothing to do with the fact that less frequent surnames are more informative

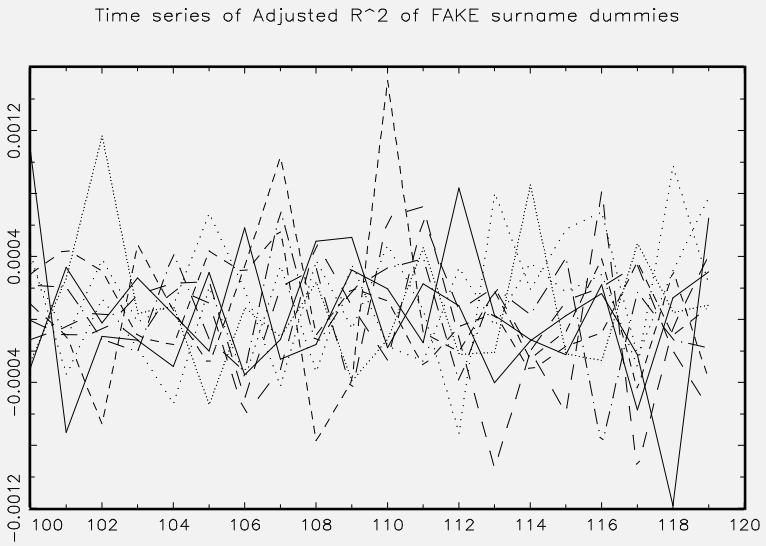
Frequency in baseline model

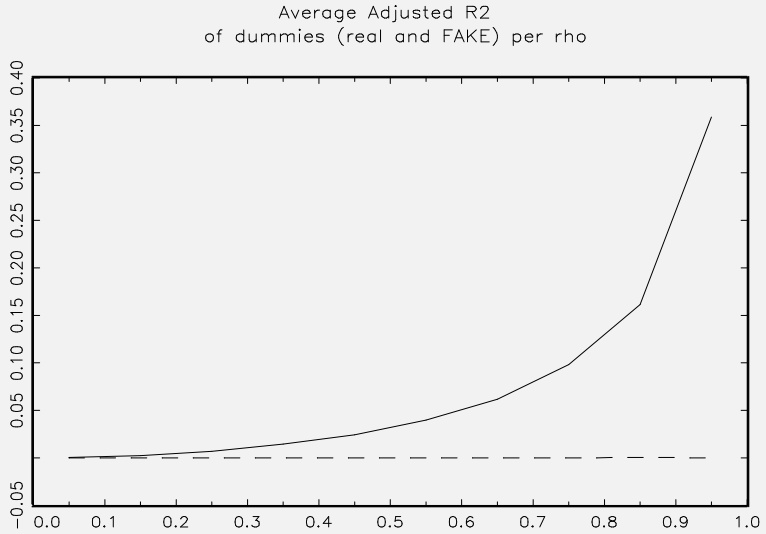


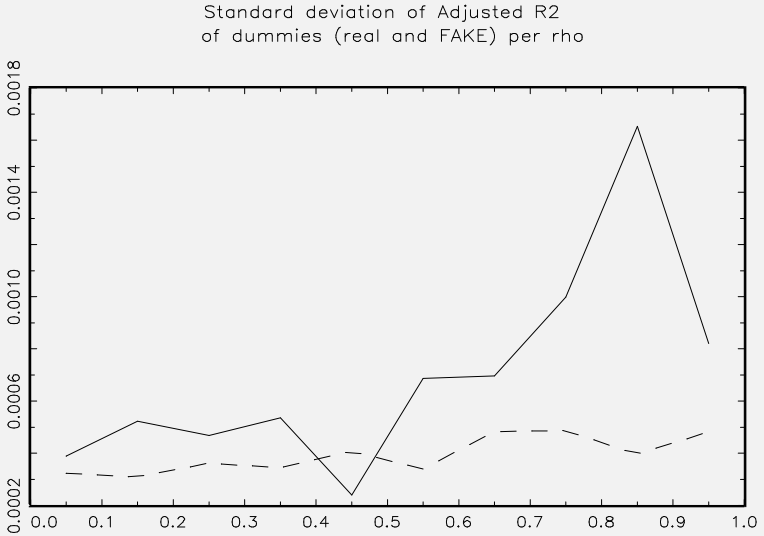


Hereu effect *(1/3)*

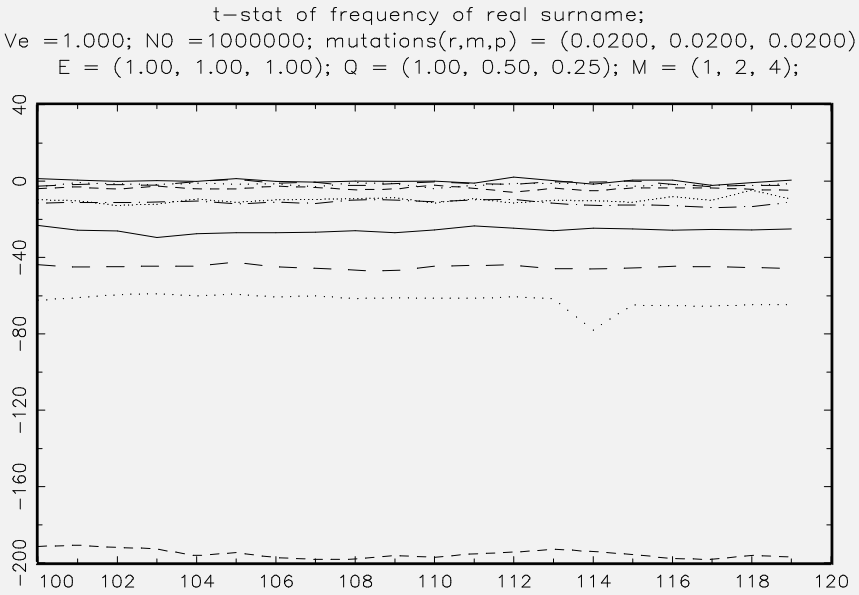


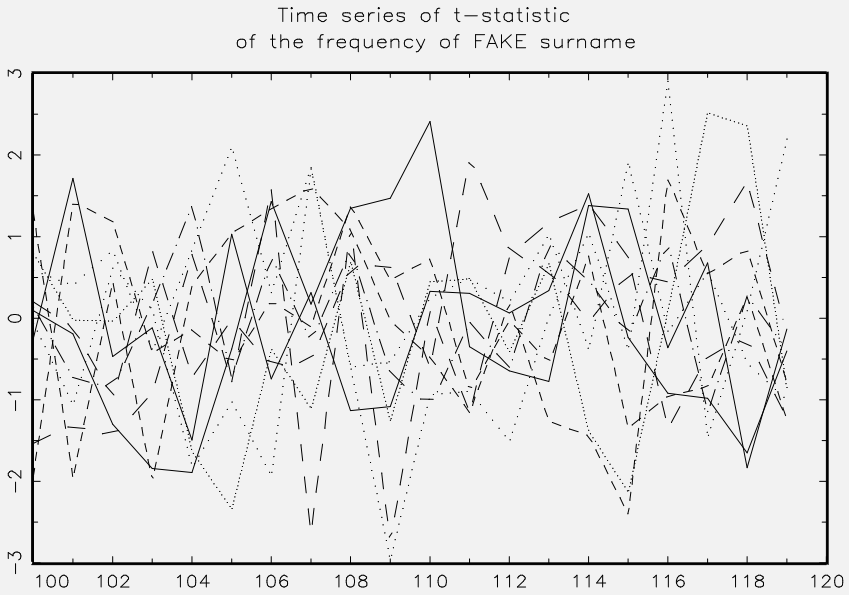


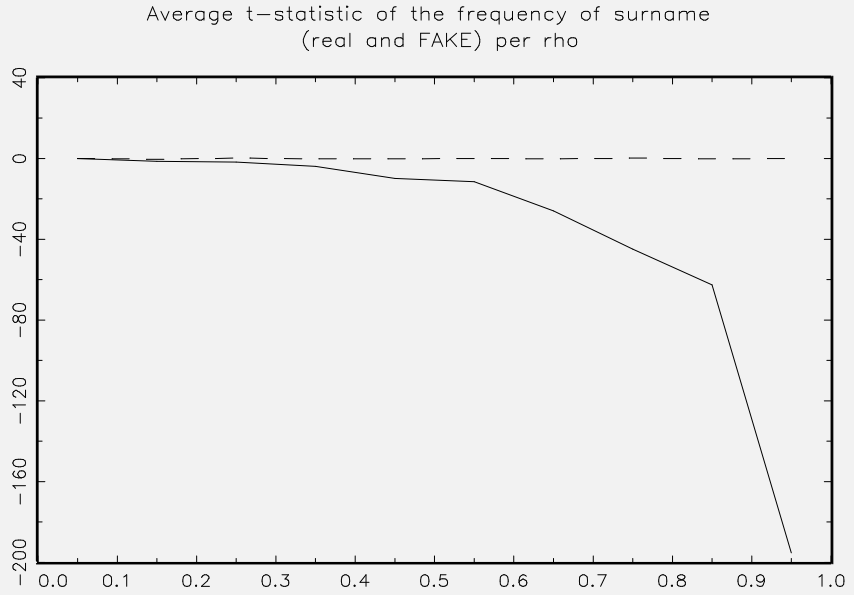


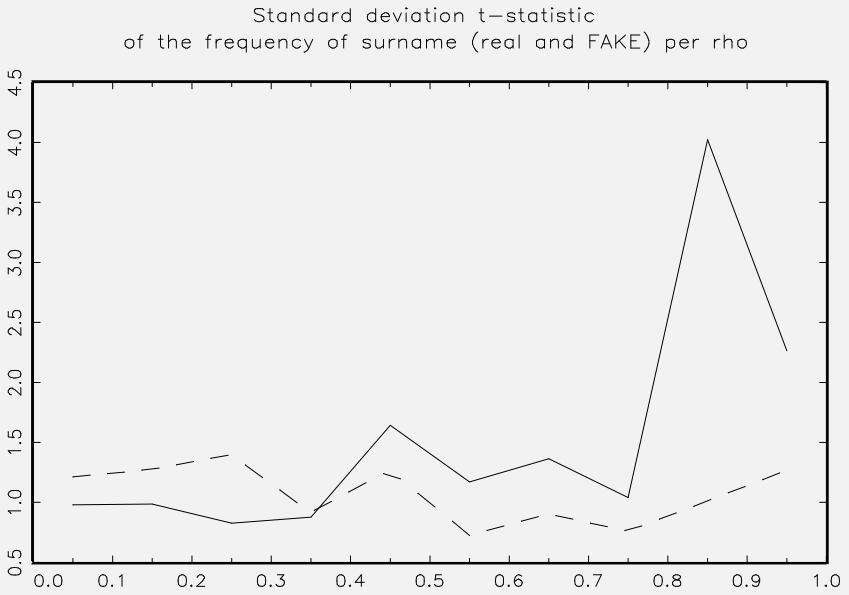


Hereu effect *(2/3)*

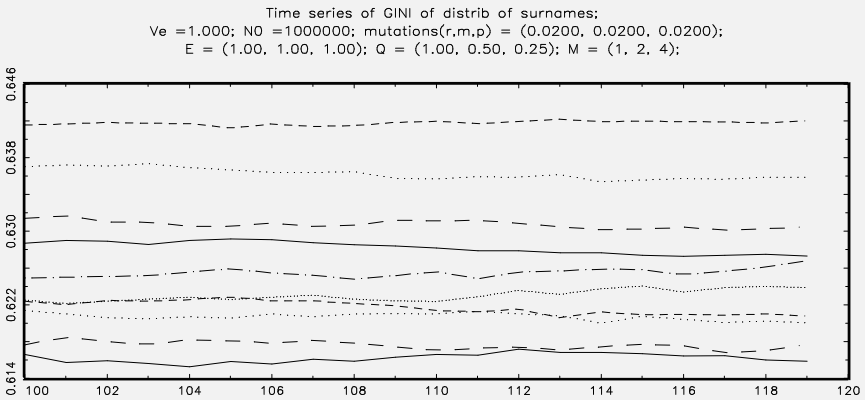


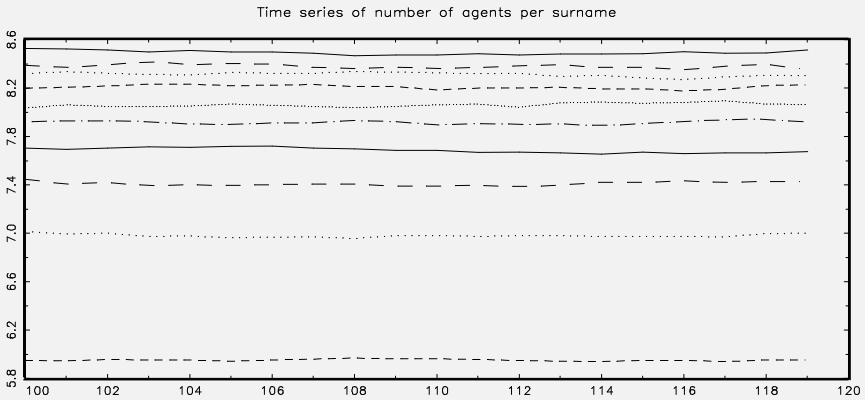


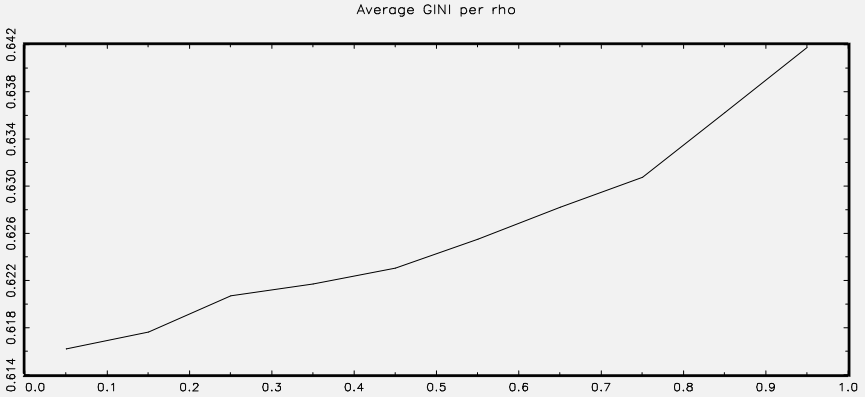


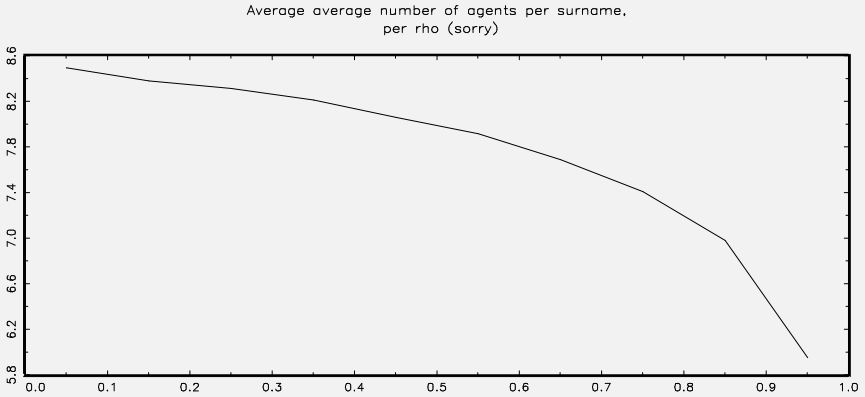


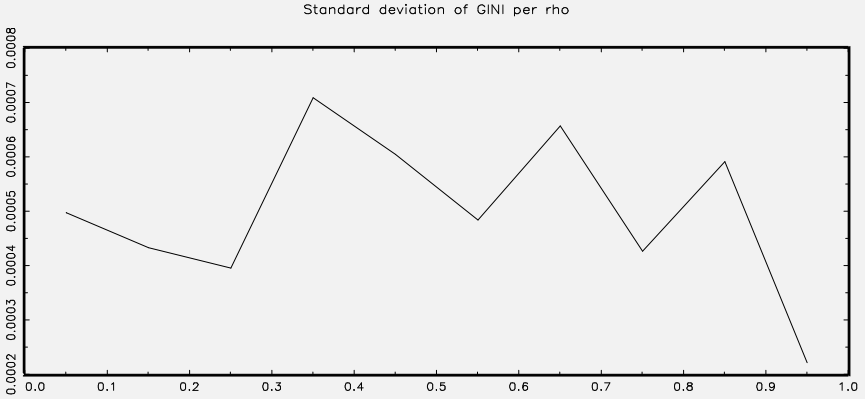
Hereu effect *(3/3)*

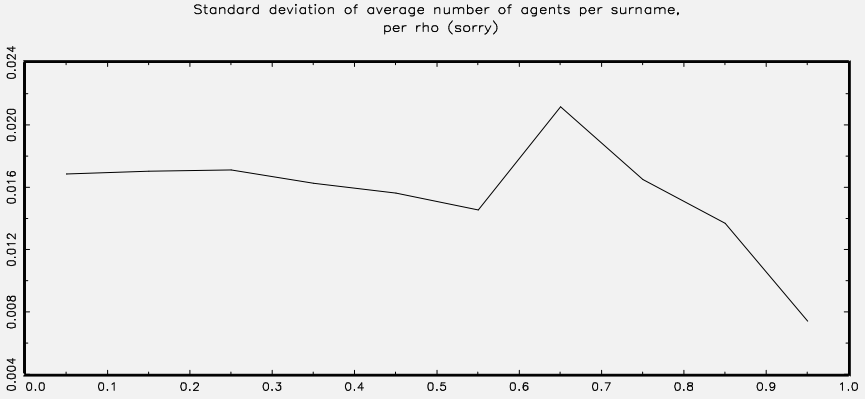




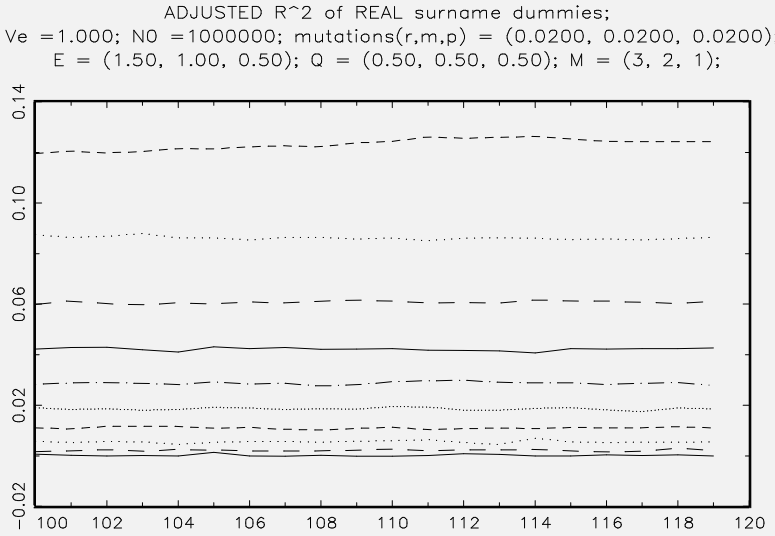


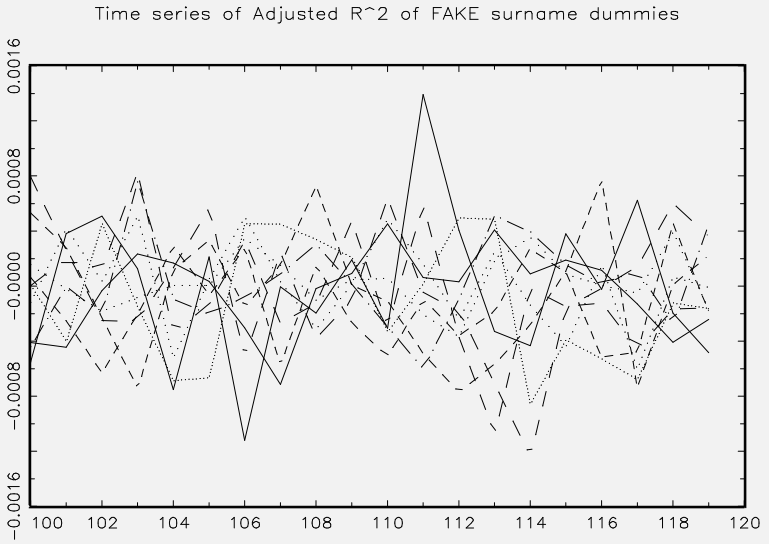


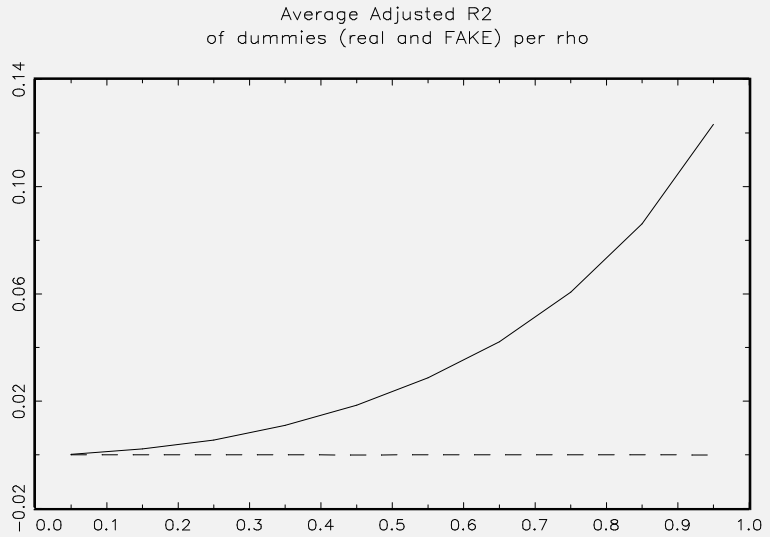


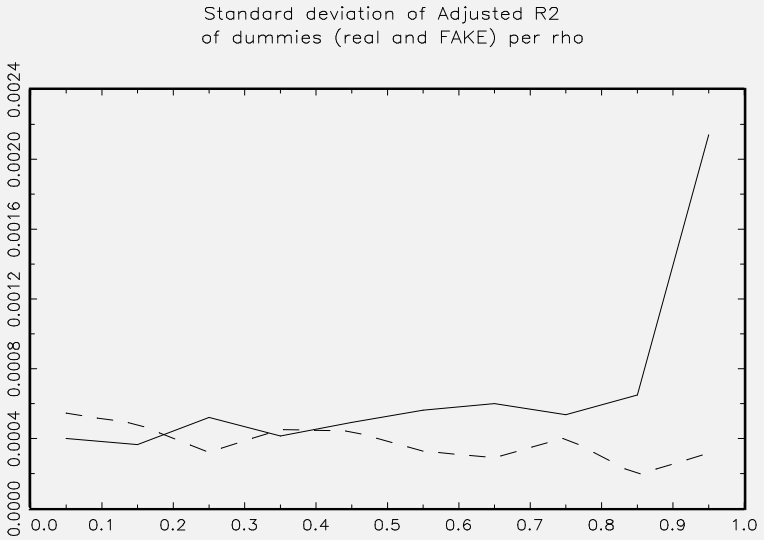


Average Fertility Differences *(1/3)*

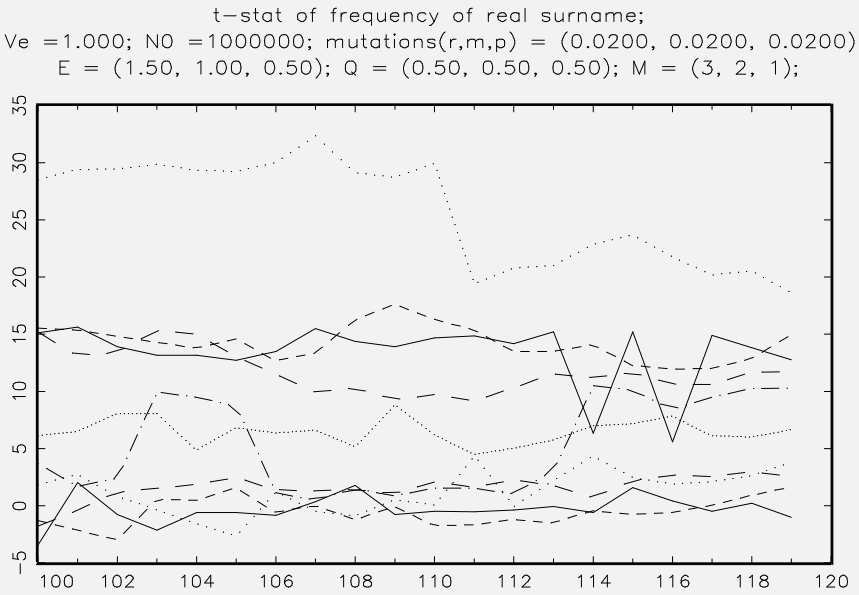


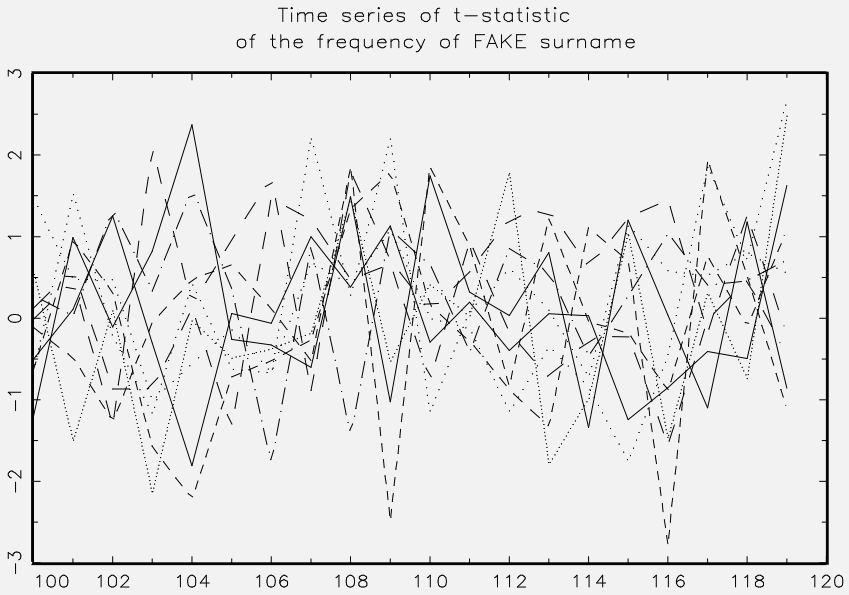


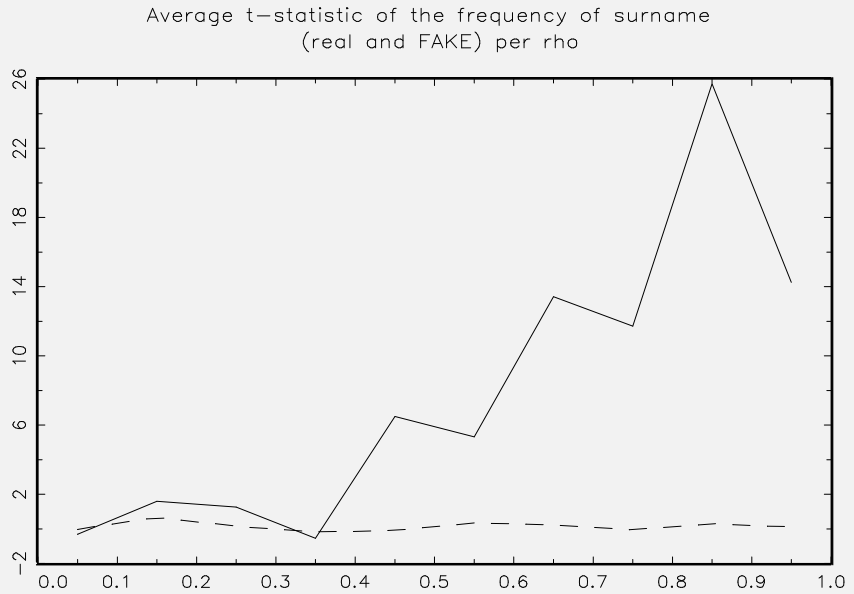


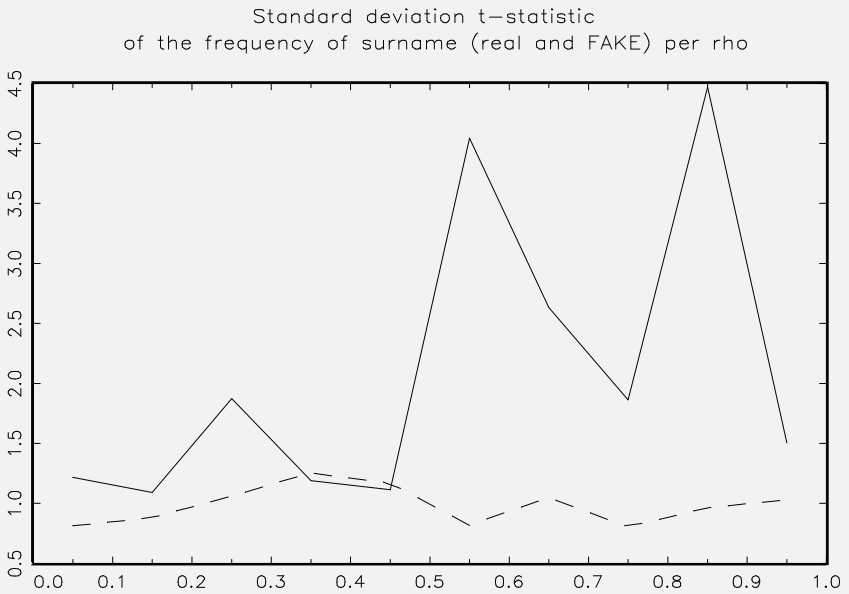


Average Fertility Differences *(2/3)*

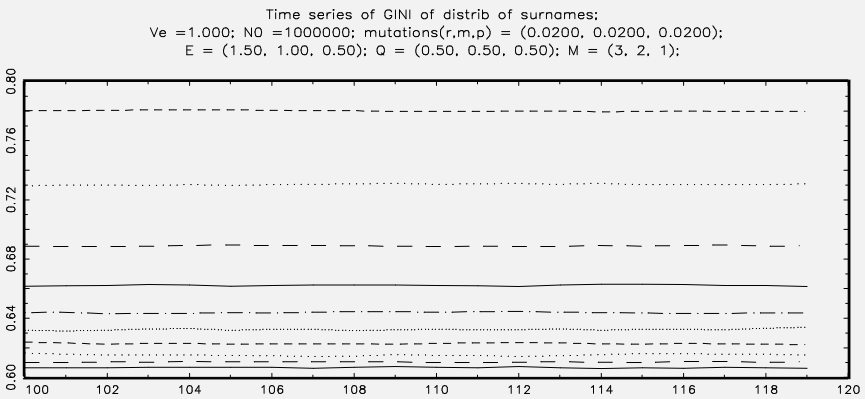


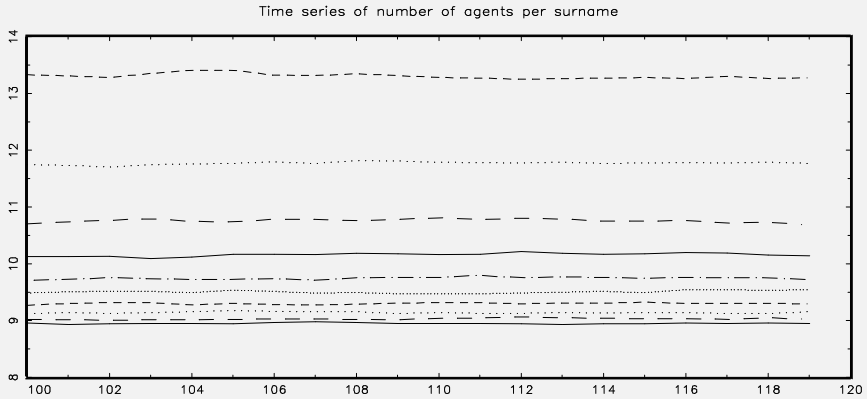


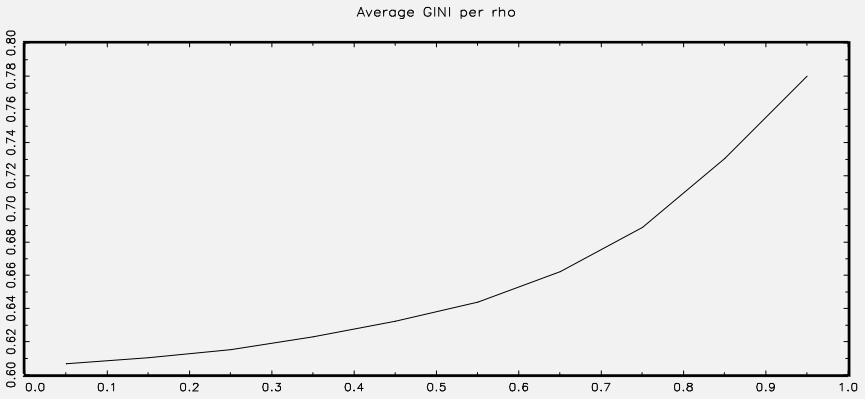


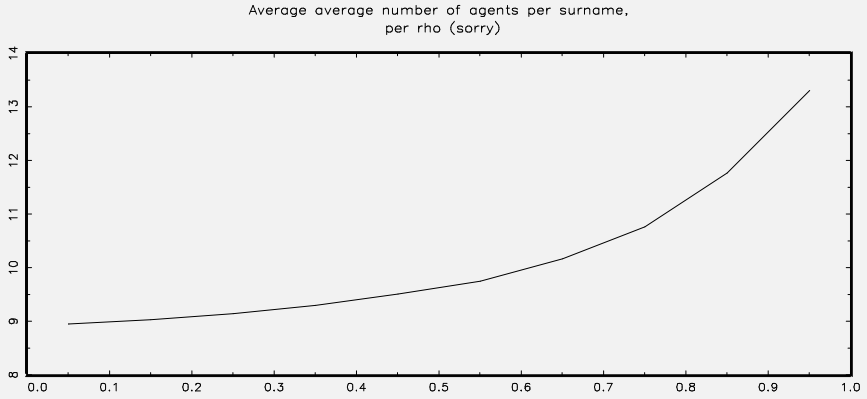


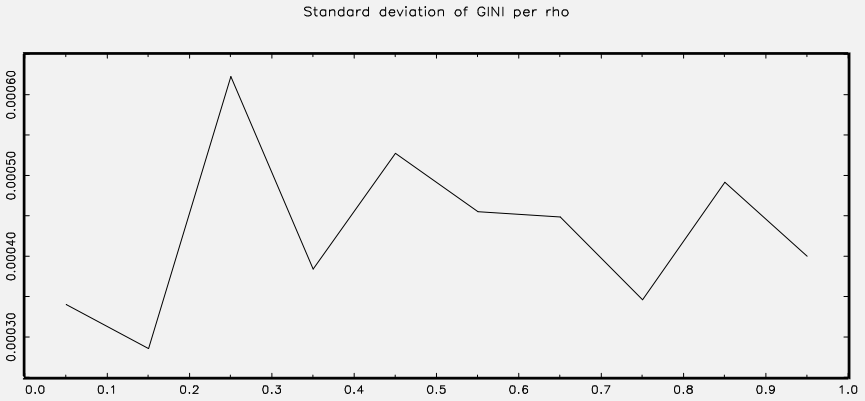
Average Fertility Differences *(3/3)*

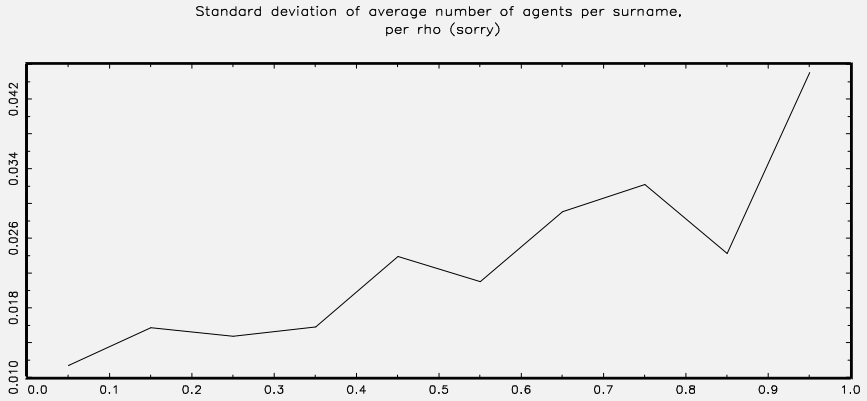












Summary. Why Surnames Inform


- Surnames are a partition of the population that informs on the family relations of the individual.
  - Birth/Death process of surname creation generates skewed distributions.
    - For many individuals, sharing surname means close family relation.
    - The less mobility, the more family explains, the more info surname contains.
    - More unfrequent surnames should have more information.
  - Assortative Mating is isomorphic to an increase of inheritance.
    - Increase the amount of information that the *father* has on the son... and surnames are transmitted through the male line.
- Surnames are a partition that informs on ethnicity of the individual.
  - If ethnicity matters, surnames do matter insofar ethnicity characteristics are inherited through the male line.
  - Surnames inform on the speed of assimilation of emigrants.
    - The slower the path of integration (less mobility), the more they matter.
- Differences in rates of birth/death of lineages according to income
  - Frequency should also have info.
  - It is difficult to learn anything because the relationship between the value of frequency and mobility might go in either way.
  - The relationship between the ICS and mobility does not depend on this.



Spanish Surnames


- Spain is an ideal case for surnames studies: Surnames have the same features than in the rest of the world, **PLUS** some extra ones that allow us to do many robustness checks of our method.
  - It is hard to change surnames (much more than, say, the US). Plus stable, simple and generalized orthographic rules. Few "mutations" due to spelling differentials. **"Mutations" are recorded.**
  - Males and Females do not change surnames along their lives, regardless of marital status.
  - Spaniards have **two surnames**, inherited from father and mother, and we can measure assortative mating, which is part of the story.
  - We can locate siblings with quite a lot of accuracy, by the combination of two surnames.
  - Before 2001 almost all the **immigration is between regions** of Spain and we can control for this since we know the distribution of surnames in the country. Otherwise we would need data from other countries.
- We use this wealth of information (not available anywhere else) to show that **our methodology works when using only the first surname**.



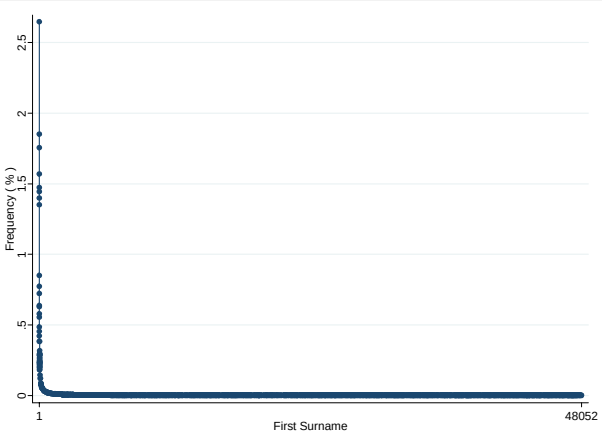
Data

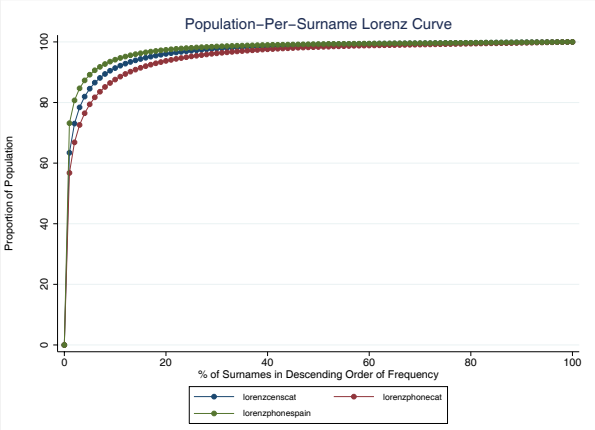
- The 2001 Catalan Census
- Distribution of Surnames
- Analyzed Population & Ethnicity Controls

The 2001 Catalan Census

- Census of the whole population, 6343110
- Variables:
  - 2 Surnames
  - Demographic characteristics (including birthplace)
  - Household characteristics
- Does not have income, but has relevant economic outcomes.
  - Education
  - House Inheritance
  - Availability of second residence.

Distribution of Surnames





- Very skewed. Gini .9

- Some interesting figures:
  - 61K different surnames.
  - On average, 34 people per surname.
  - Yearly 96 surnames die.
  - In Spain in 2001, 1.5K individuals had their surname approved for modification.
    - This amounts to a mutation rate of .26%. 107 new surnames.
  - Roughly steady state.

Analyzed Population; Ethnicity *(1/3)*

- Aged 25 and above (full time education finished)

- Focus on information in surnames due to family links:
  - Spanish citizens living in Catalonia (no foreign immigrants)
  - "Ethnicity" component of surnames (regional origin within Spain),

$$CatalanDegreeSurname(j) = \frac{\textit{Number of Phones under surname j in Catalonia}}{\textit{Number of Phones under surname j in the Spain}}$$

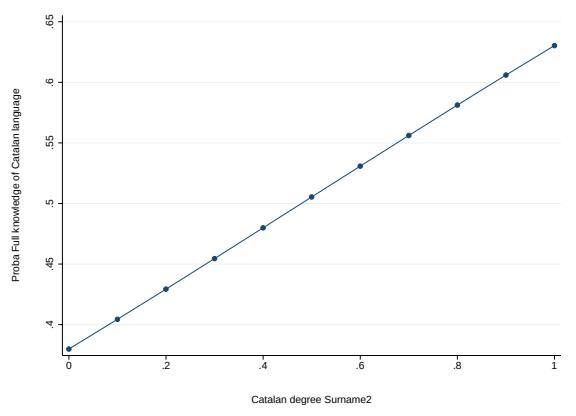
    - Captures the probability that a Spaniard holding surname j is living in Catalonia.
    - We use it as a proxy of the "catalonianess" of the holder.

Analyzed Population; Ethnicity *(2/3)*

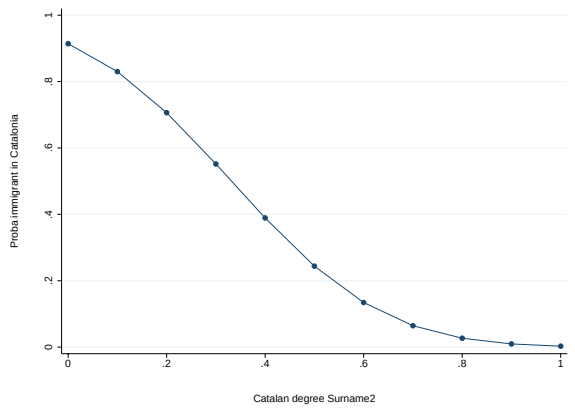
Summary Statistics

	All residents in Catalonia	Born in Catalonia before 1950	Born anywhere in Spain before 1950	Born anywhere in Spain after 1950
Mean CatalanDegreeSurname2	0.344	0.5672	0.367	0.322
Standard deviation	(0.302)	(0.3241)	(0.312)	(0.292)
Share with CatalanDegreeSurname2>0.16	0.568	0.8365	0.596	0.542

Probability of Knowledge of Catalan Language



Probability of Immigrant



Analyzed Population; Ethnicity *(3/3)*

Probability of knowledge of Catalan language

LHS: Knowledge of Catalan language	(1)	(2)
CatalanDegreeSurname2		0.639 (0.003)
Log likelihood	-2248320.8	-2219774.9
Pseudo R^2	0.2387	0.2483

Probability of being an immigrant

LHS: Immigrant	(1)	(2)
CatalanDegreeSurname2		-4.121 (0.006)
Log likelihood	-1418949.8	-903746.31
Pseudo R^2	0.0052	0.3664

Empirical Results

- The informational content of surnames is **large**, beyond ethnicity.
  - Also among very Catalan surnames; those born in Catalonia; and Catalan born before 1950.
- Calibrating the model, the ρ that matches the ICS for those born in Catalonia (3.19%) is 0.4. Rough, but reasonable number.
  - Also, the mutation rate implied is "about right" (0.2%)
  - It increases trust that the exercise is not silly.
- As the model predicts,
  - the ICS is **larger for surnames that have less individuals**, as the surname partition is then more related to family.
  - the **frequency** of surnames is also informative.
- An alternative way to look at family links: using *both* surnames (in practice grouping **Siblings**).
- Increase in provision of public education results in **more education**
- In spite of it:
  - Also **ICS has increased over time** (after controlling for ethnicity); Family Background and Ethnicity are BOTH more determinant (**table**)
  - This is very robust: **among least frequent, very Catalan surnames, siblings.**
- Explanation: The amount of **Assortative Mating** did increase one generation earlier (**table**).

ICS. Complete Population.

Born in Spain. Living in Catalonia.
More than one individual with surname

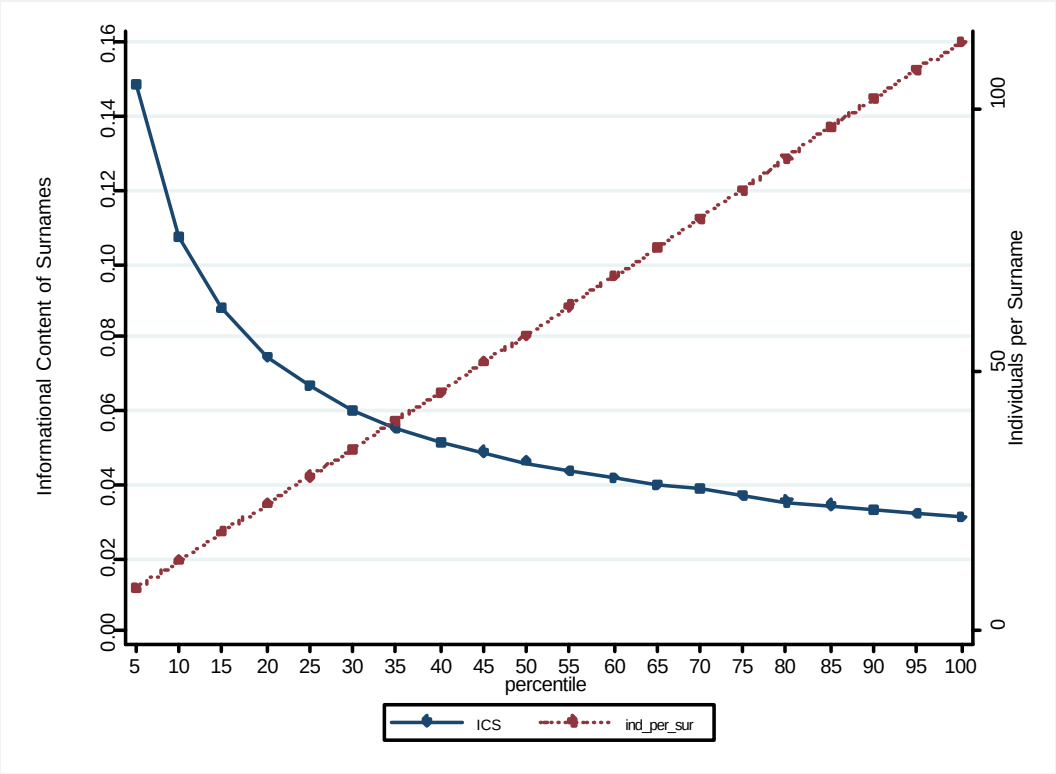
LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.692_{(0.007)}$	$1.017_{(0.008)}$	$1.692_{(0.007)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.3363	0.3440	0.3653	0.3440	0.3629	0.3363
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.384	0.000	0.332

Notes: Individual controls include gender, age dummies & county of birth.

Number of surnames is 38,024.

ICS= 2.13%.

Large ICS for unfrequent surnames



Frequency

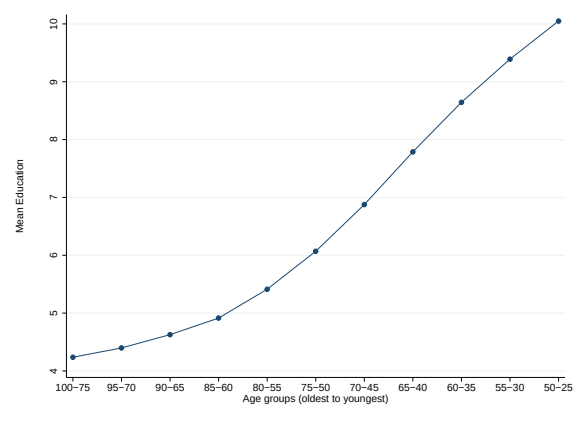
LHS: years of education	(1)	(2)	(3)	(4)
FrequencySurname1	-30.157$_{(0.309)}$	-23.696$_{(0.309)}$		
FrequencyFakeSurname1			0.148$_{(0.301)}$	0.107$_{(0.299)}$
CatalanDegreeSurname2		1.636$_{(0.007)}$		1.692$_{(0.007)}$
R-squared	0.3378	0.3449	0.3363	0.3440

Siblings

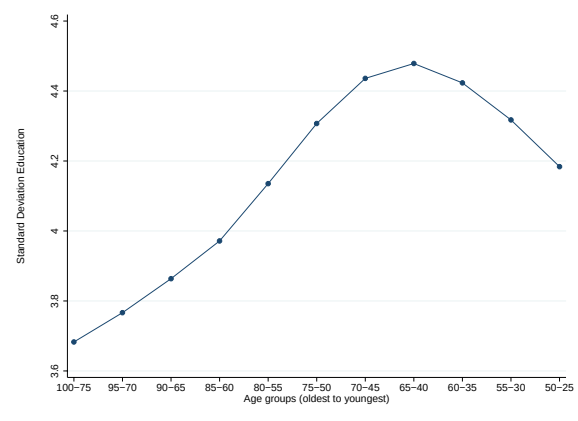
Dependent variable: years of education	(1)	(2)	(3)	(4)	(5)	(6)
Adjusted R^2, Surname Dummies	0.5486	0.5416	0.5375	0.5326	0.5283	0.4696
Adjusted R^2,Fake Surnames Dummies	0.3244	0.3224	0.3218	0.3228	0.3232	0.3354
Informational Content of Surnames	0.2242	0.2192	0.2157	0.2098	0.2051	0.1342
Observations	774,788	1,315,853	1,664,717	1,900,652	2,067,590	3,695,479
Number of surnames	387,394	567,749	654,965	702,152	729,975	811,502
Max number of people per surname	2	3	4	5	6	All sample

Evolution of Mean and S.D. of Education

Mean



Standard Deviation



ICS increases with time.

Born before 1950 (ICS= 1.98%)

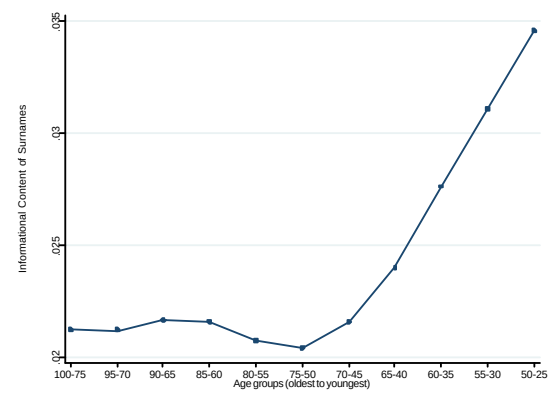
LHS: years edu. all	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.896	0.594	0.897		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.2313	0.2335	0.2533	0.2335	0.2524	0.2313
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.679	0.000	0.688

Born after 1950 (ICS= 3.5%)

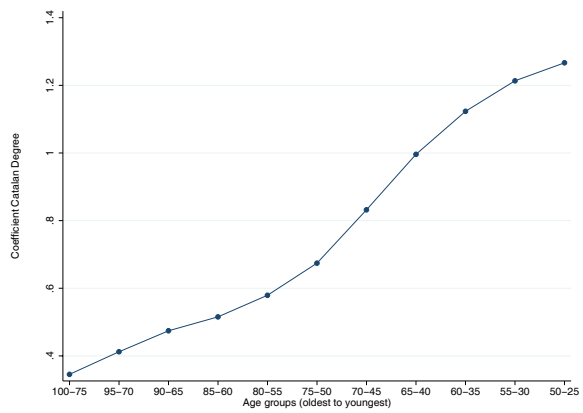
CatalanDegreeSurname2		2.143	1.271	2.145		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.1002	0.1183	0.1534	0.1184	0.1481	0.1002
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.260	0.000	0.421

More determinant Background, Ethnicity

Evolution of ICS over time

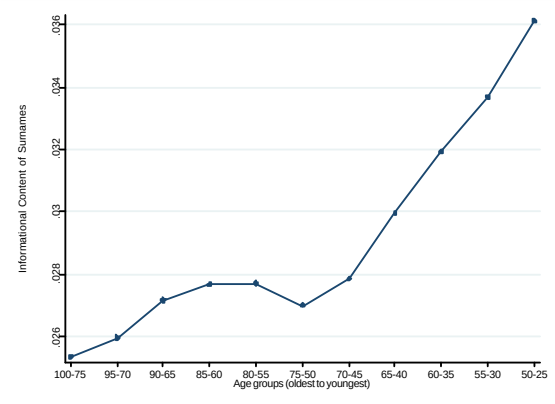


Evolution of parameter of $Catdegree$

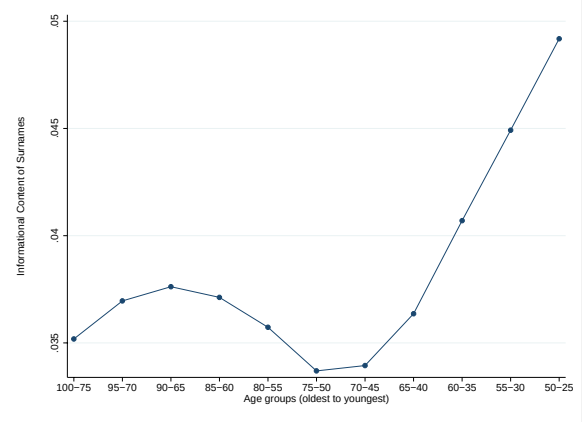


Increase of ICS *(1/2)*

45% Most Catalan Surnames

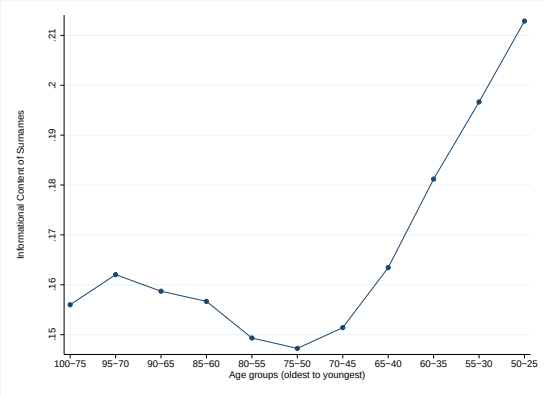


50% Least Frequent Surnames

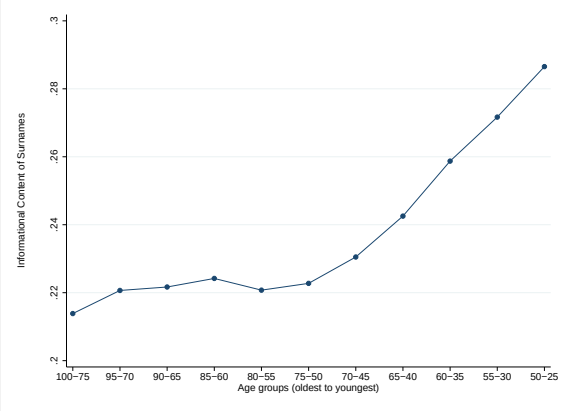


Increase of ICS *(2/2)*

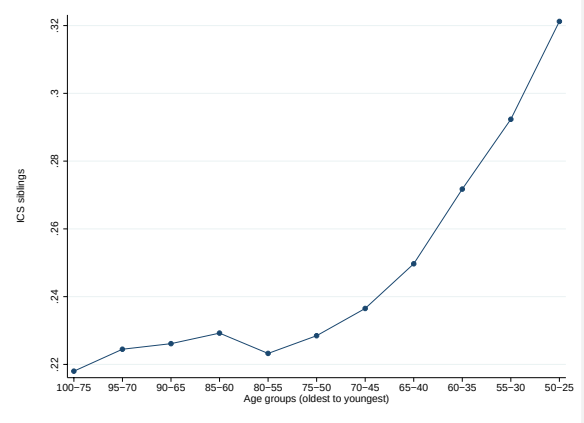
ICS of the complete surname (2 surnames). Siblings



Siblings. 45% Most Catalan Surnames



Siblings. 50% Least Frequent Surnames



Assortative mating table

- Education

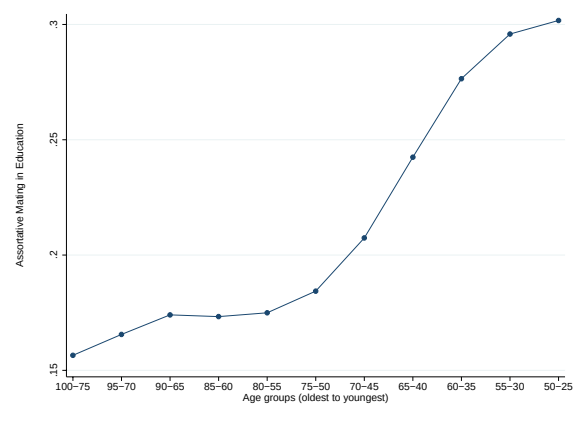
	EduSurname2	
	"Old"	"Young"
EduSurname1	$0.170_{(0.001)}$	$0.303_{(0.001)}$
Observations	2,041,044	2,222,917
R^2	0.3410	0.1997

- Catalan degree

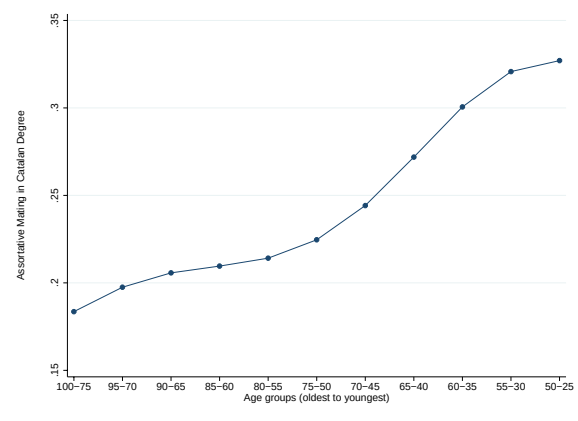
	CatDegreeSurname2	
	"Old"	"Young"
CatDegreeSurname1	$0.217_{(0.001)}$	$0.328_{(0.001)}$
Observations	2,041,044	2,222,917
R^2	0.5110	0.2778

Assortative mating graph

A.M. by Education



A.M. by ethnicity



Conclusions

- ICS, with only one surname, allows us to look at intergenerational mobility; in particular to its evolution over time.
  - Surname distributions are **necessarily skewed**. So, we can capture family links.

- In Catalonia, we find that mobility has decreased as a consequence of an increase in assortative mating. Still, our preferred reading is that the method is workable.
  - With only first surname we find increase of ICS. This is what we would have found with, say, UK data.
  - With siblings we find the same result but using a **different methodology**. We could NOT do this in the UK, but the same results with both methods.
  - We can explain why this happen. Using yet another methodology, AM... consistent results.

Ongoing Extensions

- We intent to make comparisons **across countries**. This needs of:
  - Improved Calibration of more general model.
  - **OR** determining an statistic which directly comparable across countries (using less frequent surnames)
- **Using Surnames in Censuses:** It is possible to convince census authorities.
  - Coding of surnames.
  - Skewedness helps confidentiality.
- Without Census, we still could do the reverse exercise.
- Other inheritance: health, immigrant assimilation.

Prepared with SEVISLIDES

Intergenerational Mobility and the Informative Content of Surnames

Maia Güell (University of Edinburgh)
José V. Rodríguez Mora (University of Edinburgh)
Chris Telmer (CMU)

SIRE Conference, 19/11/2007