Underrepresentation of Minorities and Women in Economics: PhD field specializations and their determinants

Karan Singhal (University of Luxembourg and LISER)
Eva Sierminska (LISER, DIW, IZA, GLO)

29th September 2023, Gdansk

Summary

- **Motivation:** While there's an increasing awareness of gender disparities in economics academia, the influence of country of origin/ethnicity, despite being recognized as crucial, remains underexplored

- My research primarily sheds light on disparities concerning field specializations among beginning economists, specifically PhD graduates from the US

- **Question:** How do the field choices of PhD graduates differ across regions? What role does gender play in field selection across these regions?

- Field specialization shows notable variation across regions, with some overarching similarities but marked concentrations in certain areas.

- Gender variations are evident across regions: field choices for women (compared to men) are considerably more concentrated in specific areas

- Prior research hasn't delved into the relationship between region/ethnicity and gender. An intersectional perspective, helps understand gender disparities better

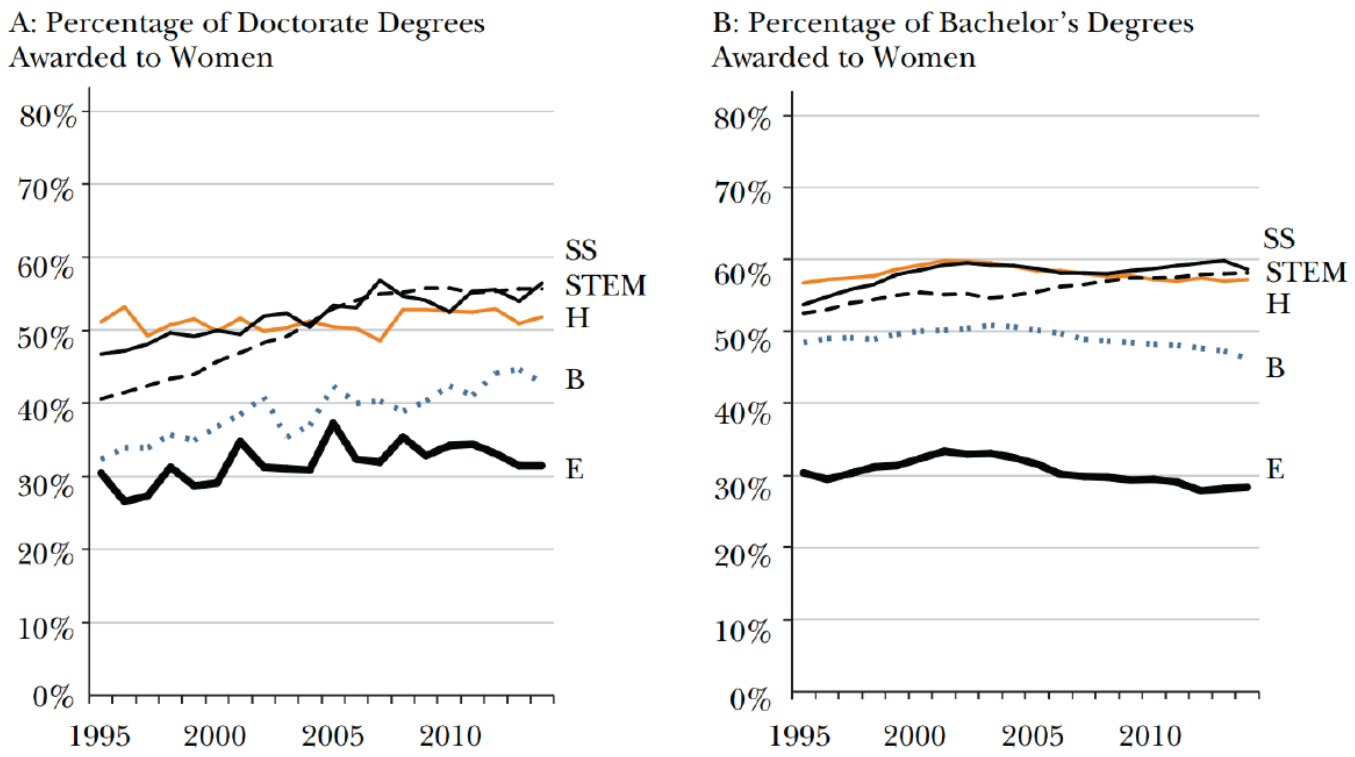
Outline

- Background: Inequality in Economics Academia & Field Specialization
- Current gaps and question
- Data & Methodology
- Understanding Specializations: JEL & Topic Modeling (quick detour)
- Tracing Country of Origin: Mapping Names to Ethnicity/Origin
- Findings: Field Specialization by Country of Origin
- Discussion/ Conclusion

- Genetic Distance (detour no.2)

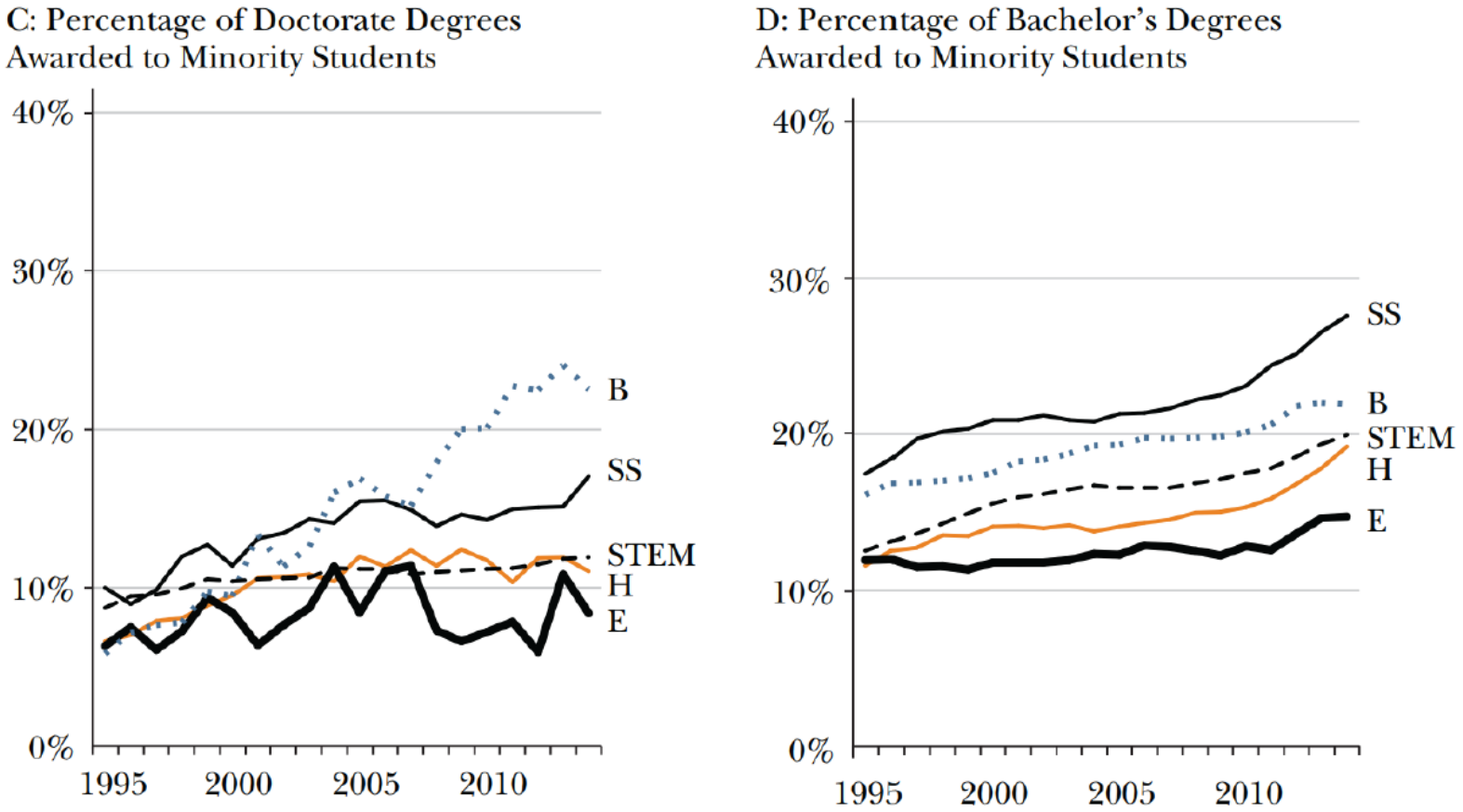
Representation in economics remains stagnant: gender

Under-representation of US women



Somanchi (2022)

Representation in economics remains stagnant: minorities



Somanchi (2022)

Econ academia is more ‘leaky’ and stagnant:

- **Stagnant share of women faculty and PhDs in economics** compared to other social sciences and low representation of minorities (Auriol et al. 2022; Lundberg and Stearns 2019; Bayer et al. 2020)

- Economics PhDs remains the least socioeconomically diverse (Schultz and Stansbury 2022)

So what?

- Economics heavily influences policy (Fourcade et al. 2015) and diversity fosters academic excellence, innovative research, and policy outcomes (May et al. 2014; Mester 2019)

- Women advocate for more policy-relevant, and interdisciplinary research (Andre and Galk 2021; Bayer and Rouse 2016)

- Diverse collaborations yield more impactful papers (Freeman and Huang 2014)

- Efforts to increase representation align with global and institutional objectives, especially in STEM and Economics

Econ academia and beyond: barriers, biases, and 'preferences'

Research, Seminars, Publications

- E.g.- Women and miniorities face **– tougher editorial standards**, lower chances of being invited to seminars, more **patronizing questions**, (Doleac et al. 2021; Dupas et al. 2021, Hengel 2022, Card et al. 2020; Koffi 2021)

Tenure, promotion, pay gaps

- E.g.- **less credit for co-authorship** (Sarsons et al. 2021), females and minorities **earn 10-15% less** than male counterparts (Foster et al. 2022); **poorer evaluation** rates in grant proposals (Witterman et al. 2019)
- **Promotion gap is the largest** across disciplines for women (Ceci et al. 2014)

Mentors, peers, role models and path dependence

- E.g.- **Fewer role models** (Bettinger and Long 2005; Porter 2020) **and path dependence** in Econ (Hale and Regev 2014), less likely to praised for their ability in **reference letters** (Eberhardt et al. 2022)

These disparities extend to sub-fields within Econ

- Uneven distribution based on gender has been noted based on published work, dissertations, conferences and seminars (e.g.- Chari and Goldsmith-Pinkham 2018; Hospido and Sanz, 2020, Fortin et al. 2021; Sierminska and Oaxaca 2022)
    - Male authors are over-represented in micro and macroeconomics, **female authors are over-represented in labor and development economics** (Onder and Yilnazkuday, 2020)
    - Women are **underrepresented in Finance and Macroeconomics** and overrepresented in Microeconomics (Chari & Pinkham, 2017; Beneito et al., 2021)
    - On the contrary, once social, economic and institutional aspects are factored in women are less likely to specialize even in labor and health (Sierminska and Oaxaca 2021, 2022)
- These differences in field specializations affect differences in placement outcomes (Fortin et al. 2021)
- Research in econ sub-fields is limited because it's difficult to conceptualize; field choices have been examined more at the undergrad level (Altonji et al. 2016)

Why might country of origin/ ethnicity matter specifically?

- **Not about us**
  - Economics produces far less race-related research that other social sciences (Advani et al. 2020, 2021)
  - Existing models and work discounts for racial discrimination (Gaule and Piacentini 2018)

  - However, research papers on gender, race, and broadly inequality has seen a substantial increase (Horpedahl and Kling 2020)

- **About us but without us**
  - Developing country authors are underrepresented (Greenspon and Rodrik 2021)
  - Less than 15% of articles authored in development journals by researchers from developing countries (Cummings and Hoebink 2016)

  - African researchers are underrepresented in publishing in development journals focused on Africa (Cummings and Hoebink 2016)
  - And researchers from the 'global south' are vastly underrepresented as presenters in prestigious conferences and authors of top development journals (Amarante et al. 2021)

Gaps and question

- Persistent disadvantages among women and minority scholars not fully explored
- Clear evidence of gender differences in field specialization (Fortin et al. 2021; Sierminska and Oaxaca 2022) but the role of ethnicity/origin in field specialization choices remains understudied

How does country of origin/ region, influence field selection? What are the significant determinants of fields (especially gender) for different regions?

- Gender and ethnicity driven preferences and biases start early and are persistent (Alesina et al. 2019; Hale and Regev 2014) and varies across cultures (Jayachandran 2021; Zafar 2013)
- Ethnicity/country and gender intersect uniquely in choices (e.g., East Asian women vs. US men)

Data

Context: PhD Graduates in the US: Fields chosen in the 2nd or 3rd year, so graduate school factors matter

Data Sources (from 2009-2018)

- Econlit: Articles in economics journals, doctoral degrees data from U.S./Canadian universities
- Academic Analytics: Publications, faculty details, and more for ~400 institutions in the US
- Proquest: Dissertation information and supervisor data
- Name-to-Gender and Origin Mapping: Gender API & Forebears (validated with manual coding),

Variables

- Fields (JEL codes), Country of origin, gender, share of females in the department, graduation year, rank, public university
- Whether a country was formerly a member of the Soviet Union, share of Muslims in the country, or a former colony of Britain, France, Portugal and Spain (view categorizations)

Limitations

- While the study focuses on graduate school experiences, it doesn't capture antecedents like family background or undergraduate education due to data constraints.
- Dropout data not captured; existing statistics indicate varying dropout rates by field, gender, and institution rank (Euler et al. 2018; Stock et al. 2010)

Country of origin/ ethnicity

Methodology:

- Manually gathered data on nationality for 1,000 random students using CVs. Key indicators include: Nationality, Citizenship, Ethnicity, Native language, Country of schooling, and Country of undergraduate education.
- Utilized the **Forebears** genealogy database to match names with nationalities, refining with individual CVs.
- Forebears Outputs: Provides 4 nationality probabilities based on surname incidence and frequency.

Rules for Classification:

- High accuracy observed when a name's probability exceeds 50% and has a >10% gap from the second nationality. This holds true for all non-U.S. nationalities
- The U.S. is an exception, requiring a 75% probability cutoff
- For ambiguous cases (20%), manual checks were repeated. If the top 3 nationalities are from the same region and their combined probability exceeds 60%, the primary nationality prediction is marked as true

Limitations:

- Does not perfectly capture multiracial or multiethnic identities (Hofstra et al. 2022)
- Difficulties arise in determining race, especially for common names in regions like the Anglosphere
- Pixel-based tools have emerged that might aid in classifying skin color, providing an alternative approach.

Example of the output from the database

Reshmi	Sengupta	India	86.06249	Banglades	8.80258	United Sta	1.37784	United Ara	0.75395
Desislava	Byanova	Bulgaria	86.05951	Russia	3.98538	Transnistr	2.32314	Moldova	2.25304
Garrett	DeSimone	United Sta	86.05536	Argentina	5.71228	Canada	2.84038	Italy	1.42735
LaPorchia	Collins	United Sta	86.00322	England	5.09323	Australia	2.39742	Canada	1.54628
Rei	Odawara	Japan	85.98977	China	2.48381	Albania	1.92556	Philippine:	1.52599
Tatsushi	Oka	Japan	85.97822	Ivory Coas	4.40737	Indonesia	2.69732	Egypt	2.19465

Field specialization: JEL

- There can be as many as 5 to 7 JEL codes listed on a dissertation entry (and on publications)
- JEL classification system was developed for use in the Journal of Economic Literature (JEL), and is a standard method of classifying scholarly literature in the field of economic
- About 20 Letter codes and number codes within: we focus on letter codes

Multi-field Specialization:

- 85% of the sample has a singular field specialization (determined by mode), 15% have more than one mode
- Those with multifield and single fields are similar in terms of characteristics – gender, region, distribution of fields
- For clarity, the analysis focuses on individual data and the subset with singular field specializations (exclusive choices), future iterations may explore more flexible forms like Composite Conditional Likelihood method (Sierminska and Oaxaca 2022)

Categorizing JEL Codes

Fields	JEL codes	Detailed JEL codes
Econometrics	C	C. Mathematical and Quantitative Methods
Micro	D	D. Microeconomics
Labor	I, J	I. Health, Education and Welfare, J. Labor and Demographic Economics
Macro/Finance	E, G	E. Macroeconomics and Monetary Economics; G.Financial Economics;
IO	L	L. Industrial Organization
Environmental & Agricultural	Q	Q. Agricultural & Natural Resource Economics, Environmental Econ
Public	H	H. Public Economics
Development/Growth/International	F, O	O. Economic Development, Innovation, Technological Change, and Growth; F. International Economics
Economic History	B, N	B. History of Economic Thought, Methodology, and Heterodox Approaches; N. Economic History
Other	P,A,K,M,R,Y,Z	P. Economic Systems; A. General Econ and Teaching; Z. Other Topics; ; K. Law and Economics; R. Urban, Reg, Real Estate & Transportation Economics

Quick detour: an alternative to JEL codes - Topic Modeling: Extracting Latent Themes from Unstructured Text

- I've also employed Latent Dirichlet Allocation (LDA) to perform topic modeling on PhD economics abstracts on Proquest (Blei et al. 2003)

- Like factor analysis for text, LDA identifies underlying 'topics' in a collection of documents by recognizing patterns in word co-occurrences

- The output gives a probabilistic distribution of words for each topic and a distribution of topics for each abstract

- This allows for a systematic and quantitative understanding of thematic structures

Example of topic modeling output based on Proquest

Difference to JEL Classification?

- Revealed Stability: Unaffected by classification changes or authors' strategic code attributions.
- Multifield Simplification: Assists in identifying dominant topics, validating multifield robustness
- Topic modeling doesn't demand a priori term definitions, offering potential advantages (Ambrosino et al. 2018; Fontana et al. 2019).

Process

- Scraped ~29,000 thesis abstracts from Proquest
- Stopwords + others removed: "model", "chapter", "data", "study", and "result" were added to enhance modeling accuracy
- Stemming: Implemented to refine word roots and enhance topic identification

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
[1,]	"inform"	"health"	"estim"	"countri"	"cost"	"labor"	"price"	"agricultur"	"econom"	"rate"
[2,]	"agent"	"insur"	"model"	"trade"	"environment"	"school"	"market"	"product"	"research"	"financi"
[3,]	"decis"	"care"	"method"	"growth"	"polici"	"educ"	"firm"	"food"	"social"	"market"
[4,]	"game"	"household"	"test"	"develop"	"water"	"wage"	"product"	"farm"	"develop"	"bank"
[5,]	"prefer"	"program"	"variabl"	"econom"	"energi"	"worker"	"consum"	"farmer"	"network"	"risk"
[6,]	"behavior"	"children"	"time"	"product"	"impact"	"student"	"cost"	"household"	"commun"	"shock"
[7,]	"incent"	"cost"	"function"	"tax"	"land"	"employ"	"demand"	"crop"	"institut"	"polici"
[8,]	"optim"	"incom"	"distribut"	"polici"	"econom"	"incom"	"competit"	"rural"	"analysi"	"price"
[9,]	"contract"	"impact"	"propos"	"economi"	"electr"	"skill"	"industri"	"land"	"theori"	"monetari"
[10,]	"choic"	"estim"	"paramet"	"sector"	"develop"	"job"	"qualiti"	"produc"	"process"	"asset"

Examples of categorizations for 10 topics

Health Economics: health, insur, care, houshold, program, children, cost, incok, impact

Finance/Macro: rate, financi, market, bank, risk, shock, polici, price, monetari, asset

Labor & Education Economics: labor, school, educ, wage, worker, student, employ, incom, skill, job

Industrial Organization: price, market, firm, product, consum, cost, demand, competit, indstri, qualiti

Econometrics & Statistical Methods: estim, model, method, test, variabl, time, function, propos, paramet.

Behavioral/ Game Theory: inform, agent, decis, game, prefer, behavior, incent, optim, contract, choic.

International Economics: countri, trade, growth, develop, econom, product, tax, polici, economi, sector.

Environmental & Energy Economics: cost, environment, polici, water, energi, impact, land, econom, electr, develop

Agricultural Economics: agricultur, product, food, farm, farmer, household, crop, rural, land, produc

Economic Theory : econo, research, develop, network, commun, institut, analysi, theory, process

Methodology

- A standard multinomial logistic regression to model field choice (JEL codes) on the characteristics of the individuals making these choices (Sierminska and Oaxaca 2022)

$$\log\left(\frac{\Pr(\text{FieldChoice}_i = j)}{\Pr(\text{FieldChoice}_i = 1)}\right) = \beta_{0j} + \beta_{1j} \times \text{CountryOfOrigin}_i + \beta_{2j} \times \text{Gender}_i$$

$$+ \beta_{3j} \times \text{YearOfGraduation}_i + \beta_{4j} \times \text{FemaleShareDept}_i + \beta_{5j} \times \text{PublicUni}_i + \beta_{6j} \times \text{UniRank}_i + \beta_{7j} \times \text{CountryLevelVars}_i + \varepsilon_i$$

- I estimate and report marginal effects for Regions for each fields

- Then I run separate regressions for different regions and discuss marginal effects of some variables that are significant (focusing on gender and a few other country level variables)

Summary Statistics

Variable	Obs	Mean	Std. dev.
US & Canada	8,849	0.2781	0.4480
Europe	8,849	0.1549	0.3618
South, and Central America	8,849	0.0765	0.2658
South and Central Asia	8,849	0.0923	0.2895
East Asia	8,849	0.3209	0.4668
Sub-saharan Africa	8,849	0.0275	0.1637
West Asia & North Africa	8,849	0.0496	0.2171
Females	8,855	0.2920	0.4547
University Rank	8,857	60.237	55.691
Share of females	8,857	0.1659	0.0979
Muslim share	8,850	0.1134	0.2559
Formerly a Portuguese colony	8,850	0.10836	0.3108
Formerly a Spanish colony	8,850	0.1026	0.3034
Formerly a French colony	8,850	0.0261	0.1594
Formerly a British colony	8,850	0.11503	0.3190
Formerly part of Soviet Union	8,850	0.03571	0.1855

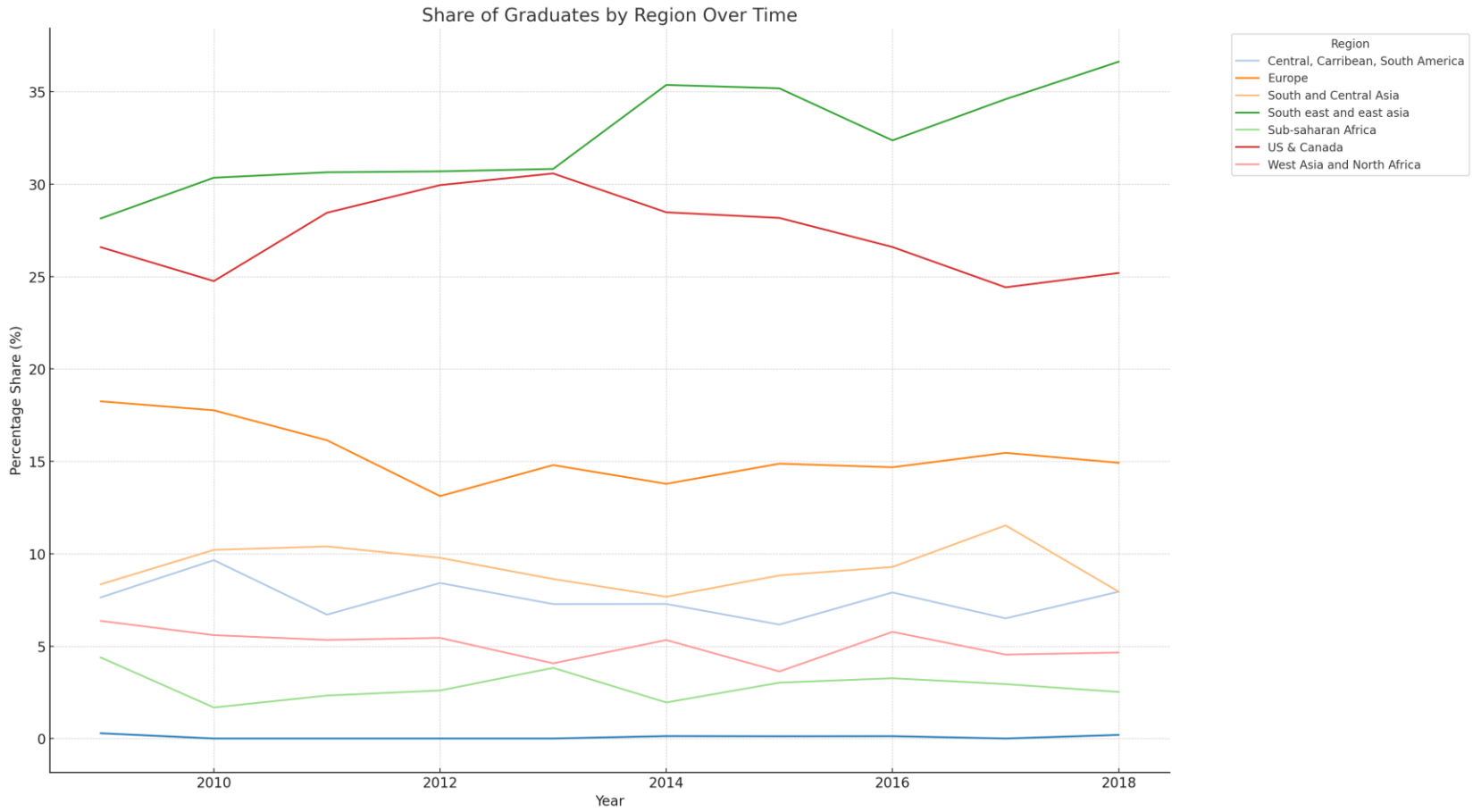
South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Share of PhD graduates by Region and Gender

%	Female	Male
US & Canada	22.81	29.88
Europe	16.92	14.91
South, and Central America, Carribean	6.04	8.31
South and Central Asia	13.13	7.63
East and South east Asia	32.88	31.76
Sub-saharan Africa	2.56	2.84
West Asia, Middle-east, North Africa	5.65	4.68
N	**2,582**	**6,266**

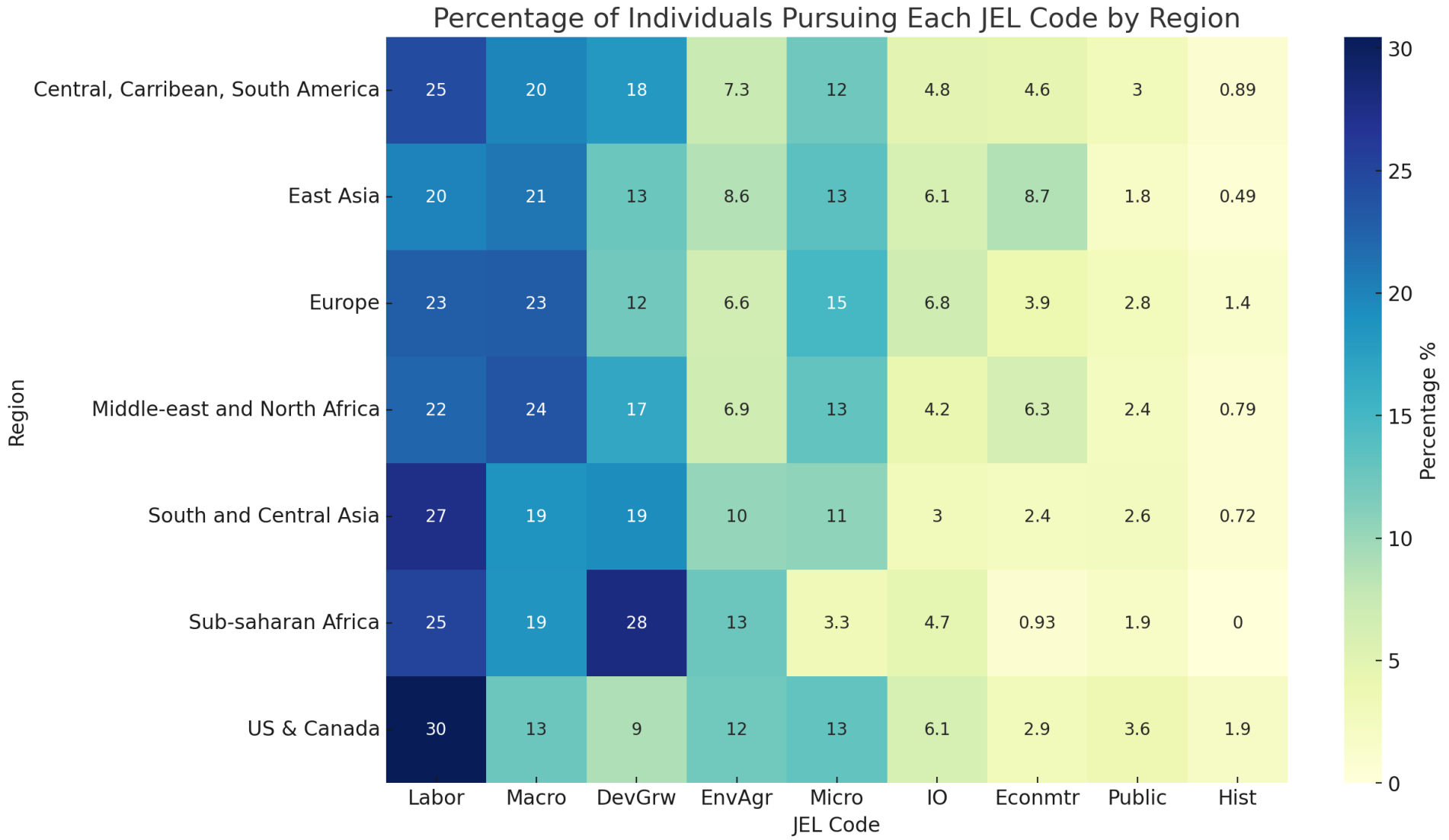
South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Share of PhD grads X region over time



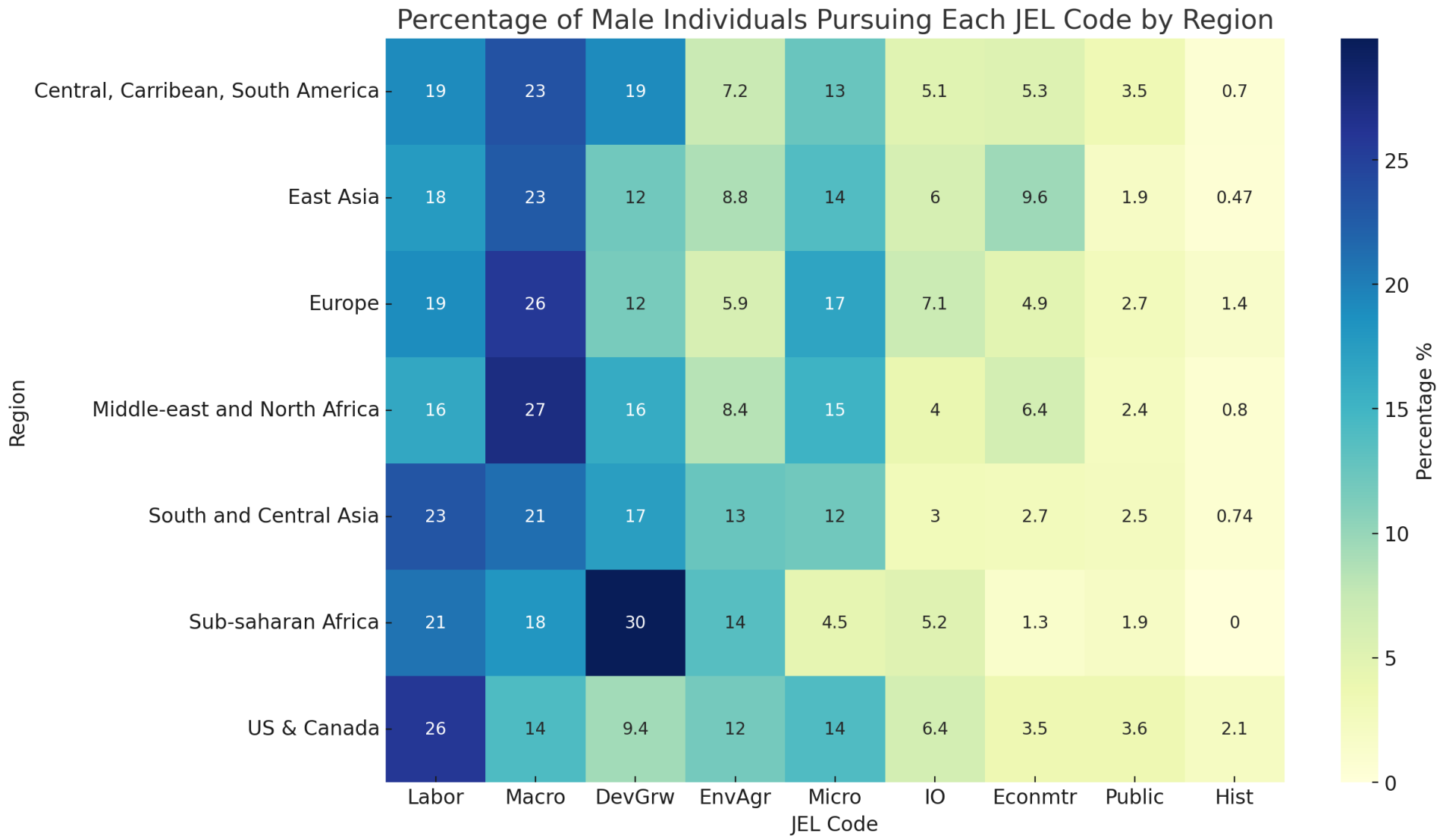
South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Field specialization (JEL codes) across region



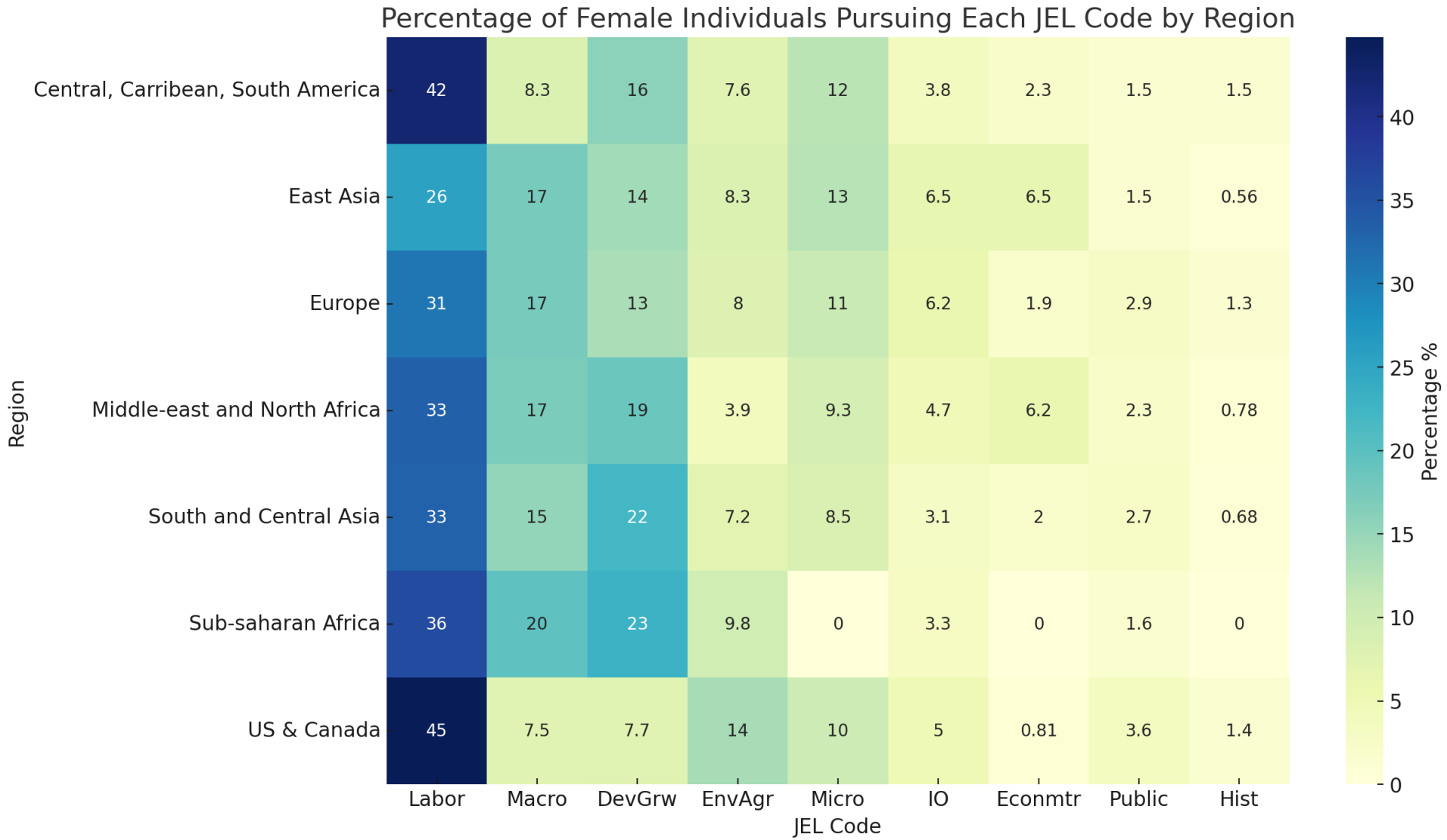
South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Field specialization among Males across regions



South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Field specialization among Females across regions



South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Marginal effects for region from the multinomial logit model

	Econometrics (C)	Micro (D)	Labor/Health (I,J)	Macro/Finance (E,G)	IO (L)	Environ & Agric (Q)	Public (H)	Dev/Growth/Int (O,F)	Econ History (B,N)
Ref: US & Canada									
Europe	0.017	-0.006	-0.062***	0.11***	0	-0.052***	-0.007	0.029*	-0.003
	(0.011)	(0.014)	(0.018)	(0.015)	(0.008)	(0.017)	(0.006)	(0.017)	(0.004)
South and Central America	0.054**	-0.036	-0.046	0.081***	-0.032**	-0.047**	-0.003	0.067***	-0.008
	(0.022)	(0.022)	(0.028)	(0.029)	(0.014)	(0.023)	(0.01)	(0.023)	(0.009)
South and Central Asia	0.037**	-0.014	-0.091***	0.094***	-0.043**	-0.009	-0.021*	0.08***	-0.013*
	(0.015)	(0.022)	(0.03)	(0.026)	(0.018)	(0.022)	(0.011)	(0.02)	(0.007)
East Asia	0.057***	0.001	-0.119***	0.097***	0	-0.037***	-0.019***	0.043***	-0.014**
	(0.011)	(0.009)	(0.018)	(0.015)	(0.007)	(0.014)	(0.005)	(0.014)	(0.006)
Sub-saharan Africa	-0.01	-0.113**	-0.017	0.125***	0	0.033	-0.013	0.164***	-0.156***
	(0.036)	(0.046)	(0.04)	(0.031)	(0.019)	(0.031)	(0.015)	(0.023)	(0.04)
Middle East and North Africa	0.048***	0.011	-0.098***	0.12***	-0.02	-0.053***	-0.011	0.073***	-0.012
	(0.012)	(0.024)	(0.028)	(0.025)	(0.015)	(0.019)	(0.009)	(0.022)	(0.009)

South America: Argentina, Brazil, Chile, Mexico; **Europe:** England, France, Germany, Italy, Poland, Romania, Spain; **South and Central Asia:** Bangladesh, India, Iran, Pakistan, Sri Lanka, Uzbekistan; **East Asia:** China, Japan, South Korea, Thailand, Vietnam; **Sub-Saharan Africa:** DR Congo, Ethiopia, Ghana, Kenya, Nigeria, South Africa; **US & Canada:** Canada, United States + Australia and New Zealand **Middle-east and North Africa:** Egypt, Israel, Saudi Arabia, Sudan, Turkey

Key takeaways

What we knew already?

- **Publications:** Male authors are over-represented in micro and macroeconomics, female authors are over-represented in development economics (Onder and Yilnazkuday, 2020)
- **Conferences:** Women are underrepresented in Finance and Macroeconomics and overrepresented in Microeconomics (Chari & Pinkham, 2017; Beneito et al., 2021)
- **Dissertations:** Sierminska and Oaxaca (2022): Less likely Macro, IO, Dev/Growth and even in Labour; More likely in Agri

What we find (some examples):

- East Asians, South Asians and Middle-east and North Africans are more likely to pursue Econometrics
- Labor/Health most prevalent in the US and East Asians least likely to do Labour/Health
- Macro/finance least prevalent in the US
- Dev/growth most popular in Sub-Saharan Africa, least in the US

- **Gender as a determinant**
- Being a female increases the likelihood of specializing in Labor in US & Canada, South Asia and Europe BUT reduces the likelihood of pursuing it in South and Central America
- Being a female reduces the likelihood of pursuing Macro in most regions, other than East Asia where it's the opposite

& more – all indicating the role of region in mediating gender level differences across different fields

Determinants for field X region: quite some variation (based on marginal effects)

US & Canada: **Female**: Less likely in Econometrics & Macro; more likely in Labor/Health

Europe: **Female**: Less likely in Econometrics, Micro & Macro; more likely in Labor/Health | **Public University**: Less chance in Micro & Macro; more in Labor & Environment | **Formerly Soviet**: More likely in Econometrics & Micro

South and Central America : **Female**: Less likely in Labor/Health, Macro & Finance

South and Central Asia: **Female**: More likely in Labor/Health; less in Environment/Agri | **Muslim Share**: Reduces chances in Environmental & Agri and Dev/Growth; **Formerly Soviet**: increases in Macro/Finance & IO

East Asia: **Female**: Less likely in Econometrics & IO; more in Macro/Finance | **Muslim Share**: More likely in Dev/Growth & Econ History

Sub-Saharan Africa: **Muslim Share**: More likely in Environment & Agri

Middle-east and North Africa: **Female**: Less likely in Environmental & Agri; more in Labour/Health | **Muslim Share**: Reduces chances in Environment & Agri and Public; increases in Macro/Finance

Conclusion

- There's a surge in understanding gender focused disparities as but region (read- ethnicity, economy, culture, norms) could play an important role in determining the experiences but hasn't been explored
  - "women avoid theory" don't fit all (e.g.- East Asian women more likely to pursue Econometrics than Western men)

- We find that field specialization among beginning economists varies substantially across country/region of origin and may be an important component in understanding existing gender-based differences in fields

- Could inform efforts to increase representation (at the level of collectives and institutions), in STEM and Economics

- More intersectional understanding may help better address disparities and institutional barriers
  - Women, especially those from foreign origins, are observed to have pronounced tenure gaps (Chen et al. 2008; 2020)
  - Gaps in promotion and career outcomes not solely due to productivity differences, suggesting field specialization's potential role (Ceci et al. 2014; Ginther and Kahn 2021)

- Future work would include exploring other determinants of field specific choices – employment and salaries (a la Sierminska and Oaxaca 2022) + citation metrics, research grant availability and other categorizations at the field and supervisor level
  - & exploring topic modeling and Genetic distance for sensitivity

Thank you!

karansinghal1993@gmail.com

Appendix

Other country-level categorizations

- **Formerly part of Soviet Union:** Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan
- **Portugal and former Portuguese Colonies:** Angola, Argentina, Brazil, Cape Verde, Ecuador, Ghana, Guinea, India, Indonesia, Macau, Mozambique, Paraguay, Peru, Sri Lanka, Uruguay, Venezuela
- **Spain and former Spanish Colonies:** Argentina, Belgium, Bolivia, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Guinea, Haiti, Honduras, Italy, Jamaica, Luxembourg, Mexico, Morocco, Netherlands, Paraguay, Peru, Philippines, Portugal, Spain, Trinidad and Tobago, Uruguay, Venezuela
- **UK and former British Colonies:** Australia, Bangladesh, Cameroon, Cyprus, Egypt, Ghana, Honduras, Hong Kong, India, Iraq, Ireland, Jamaica, Jordan, Kenya, Kuwait, Malawi, Malaysia, Myanmar, New Zealand, Nigeria, Oman, Pakistan, Palestine, Qatar, Singapore, South Africa, Sri Lanka, Sudan, Uganda, Zambia
- **France and former French Colonies:** Algeria, Cambodia, Cameroon, DR Congo, Dominican Republic, France, Guinea, Haiti, Honduras, Ivory Coast, Laos, Lebanon, Madagascar, Mauritania, Morocco, Niger, Senegal, Sudan, Syria, Trinidad and Tobago, Vietnam
- **Share of Muslim population** (in 2015)

Sources: World Population Review, World Atlas, Angeles (2012) (Data)

Marginal effects

$$\mathrm{ME}_{jk} = \frac{\partial \Pr(\mathrm{FieldChoice}_i = j)}{\partial x_k} = \Pr(\mathrm{FieldChoice}_i = j) \times [\beta_{kj} - \sum_{m=1}^{10} \Pr(\mathrm{FieldChoice}_i = m) \times \beta_{km}]$$

ME_{jk} is the marginal effect of predictor x_k on the probability of choosing field j.
$\Pr(\mathrm{FieldChoice}_i = j)$ is the predicted probability of individual i choosing field j.
β_{kj} is the coefficient of predictor x_k for field j from the multinomial logit regression.
The term $\sum_{m=1}^{10} \Pr(\mathrm{FieldChoice}_i = m) \times \beta_{km}$ is the weighted average of the coefficients of x_k across all fields, where the weights are the predicted probabilities of choosing each field.

This equation calculates the change in the predicted probability of individual i choosing field j for a one-unit change in x_k, while holding other predictors constant.

Categorization of countries

- **South and Central America:** Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Paraguay, Peru, Trinidad and Tobago, Uruguay, Venezuela
- **Europe:** Albania, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, England, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Kosovo, Latvia, Lithuania, Luxembourg, Moldova, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Russia, Scotland, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Transnistria, Ukraine, Wales
- **South and Central Asia:** Afghanistan, Bangladesh, India, Iran, Kazakhstan, Kyrgyzstan, Nepal, Pakistan, Sri Lanka, Tajikistan, Turkmenistan, Uzbekistan
- **East Asia:** Cambodia, China, Hong Kong, Indonesia, Japan, Laos, Macau, Malaysia, Mongolia, Myanmar, North Korea, Philippines, Singapore, South Korea, South Korean, Taiwan, Thailand, Vietnam
- **Sub-Saharan Africa:** Angola, Benin, Burkina Faso, Cameroon, Cape Verde, DR Congo, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Ivory Coast, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mozambique, Namibia, Niger, Nigeria, Senegal, Sierra Leone, Somalia, South Africa, Tanzania, Togo, Uganda, Zambia, Zimbabwe
- **US & Canada:** Canada, United States + Australia and New Zealand
- **West Asia, Middle-east, and North Africa:** Algeria, Armenia, Azerbaijan, Cyprus, Egypt, Georgia, Iraq, Israel, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Turkey, Yemen

Alternate measure – Genetic Distance

- Genetic distance, as introduced by Spolaore and Wacziarg (2009) and based on data from Cavalli-Sforza et al. (1994), measures the time since two populations shared a common ancestor, akin to a molecular clock

What Does Genetic Distance Capture?

- Represents divergence in implicit beliefs, customs, habits, and biases transmitted across generations, both biologically and culturally. Serves as a metric for differences in characteristics conveyed over time.
- Measures the difference in gene distributions between two populations, focusing on neutral genes that change independently of selection pressures.
- Rationale: Divergence in neutral genes provides insights into lines of descent, with most random genetic changes occurring consistently

Methodological Considerations

- Two measures considered: Distance between the largest ethnic groups in paired countries & Weighted genetic distance, accounting for countries with multiple genetically distant subpopulations.
- Notably, there's a strong correlation between genetic and geographical distances (Saha and Mishra 2020)

Application and Interpretation:

- Potential to use country-level genetic distance relative to the USA, with a score of 0 for the USA.
- Advantage: Continuous measure, but careful interpretation is required.

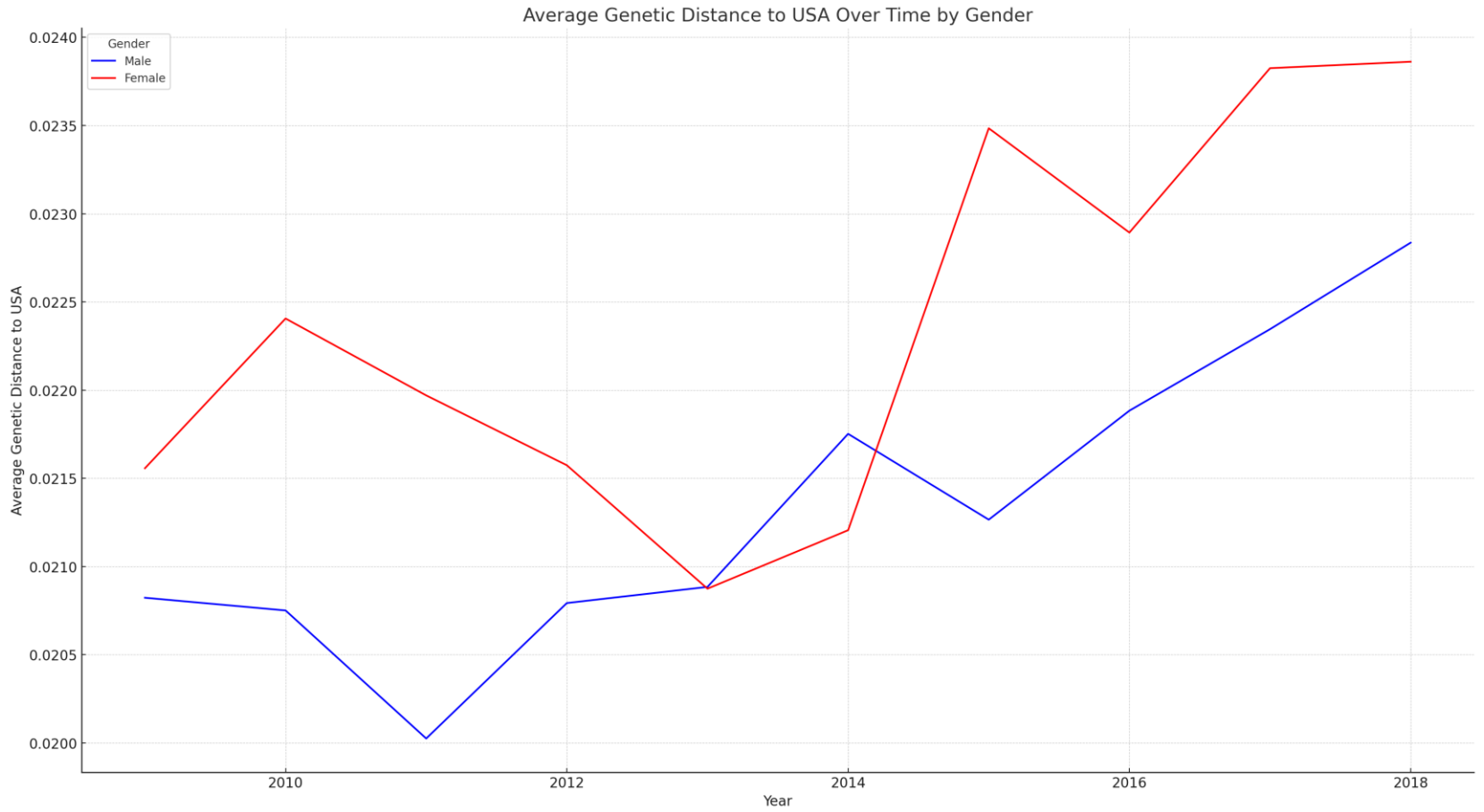
Genetic distance and regions

	Mean	Std. err.	[95% conf.	interval]
Us, Canada	0.000157	2.89E-05	9.98E-05	0.000213
Europe	0.013451	8.15E-05	0.013291	0.01361
West Asia and North Africa	0.018546	0.000213	0.018128	0.018963
South and Central Asia	0.021085	0.000104	0.020881	0.021289
South, and Central America	0.028317	0.00043	0.027474	0.029159
East Asia	0.04069	3.18E-05	0.040627	0.040752
Sub-saharan Africa	0.046304	0.000361	0.045596	0.047012

JEL X genetic distance

JEL	Mean score
Econ History (B,N)	0.012
Public(H)	0.017
Labor/Health (I,J)	0.0191
Environ & Agric (Q)	0.02
Others	0.021
Micro (D)	0.021
IO (L)	0.021
Macro/Finance (E,G)	0.024
Dev/Growth/Int (O,F)	0.024
Econometrics (C)	0.028

Genetic distance X gender over time



Single vs Multifield sample

- Compared the restricted sample – 85% to the 115% (where those with two fields are counted twice)
- Sample is largely similar across gender, region, and JEL

	Multiple fields	**Single Field**
JEL		
Others	7.34	6.42
Econometrics (C)	5.77	5.12
Micro (D)	14.68	12.95
Labor/Health (I,J)	21.41	24.57
Macro/Finance (E,G)	18.11	18.86
IO (L)	6.32	5.68
Environ & Agric (Q)	8.52	9.41
Public(H)	3.38	2.63
Dev/Growth/Int (O,F)	13.39	13.27
Econ History (B,N)	1.08	1.08
Total	100	100
N	10,218	7,496

	Multiple fields	**Single Field**
Females	29.05	29.42
US & Canada	28.08	27.45
Europe	15.61	15.34
South and Central America	7.75	7.52
South and Central Asia	9.17	9.32
East Asia	31.84	32.44
Sub-saharan Africa	2.67	2.87
West Asia and North Africa	4.89	5.06
Genetic distance	0.0215	0.0217
Total	100	100
N	10,218	7,496

Technical explanation of topic modeling

1. **Methodology**:
    1. Used LDA, a generative probabilistic model, to segment text into discrete topics.
    2. LDA assumes each document is a mix of topics, and a topic is a mix of words.
2. **Bayesian Framework**:
    1. LDA relies on two key Dirichlet-distributed priors:
        1. Document-Topic (θ): Proportions of topics in a given document.
        2. Topic-Word (β): Proportions of words in a given topic.
    2. Bayesian inference is used to obtain the posterior distribution of the latent variables given the observed data.
3. **Hyperparameters**:
    1. α (Alpha): Controls the mixture of topics for any given document.
    2. β (Beta): Influences the mixture of words describing a topic.
    3. Adjusting these can influence topic granularity.
4. **Model Evaluation & Calibration**:
    1. Used perplexity and coherence scores to fine-tune the model.
    2. Iterative approach with Gibbs sampling or variational inference for parameter estimation.

15 topics

Public Finance: tax, polici, cost, regul, effici, public, govern, environment, program, incent.

Behavioral & Game Theory: inform, social, decis, agent, network, behaviour, individu, game, theori, prefer.

Labor Economics: labor, incom, wage, worker, employ, household, market, job, inequ, skill.

Agricultural Economics: agricultur, product, land, farm, water, food, farmer, crop, econom.

Finance & Banking: bank, financi, hous, credit, market, debt, loan, borrow, default, risk

Education Economics: school, student, educ, colleg, program, public, enrol, impact, score, teacher

Financial Markets & Asset Pricing: risk, market, return, stock, price, asset, financi, volatil, investor, inform

Econometrics & Statistical Methods: estim, model, method, test, variabl, function, distribut, time, propos, paramet.

Political Economy: econom, develop, polit, govern, institut, polici, countri, growth, region, local

International Trade/ Energy Economics: trade, export, price, countri, market, product, energi, oil, import, electr

Health Economics: health, care, insur, cost, patient, hospit, medic, drug, servic, treatment

Macroeconomics: rate, polici, shock, monetari, economi, exchang, countri, inflat, real, macroeconom

Family & Gender Economics: household, women, children, health, famili, child, parent, time, impact, age.**Industrial Organizatio:** price, market, consum, product, demand, cost, competit, firm, qualiti, auction.

On the challenges of dropouts (Jaksztat et al. 2021)