CAPÍTULO IX

Diferencias de género en la investigación económica. Un enfoque de aprendizaje automático

J. Ignacio Conde-Ruiz
Juan-José Ganuza
Manu García
Luis A. Puch

Utilizando técnicas de aprendizaje automático no supervisado y una nueva base de datos compuesta por los resúmenes de todos los artículos publicados en revistas *Top5* (T5) de economía para el periodo (2002-2019), se demuestra que existen diferencias horizontales persistentes y significativas en la forma en que hombres y mujeres abordan la investigación en economía. Utilizando el modelo temático estructural (*STM*) se estiman los temas de investigación de los artículos, y se asignan los artículos (y los autores) a dichos temas. La distribución resultante de temas de investigación por sexo demuestra que hay diferencias horizontales significativas de género en la eleccion de temas de investigación. Estos resultados son importantes por varias razones: i) las publicaciones de T5 son clave para la carrera investigadora y también para determinar la trayectoria de la investigación económica; ii) los resultados son robustos en el sentido de que se generan automáticamente con un modelo probabilístico sin ninguna asignación arbitraria de los trabajos a categorías o campos de investigación preestablecidos; iii) por último, los recientes resultados teóricos de Conde-Ruiz y Ganuza (2017); Conde-Ruiz *et al.* (2022) y Siniscalchi y Veronesi (2020) muestran que las diferencias "horizontales" de género en la elección del tema de investigación pueden conducir a una brecha de género permanente.

Palabras clave: brecha de género, aprendizaje automático, estimación de temas latentes, revistas de gran impacto en economía.

1. INTRODUCCIÓN

La economía digital ha venido acompañada por una gran capacidad de generar, analizar y gestionar gran cantidad de datos. En el pasado, se usaban complejas técnicas econométricas para resolver problemas de inferencia con bases de datos muy limitadas. En el nuevo paradigma, la idea es utilizar la capacidad computacional de algoritmos de inteligencia artificial para explotar ricas bases de datos, con modelos probabilísticos relativamente sencillos. Esto permite obtener unos resultados empíricos que son robustos en una dimensión particular, provienen de algoritmos sin hipótesis o intervención humana que *a priori* condicionen dichos resultados obtenidos.

En este capítulo vamos a presentar un análisis empírico llevado a cabo en (2022) que tiene estas características. El objeto del análisis es identificar diferencias de género en la elección de temas de investigación en economía y finanzas. Esta pregunta ha sido analizada con anterioridad utilizando los códigos JEL (*Journal of Economic Literature*) que los propios autores elijen para describir sus artículos o las áreas de conocimiento de los departamentos de economía. Esta metodología depende de un juicio subjetivo humano y además son frecuentemente descripciones poco precisas del objeto de la investigación. Por el contrario, Conde-Ruiz *et al.* (2022) utiliza un enfoque de aprendizaje automático no supervisado que tiene las siguientes ventajas. En primer lugar, analiza todos los artículos publicados en las mejores revistas de economía, las denominadas *Top-five*, T5, entre 2002-2019. Esta es una gran base de datos, cuyo análisis de textos sólo se ha hecho factible con el desarrollo de estas nuevas herramientas de inteligencia artificial. En segundo lugar, no se introduce ninguna guía sobre el contenido de los temas de investigación. El algoritmo sin ningún tipo de supervisión e intervención humana determina cuál es el conjunto de temas que mejor se adapta a los datos. El algoritmo también asigna los artículos a los diferentes temas de investigación. Una vez asignados el artículo (y sus autores) a los diferentes temas estimados, se puede analizar si los autores masculinos y femeninos presentan una distribución entre los temas de investigación diferente. Este último paso tampoco es trivial, dado que el género no es observable directamente. De nuevo, los algoritmos resuelven este problema, comparando los nombres de pila de los autores con grandes bases de datos que relacionan nombres de pila y género. El algoritmo estima una probabilidad de que el sexo esté bien asignado, reduciendo la intervención humana a un conjunto muy pequeño de autores donde esta probabilidad es significativamente diferente de 1.

El capítulo comienza con la motivación de la pregunta de investigación. ¿Por qué es importante identificar diferencias horizontales de género en la elección de temas de investigación? Recientes artículos de investigación muestran mecanismos teóricos por los que las diferencias horizontales (que no afectan a la calidad o la productividad) se trasladan a discriminación y brechas de género. También, se documenta la existencia de esta brecha de género en numerosos sectores de actividad económica y específicamente, en las carreras académicas en las áreas de economía y finanzas.

La tercera sección explica en detalle la metodología seguida. La construcción de la base de datos y su preparación para el análisis de textos. Se explica la elección del mecanismo de aprendizaje automático utilizado (*STM, Structural Topic Model*) y su funcionamiento. La sección cuarta presenta el análisis de textos llevado a cabo en Conde-Ruiz *et al.* (2022) donde se presentan los temas de investigación estimados y la distribución de artículos, autores y género sobre ellos. El principal resultado que se obtiene es que las distribuciones de hombres y mujeres sobre los temas de investigación difieren significativamente. La única variable que se introduce al algoritmo es el número de temas a investigar. En el análisis principal se estiman 54 temas, que es el número que mejor se adapta al conjunto de datos (artículos). Mientras que esta es la estrategia adecuada para ver si difieren hombres y mujeres en la distribución de los temas de investigación, tiene el inconveniente de que no siempre es fácil interpretar semánticamente dichos temas. En la sección quinta, se fuerza al algoritmo a estimar 15 temas, lo que permite relacionar mejor los temas con las áreas tradicionales de investigación. Gracias a ello, podemos analizar específicamente el campo de las finanzas, y demostrar que la importancia de este campo en las T5 ha aumentado, pero esto no ha llevado a reducir la poca representación de las mujeres en esta área de investigación. Por último, la sección sexta presenta las conclusiones.

2. BRECHAS DE GÉNERO EN PROMOCIONES PROFESIONALES, DIFERENCIAS HORIZONTALES Y COMITÉS

La discriminación ocurre cuando algunos trabajadores son tratados de manera diferente a otros por sus características personales, tales como género, raza, edad, nacionalidad, orientación sexual, etc., que no están relacionadas con su productividad (Arrow, 1973). La discriminación no sólo conduce a una mayor desigualdad y falta de equidad, también puede generar pérdidas de eficiencia: pérdida de talento, falta de incentivos para invertir en capital humano por parte del grupo discriminado e ineficiente asignación de recursos.

En este capítulo vamos a centrarnos específicamente en la discriminación por razones de género. Observamos brechas de género en el desarrollo profesional en muchos ámbitos. Brechas de género que existen a pesar de los esfuerzos realizados por toda la sociedad para luchar contra la discriminación en las últimas décadas. Gracias a estos avances, gran parte de las barreras y mecanismos de discriminación directa y explícita, han sido eliminados o reducidos. Sin embargo, la brecha de género todavía persiste; por ejemplo, la diferencia salarial es del 17 % sin tener en cuenta las características personales (y del 14 % si se tienen en cuenta). Las mujeres directivas solamente representan el 16 %. Si nos fijamos en los consejos de administración de las empresas cotizadas, apenas llegan al 26 %, cifra que se reduce al 5,5 % cuando nos fijamos en las consejeras ejecutivas. También, encontramos que las mujeres están infrarrepresentadas en las carreras STEM (ciencia, tecnología, ingeniería y matemáticas). A pesar de que más mujeres que hombres acceden a la universidad, únicamente el 11 % de las mujeres obtuvieron un título en STEM, mientras que en el caso de los hombres, ese porcentaje se eleva al 36 %.

Para cerrar la brecha, primero tenemos que entender cuáles son las barreras que impiden a las mujeres alcanzar altos cargos. Existe una amplia bibliografía que provee argumentos del lado de la oferta como: a las mujeres no les gusta la competencia por los ascensos (Niederle y Vesterlund, 2010); las mujeres evitan el estrés y el desequilibrio entre la vida laboral y personal de los altos cargos (Azmat y Ferrer, 2017); interrupciones de la carrera profesional debido al cuidado de los hijos (Bertrand, Goldin y Katz, 2010)[1].Sin embargo, los mecanismos del lado de la demanda del llamado techo de cristal son mucho menos conocidos, aunque se ha mostrado que puede existir sesgos por estereotipos (Reuben, Sapienza y Zingales, 2014; Bordalo, Coffman y Gennaioli, 2019, y Bohren y Rosenberg, 2019).

En un reciente artículo Conde-Ruiz *et al.* (2022) intentan contribuir a esta literatura, mostrando que el fenómeno del techo de cristal puede ser debido a diferencias "horizontales" entre hombres y mujeres. Por diferencias horizontales nos referimos a diferencias entre hombres y mujeres que no afectan a su productividad, por ejemplo, su cultura, sus aficiones o el área en que están especializados o están trabajando. Conde-Ruiz *et al.* (2022) proponen un modelo teórico que explica cómo las diferencias horizontales pueden generar brechas de género. El modelo tiene las siguientes características. Hombres y mujeres pertenecen a dos grupos que son iguales en términos de productividad esperada (aunque dentro de los grupos los individuos son heterogéneos en términos de talento, pero exanteidénticos) pero que se diferencian horizontalmente. La productividad de los trabajadores se observa imperfectamente y la evalúa un comité de evaluación mediante entrevistas e indicadores similares. El resultado de este proceso de evaluación determina las retribuciones de los trabajadores (oportunidades de carrera) y sus promociones. Los miembros del comité tienen un sesgo en la precisión de la evaluación, aunque son imparciales y su objetivo es promocionar a los trabajadores con la mayor productividad. No obstante, son mejores evaluando a trabajadores del mismo grupo (por ejemplo, porque la comunicación es mejor). Hombres y mujeres tienen el mismo peso en el conjunto de la población, pero las mujeres pueden estar infrarrepresentadas en los comités de promoción (por ejemplo, tal como hemos comentado anteriormente, la proporción de directivas es menor que la de los hombres). Si esto sucede inicialmente, aunque no son discriminadas explícitamente, son evaluadas peor.

Si suponemos que la productividad de los trabajadores depende de la inversión en capital humano específico[2], y los incentivos a invertir en capital humano dependen de la calidad de la evaluación por parte de estos comités, podemos concluir que la infrarrepresentación en los comités conlleva que las mujeres tienen menos incentivos a invertir en capital humano y Conde-Ruiz *et al.* (2022) muestran que esto puede llevar a una "trampa" dinámica. Consideremos que la composición de los comités se determina endógenamente según la proporción de cada grupo entre los trabajadores identificados como más productivos en el periodo

[1] Véase Matsa y Miller (2011) para una breve revisión de esta literatura.

[2] El capital humano va más allá del nivel de educación observable, lo debemos considerar como un concepto amplio que recoge todos los atributos y las inversiones en el aumento de la productividad en dimensiones difíciles de evaluar. Siguiendo a Arrow (1973), podemos incluir entre ellos la constancia, la puntualidad, la capacidad de respuesta, el liderazgo, el esfuerzo en la experiencia laboral previa o la iniciativa.

anterior. En ese caso, si inicialmente las mujeres están infrarrepresentadas en los comités, las mujeres invertirán menos en capital humano, promocionarán en menor proporción que los hombres y como consecuencia seguirán estando infrarrepresentados en los comités, generando un estado estacionario en el que existe una brecha de género.

En definitiva, Conde-Ruiz *et al.* (2022) demuestran que diferencias horizontales entre géneros pueden generar una brecha de género permanente. Siniscalchi y Veronesi (2020) obtienen un resultado similar centrándose en el mercado laboral académico y señalan una trampa de discriminación involuntaria vinculada al que llaman sesgo de autoimagen. Estos autores construyen un modelo de generaciones solapadas con dos grupos de investigadores con características de investigación igualmente deseables (pero ligeramente diferentes) y distribuciones de productividad *ex ante* idénticas. En el modelo, los investigadores veteranos (que estén en su segundo periodo) evalúan a los investigadores jóvenes y tienen un sesgo hacia aquellos que investigan en un campo similar al suyo. Como en Conde-Ruiz *et al.* (2022), en este entorno si un grupo está ligeramente sobrerrepresentado entre los evaluadores, este grupo (y sus características específicas de investigación) pueden dominar para siempre.

Por lo tanto, Conde-Ruiz *et al.* (2022) y Siniscalchi y Veronesi (2020) demuestran que diferencias horizontales de género independendientes de la productividad, pueden generar brechas en las remuneraciones y promociones entre hombres y mujeres. En este trabajo, vamos a identificar diferencias horizontales entre hombres y mujeres en el entorno académico de la economía (y luego específicamente en las finanzas) que pueden potencialmente ayudar a explicar la clara evidencia de brecha de género en este sector. En las universidades americanas en el campo de la economía, a principios de este siglo, el 35 % de los estudiantes de doctorado y el 30 % de los profesores ayudantes eran mujeres. Desde entonces, estas cifras no han aumentado y la proporción de profesoras ayudantes en las 10 mejores escuelas ha descendido, siendo menos del 20 % en 2019. Las cifras de titulares y catedráticas no son signigficativamente mejores.

3. APRENDIZAJE AUTOMÁTICO, ANÁLISIS DE DATOS Y DIFERENCIAS HORIZONTALES EN LA INVESTIGACIÓN EN ECONOMÍA

Dados estos resultados es importante indagar si efectivamente existen diferencias horizontales entre hombres y mujeres en la investigación en economía que pudieran ayudar a explicar las brechas de género observadas. Para ello, vamos a analizar (siguiendo a Conde-Ruiz *et al*., 2022) todos los artículos publicados en las principales revistas de investigación económica durante el periodo 2002-2019. Estas revistas denominadas *Top Five* (T5) son: *The American Economic Review, Econometrica, The Journal of Political Economy, The Quarterly Journal of Economics*, y *The Review of Economic Studies*. En total se trata de 5.311 artículos, que utilizamos para crear una base de datos donde para cada artículo recogemos los nombres de los autores, la fecha de publicación, la revista y el resumen del mismo. La idea central es utilizar esta base datos para caracterizar los patrones de publicación de hombres y mujeres

en estas revistas líderes. La idea de centrarnos en publicaciones en las T5 para este objetivo se puede justificar por diferentes razones. Primero, el proceso de publicación se parece mucho a los procesos de evaluación por comités de los artículos teóricos que fundamentan la investigación. En segundo lugar, la influencia de las publicaciones T5 es enorme, tanto en lo referente a la investigación en sí misma (dado que estas publicaciones son a su vez las más citadas) como en las carreras profesionales y la promoción de los investigadores. Heckman y Sidharth (2020) analizan las decisiones de *tenure* (la promoción a un contrato permanente de un investigador) de los 35 departamentos de economía más importantes de EE. UU. y concluyen que las publicaciones en T5 son una variable explicativa muy potente de la promoción a *tenure*.

La estrategia que vamos a seguir en nuestro análisis empírico es identificar los temas de investigación que hay detrás de las publicaciones seleccionados y verificar si la distribución de estos temas, difiere entre hombres y mujeres. Una primera forma de llevar a cabo este análisis sería centrarnos en los denominados códigos JEL de los propios artículos. Sin embargo, estos códigos tienen muchas limitaciones, primero son demasiado amplios y frecuentemente no son una buena descripción de los temas tratados en el artículo. Luego son asignaciones arbitrarias realizadas por los mismos autores. En su lugar, nuestra estrategia empírica consiste en utilizar técnicas de aprendizaje automático no supervisado para descubrir la estructura oculta de nuestros documentos de texto[3]. El aprendizaje automático además de modelos predictivos, se utiliza para identificar patrones en grandes bases de datos, reduciendo su dimensionalidad en un número limitado de variables. Con otras palabras, queremos proporcionar una representación de baja dimensión (temas) de un objeto de alta dimensión (resúmenes), conservando en lo posible su contenido informativo. En nuestro problema esto significa que tomando como variable de entrada los propios resúmenes de los artículos, identifique los temas de investigación y asigne los artículos a ellos. Por no supervisado denotamos la ausencia de intervención humana para identificar dichos temas latentes.

Del universo de algoritmos de análisis de textos que existen para la estimación de temas, hemos elegido el *Structural Topic Model (STM)* desarrollado por Roberts, Stewart y Tingley (2019). La ventaja de este algoritmo es que permite incorporar metadatos a nivel de documento en un modelo de texto probabilístico. En nuestro caso, para mejorar la estimación de los temas, se puede utilizar la información adicional de los artículos, referente a los nombres de las revistas y las fechas de publicación. Aunque vamos a utilizar el algoritmo *STM*, este puede interpretarse como un refinamiento del algoritmo *LDA* (*Latent Dirichlet Allocation*) desarrollado por Blei (2003), que es el algoritmo de aprendizaje automático más popular en la reducción de la dimensionalidad de los documentos de texto[4]. A continuación, queremos explicar la idea central del funcionamiento del algoritmo y cómo se estiman los temas.

Comenzamos describiendo el tratamiento de los datos. Extraemos todas las palabras de nuestros 5.311 resúmenes (o documentos). En primer lugar, tenemos que "limpiar" este

[3] Para una excelente introducción no técnica al aprendizaje automático, véase Hansen, McMahon y Prat (2017).

[4] Para la descripción técnica del algoritmo *LDA*, véase el artículo original de Blei (2003) y también Hansen, McMahon y Prat (2017) que es el primer documento que utiliza el algoritmo *LDA* en la literatura económica.

conjunto de palabras para reducir el vocabulario y seleccionar los términos con más contenido informativo. Centrarnos en las palabras con más significado semántico, nos ayuda a estimar mejor los temas. El *corpus* es el conjunto de palabras únicas que obtenemos, después de convertirlas a minúsculas y eliminar del texto original en bruto las palabras sin significado semántico utilizando la lista SMART, desarrollada en la Universidad de Cornell en 1960. Eliminamos, por ejemplo, preposiciones como "para" o "en". Además, reducimos las palabras hasta obtener su raíz lingüística original ("educ" en lugar de "education") y eliminamos las palabras que aparecen solamente una o dos veces. Este proceso en nuestro caso, transformó un conjunto de 13.835 términos diferentes a un *corpus* de 4.241 palabras únicas.

El segundo paso es representar nuestros datos de texto en una matriz documentos-términos de D filas (5.311 resúmenes) y V columnas (4.182 palabras únicas en nuestro *corpus*) donde el elemento (d, v) de la matriz es el número de veces que la palabra v_{th} aparece en el resumen d_{th}. Esta matriz documentos-términos que reduce la dimensionalidad de nuestras variables de texto originales es el *input* del algoritmo. Nuestro objetivo es encontrar un modelo probabilístico que sea capaz de explicar la matriz documentos-términos en dos pasos adicionales. Primero, identificando K temas en nuestros *corpus* y luego representando los documentos como una combinación de esos temas. ¿Qué es un tema? El tema k es una distribución de probabilidad β_k sobre todas las palabras únicas de nuestro *corpus*, donde β_k^v es la probabilidad de que el tema k genere la palabra v. Cada documento d tiene su propia distribución sobre el conjunto de temas θ_d. Esto significa que cada documento/resumen es una combinación lineal de los temas. Por tanto, θ_d^k significaría el peso del tema k en el documento d. El modelo probabilístico de temas se describe mediante estas distribuciones de temas β_k y de documentos θ_d. Teniendo en cuenta esto, la probabilidad de que una palabra elegida arbitrariamente en el documento d coincida con el término v_{th} es $p_{dv} = \sum_k \beta_k^v \theta_d^v$. Utilizando estas probabilidades, podemos obtener la probabilidad total de nuestros datos, $\prod_d \prod_v p_{d,v}^{n_{d,v}}$, donde $n_{d,v}$ corresponde a los elementos de la matriz documentos-términos (el número de veces que la palabra v_{th} aparece en el documento/resumen d_{th})[5] .

Una variable importante del algoritmo es el número de temas a estimar. Podemos seguir dos estrategias. Una, encontrar el número de temas que mejor se ajuste a los datos, lo que suele conducir a un K óptimo que normalmente es un número grande. En nuestro caso este número es $k = 54$. El problema de esta estrategia es que los temas no son fácilmente interpretables. La alternativa es forzar al algoritmo a utilizar un número determinado de temas para facilitar el contenido semántico de los mismos. En la última parte de este artículo forzaremos el algoritmo a estimar 15 temas y con ello será más sencilla la identificación de estos temas[6].

[5] Véase Hansen, McMahon y Prat (2017) para una descripción precisa del cálculo de la probabilidad total.

[6] En Conde-Ruiz *et al.* (2022) ampliamos la muestra original para incluir los resúmenes de 1.117 artículos publicados como *Papers* y *Proceeding* en AER, entre 2011 y 2018 (antes de 2011 este tipo de trabajos no tienen resúmenes y después de 2018 se publican en otra revista). Con esta nueva muestra ampliada el número óptimo de temas aumenta hasta $K = 70$. Aunque estos nuevos artículos incorporados son muy cortos y con procesos editoriales muy diferentes a los de los envíos regulares, esta muestra ampliada genera interesantes resultados adicionales a los presentados en este capítulo.

Una vez estimados los temas y asignados los resúmenes a ellos, para analizar cómo esta distribución depende del género, debemos determinar el género de los autores, dado que no observamos directamente el género en nuestros datos. Para resolver este problema, clasificamos a los autores por género según su nombre de pila. Nos basamos en tres bases de datos diferentes: la base de datos de nombres de pila publicada por la Administración de la Seguridad Social de EE.UU., creada a partir de los datos de las solicitudes de tarjetas de la Seguridad Social; la base de datos construida por Tang *et al.* (2011), que utiliza Facebook para recopilar datos sobre los nombres de pila y el género autodeclarado; y, por útimo, la base de datos de nombres desarrollada por Bagues (2017). Comprobamos manualmente cualquier candidato que (a) esté dentro del intervalo de probabilidad [0,05 0,95] de ser hombre/mujer o (b) no se encuentre en ninguna de las bases de datos. Una vez hecho esto, convertimos la muestra original de artículos en una muestra de artículos-autores. Transformamos los 5.311 artículos originales en una muestra total de 11.721 (con 9.840 artículos-autores masculinos y 1.881 artículos-autoras). Salvo que se indique lo contrario, todas las medidas que figuran a continuación se calculan sobre esta muestra aumentada de artículos-autores.

4. DIFERENCIAS DE GÉNERO EN LOS TEMAS DE INVESTIGACIÓN EN ECONOMÍA

Siguiendo Conde-Ruiz *et al.* (2022) empezamos presentando los 54 temas estimados por el algoritmo STM en la figura 1 que muestra cada tema asociándolo a sus palabras clave. A pesar de que los temas estimados recogen además del objeto de la investigación, potencialmente la metodología o el estilo, podemos identificar algunos temas por la prevalencia de sus palabras clave. Por ejemplo, el tema 28 parece estar relacionado con el Comercio Internacional, mientras que el tema 9 puede asociarse a la Teoría Econométrica.

4.1. Prevalencia de los temas

Una vez estimados los temas de investigación, el siguiente paso es asignar a los mismos los documentos. En particular, asociamos un resumen d a diferentes temas de acuerdo con la distribución subyacente θ_d. La figura 2 muestra los temas latentes estimados, donde el tamaño del círculo es proporcional al número esperado de documentos en el tema (también hemos reproducido numéricamente esta información en una columna de la figura 1).

La figura 2 también contiene información sobre la conectividad entre los temas. Por ejemplo, si el tema latente k está más cerca de k' que de k'', significa que la distribución β_k se parece más a la distribución $\beta_{k'}$ que a la distribución $\beta_{k''}$. Observando la figura 1 y la descripción de los temas latentes en la figura 2, surgen algunos patrones interesantes. Por ejemplo, los temas 11, 9 y 21 ("Teoría Econométrica"), que ya se han discutido anteriormente, están en cierto modo aislados del resto de temas. En la figura 2 también podemos identificar algunos otros clusters de temas, por ejemplo (al este en la figura 2) 51, 34, 23, 2, etc., son temas relacionados con macrofinanzas, más cercanos a los de Teoría Econométrica, pero no tanto; (oeste en la figura 2) 50 es un nodo central de un conjunto de temas relacionados con econo-

Figura 1.

Los K temas óptimos ordenados por su prevalencia en el *corpus*

Topic 28	trade	countri	product	export	intern	import	firm	sector	factor	develop	3.8%	17.8%
Topic 9	estim	method	sampl	data	asymptot	paramet	consist	use	error	bias	3.5%	13.4%
Topic 11	condit	variabl	function	identif	identifi	restrict	estim	distributinstrument		bound	3.3%	15.5%
Topic 29	experi	subjectexperimenbehaviortreatment				predict	learn	evid	theori	differ	2.8%	17.5%
Topic 22	prefer	choic	decis	util	individu	make	altern	behavior	set	maker	2.7%	14.1%
Topic 21	test	statist	asymptot	distribut	method	paramet	confid	propos	forecast	bootstrap	2.7%	15%
Topic 19	school	student	effect	educ	colleg	score	test	teacher	program	assign	2.6%	17.8%
Topic 48	wage	worker	employ	firm	product	job	increas	labor	plant	skill	2.6%	15.4%
Topic 37	equilibrium	dynam	general	equilibria	exist	economi	condit	stochast	solut	uniqu	2.5%	10.8%
Topic 51	shock	polici	monetari	inflat	aggreg	respons	money	real	nomin	volatil	2.4%	13.2%
Topic 16	belief	agent	expect	prior	ration	probabl	signal	util	set	learn	2.3%	10.1%
Topic 6	game	player	strategi	payoff	equilibrium	play	bargain	repeat	cooper	equilibria	2.3%	10.4%
Topic 2	price	cost	adjust	chang	data	firm	demand	good	markup	relat	2.3%	17%
Topic 49	women	children	parent	femal	men	famili	educ	marriag	child	birth	2.2%	32.8%
Topic 53	market	inform	trade	price	asset	valu	trader	privat	advers	select	2.2%	15.1%
Topic 15	welfar	cost	benefit	insur	gain	polici	estim	loss	reduc	use	2.2%	18.7%
Topic 33	return	firm	stock	manag	asset	equiti	investor	portfolio	predict	size	2.2%	18.4%
Topic 32	contract	agent	princip	commit	optim	hazard	incent	moral	inform	problem	2.1%	11.6%
Topic 50	polici	polit	govern	parti	elect	voter	power	politician	elector	public	2%	13.5%
Topic 34	financi	invest	constraint	recess	shock	asset	firm	aggreg	credit	financ	2%	15.6%
Topic 3	risk	avers	consumpt	ambigu	util	discount	prefer	expect	asset	intertempor	2%	14.5%
Topic 47	consum	firm	product	demand	market	good	price	profit	advertis	competit	1.9%	15.1%
Topic 41	percent	health	insur	increas	hospit	estim	care	patient	drug	use	1.9%	22%
Topic 18	region	econom	area	local	growth	land	agricultur	locat	develop	data	1.9%	14.4%
Topic 43	household	hous	consumpt	spend	incom	expenditur	increas	effect	respons	data	1.8%	15.1%
Topic 45	cycl	busi	product	industri	fluctuat	chang	demand	volatil	aggreg	entri	1.8%	14.4%
Topic 40	optim	alloc	effici	distort	economi	privat	condit	ineffici	resourc	polici	1.8%	14.1%
Topic 27	incom	earn	inequ	data	differ	measur	survey	distribut	use	mobil	1.7%	17.4%
Topic 52	capit	human	invest	skill	growth	accumul	differ	labor	account	life	1.7%	14.5%
Topic 26	market	match	stabl	friction	competit	labor	agent	labour	side	type	1.7%	15.6%
Topic 25	technolog	innov	product	new	firm	patent	research	adopt	knowledg	spillov	1.7%	19.5%
Topic 44	inform	coordin	action	communic	strateg	payoff	game	outcom	sender	signal	1.6%	14.9%
Topic 10	mechani	implement	incent	transfer	type	design	compat	post	agent	problem	1.6%	10.9%
Topic 5	auction	bid	bidder	buyer	seller	valu	price	revenu	privat	inform	1.6%	14.8%
Topic 4	state	unit	right	issu	econom	protect	problem	institut	properti	resourc	1.5%	15.3%
Topic 12	social	network	individu	incent	interact	opportun	depend	connect	link	secur	1.5%	19.9%
Topic 17	bank	credit	polici	fund	crisi	lend	liquid	loan	financi	market	1.5%	14.5%
Topic 42	public	regul	enforc	good	privat	law	provis	punish	legal	cost	1.5%	18%
Topic 13	work	program	labor	suppli	hour	increas	transfer	time	particip	home	1.4%	18%
Topic 20	tax	reform	incom	rate	increas	taxat	margin	chang	optim	effect	1.4%	16.9%
Topic 23	debt	default	borrow	govern	credit	bond	fiscal	sovereign	market	matur	1.4%	16.8%
Topic 1	econom	studi	name	correct	bias	black	measur	data	signific	racial	1.3%	18.7%
Topic 30	firm	contractownership		vertic	integr	adopt	industri	cost	supplier	exclus	1.3%	21.8%
Topic 38	group	ethnic	member	trust	evid	segreg	countri	increas	cultur	chang	1.3%	19.8%
Topic 36	inform	vote	signal	voter	aggreg	bias	privat	strateg	elect	larg	1.3%	15%
Topic 39	rate	exchang	interest	currenc	countri	real	patient	donor	regim	transplant	1.2%	13.6%
Topic 31	save	citi	retir	account	popul	life	increas	german	individu	rate	1.2%	19.4%
Topic 7	vote	news	voter	media	candid	elect	estim	committe	newspap	bias	1.2%	17.5%
Topic 8	search	unemploy	worker	job	distribut	durat	wage	rate	employ	benefit	1.2%	14.7%
Topic 35	conflict	increas	violenc	crime	war	polic	outsid	option	effect	attack	1.1%	17.1%
Topic 14	rule	demand	set	ration	problem	yield	constitut	optim	function	util	1%	10.5%
Topic 46	project	effort	team	perform	redistribut	outcom	win	competit	one	prize	0.9%	19.4%
Topic 24	qualiti	delay	probabl	accept	fee	order	card	offer	paper	higher	0.8%	14.8%
Topic 54	import	use	addit	data	sever	relat	support	analys	find	limit	0.3%	15.7%
											Topic Prop.	Female Prop.

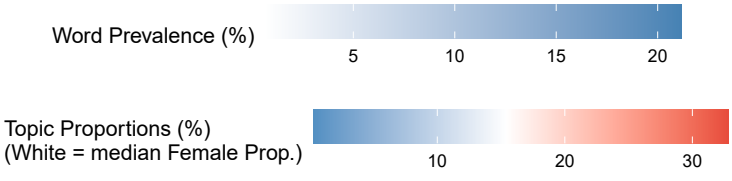
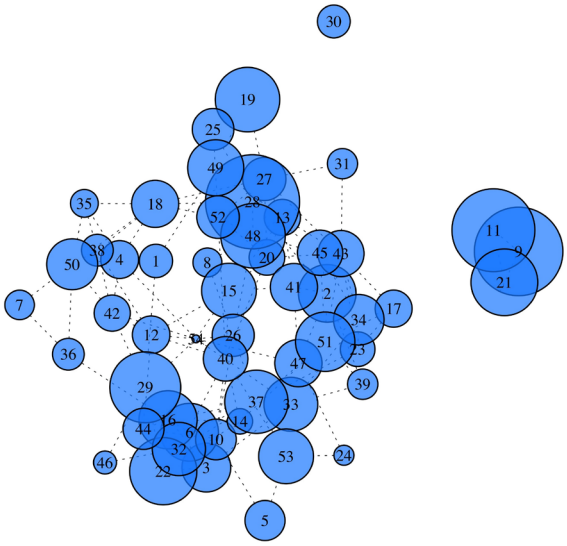


Figura 2.

Conectividad entre los temas y la proporción de documentos en cada tema (la distribución θ_d)



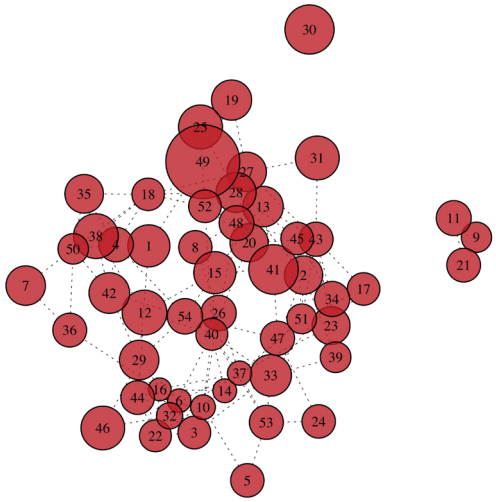
mía política e instituciones), (suroeste en la figura 2) 29, 32, 2 2, etc., son temas relacionados con la microeconomía (teoría de contratos, teorema de la decisión, etc.). Por último, las áreas aplicadas como la economía del trabajo, el desarrollo internacional o la economía política se sitúan en torno a los temas 19, 49, 28 y 48 (norte en la figura 2). En Conde-Ruiz *et al.* (2022), realizamos un análisis más formal de la distancia entre temas utilizando un análisis de correspondencia simple.

La figura 3 repite el análisis pero el tamaño del círculo representa esta vez la proporción de autoras mujeres en el tema. Los tamaños relativos de los temas han cambiado dado que encontramos que la distribución por género es diferente entre las diferentes áreas. La figura 3 es la primera evidencia de que existen diferencias horizontales entre hombres y mujeres en las publicaciones, y por lo tanto en los temas de investigación. Esto se ilustra bien, con un pequeño subconjunto de temas (norte en la figura 3) aparentemente relacionados con Economía Aplicada, especialmente el tema 49, con una proporción relativamente alta de mujeres, y por el contrario otro conjunto de temas (por ejemplo, el sudoeste en la figura 3) que están estrechamente conectados y en los que la presencia de mujeres es escasa (en esos temas aparecen términos frecuentemente utilizados en la Teoría Económica).

Para ilustrar esta intuición vamos a analizar con mayor detalle, los temas donde la proporción de autoras es mayor, el tema 49 (32,8 %), y el tema 16 donde el porcentaje de autoras sólo llega al 10,1 %. Para ello, la figura 4 representa estos dos temas como nubes de palabras, donde el tamaño de los términos en la nube es equivalente a su probabilidad en la distribu-

Figura 3.

Conectividad entre los temas y la proporción de autores mujeres en cada tema



ción de temas latentes β_k. Las palabras que parecen más destacadas en la nube 49 son mujer, hombre, padre, hijo, salud, etc. Estas palabras podrían relacionarse fácilmente con campos de investigación, como la Economía de Género o la Economía de la Salud, tradicionalmente asociados a las mujeres. Del mismo modo, la nube de palabras del tema 16 parece estar relacionada con teoría Microeconómica.

Figura 4.

Nubes de palabras por tema: tema 49 vs. tema 16

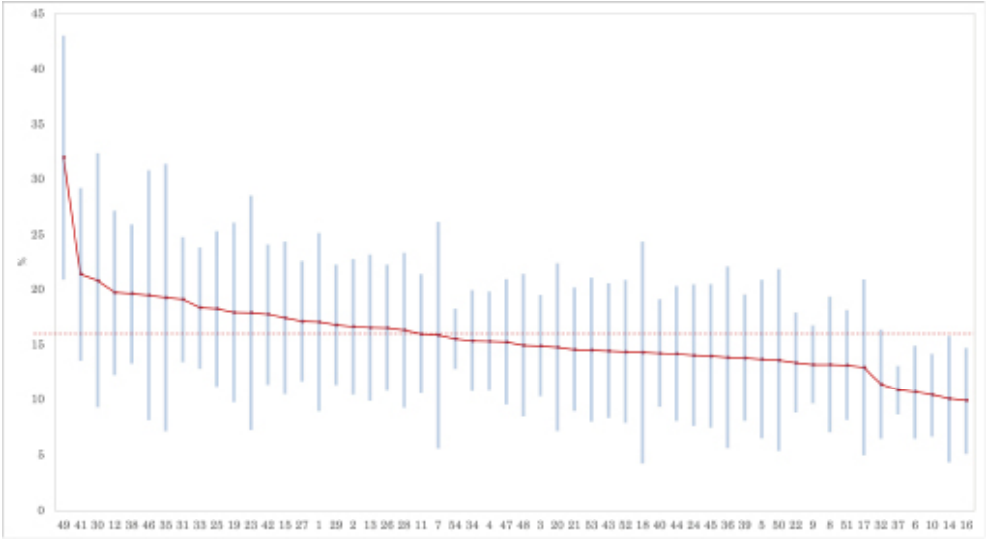
(a) Topic 49 (El tema con una mayor proporción de autores mujeres)

(b) Topic 16 (El tema con una mayor proporción de autores mujeres)

Conde-Ruiz *et al.* (2022) ilustra las diferencias horizontales en temas de investigación con dos figuras muy informativas. La figura 5 muestra la media de la presencia de autoras por tema ordenadas de mayor a menor (la media es del 15,9 % en el periodo 2002-2019), junto con la desviación estándar de esta presencia a lo largo de la muestra de años. Esta figura muestra claramente la prevalencia relativa de las mujeres con respecto a los hombres en algunos temas, mientras que en otros, la representación es muy pequeña.

Figura 5.

Sobre la prevalencia de autores mujeres por tema: media y una desviación estándar a lo largo del tiempo

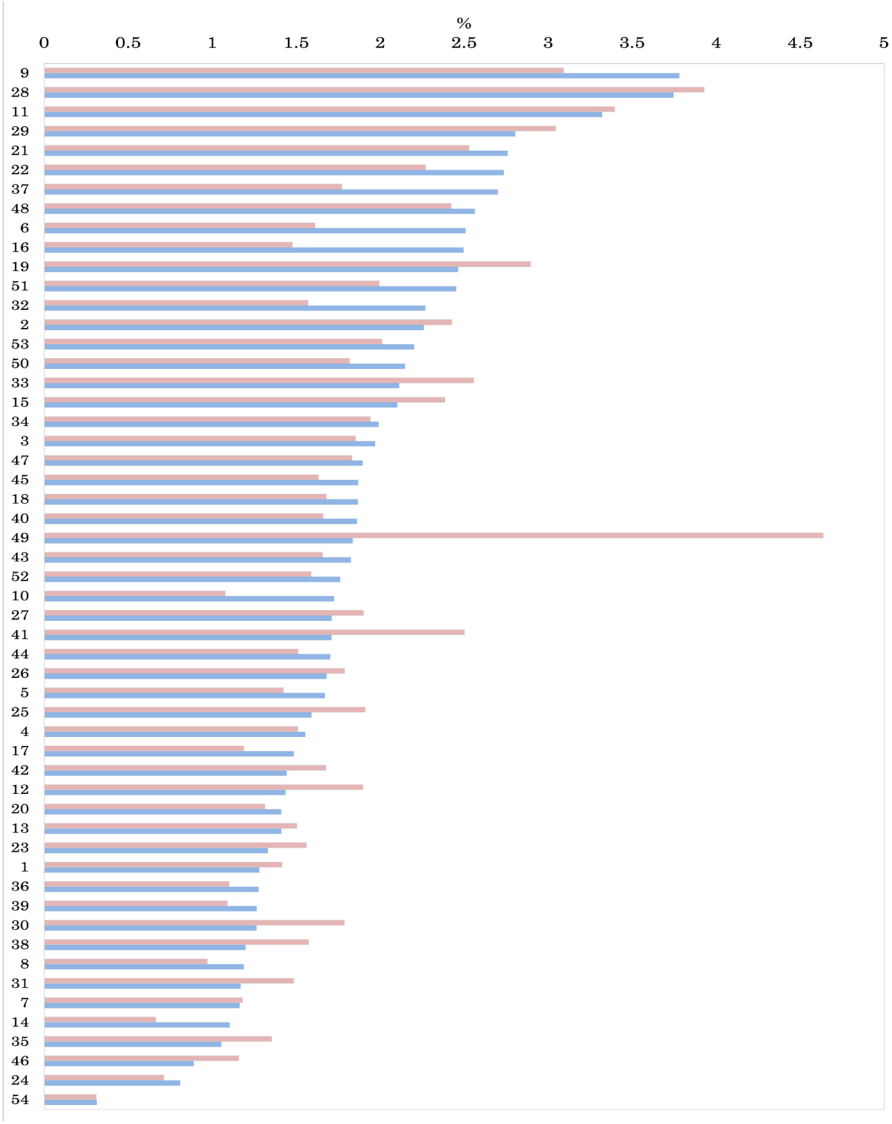


No obstante, para tener una imagen más precisa de estas diferencias "horizontales", debemos añadir la información relativa a la prevalencia de los temas. Es posible que las mujeres no estén representadas en un tema concreto y que esta circunstancia tenga poco impacto en la medida en que este tema contenga muy pocos trabajos publicados. La figura 6 muestra la distribución entre hombres y mujeres en los temas normalizados por tener el mismo tamaño. Esto nos da la probabilidad de que, por ejemplo, un artículo de autoría femenina pertenezca a cualquiera de los 54 temas. Clasificamos los temas según la probabilidad de que los elija un autor masculino. Esta figura demuestra que los autores masculinos y femeninos tienen preferencias diferentes o siguen estrategias distintas a la hora de realizar y publicar sus investigaciones. Observamos que los temas con mayor "demanda" por parte de los hombres también son muy solicitados por las mujeres. Sin embargo, hay un conjunto de temas, para los que la proporción de artículos publicados por los hombres es alta, que son menos atractivos (o más difíciles de publicar) para las mujeres. En general, las distribuciones masculina y femenina son diferentes, con la característica sobresaliente del tema 49 para las mujeres, que es un claro pico en la distribución femenina de los trabajos publicados.

El análisis de los textos sobre todas las publicaciones en T5 realizadas entre 2002 y 2019 provee una evidencia clara de que entre los hombres y las mujeres existen diferencias horizontales respecto a la elección de temas de investigación.

Figura 6.

Distribuciones empíricas por temas entre hombres y mujeres (con la condición de haber publicado un artículo en el *Top 5)*

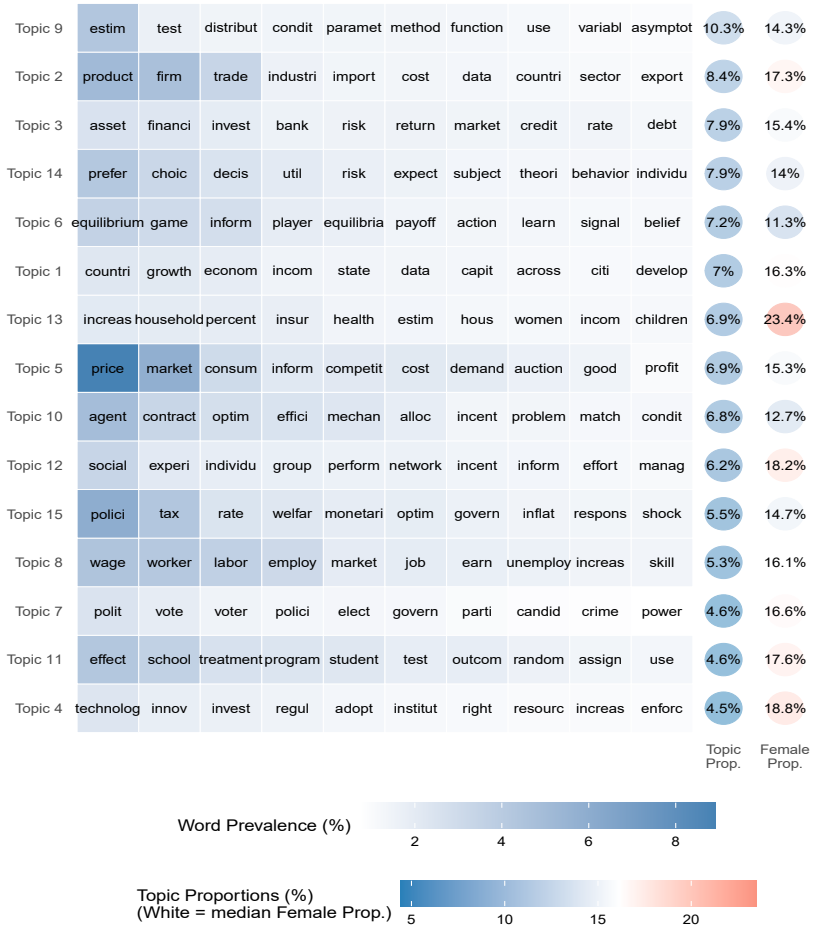


5. TEMAS ESTIMADOS COMO ÁREAS DE INVESTIGACIÓN: FINANZAS, MAYORITARIAMENTE COSA DE HOMBRES

En esta sección vamos a reducir el número de temas hasta el punto de que podemos identificar cada tema con un campo de investigación. Hemos visto como cuando el número de temas es alto, no sólo identificamos el área de investigación sino también aspectos semánticos, o forma de enfocar la investigación. Al reducir *K* o el número de temas, por un lado, facilita la interpretación semántica de los temas y luego permite analizar, por ejemplo, si el peso de un campo concreto en el *Top 5* ha aumentado con el tiempo. Por último, un número bajo de temas nos permitirá enmarcar nuestros resultados con la literatura anterior que ha utilizado un número reducido de categorías vinculadas a los códigos JEL y las áreas de investigación de los principales departamentos.

Figura 7.

Temas latentes ordenados por prevalencia en el *corpus* con *K* = *15*



Analizando de forma manual distintos números de temas, encontramos que para $k = 15$ el modelo estimado funciona mejor en términos de ajuste con los datos, y en términos de coincidencia del *topic* latente y el campo de investigación. El modelo con $k = 15$ temas latentes se resume en la figura 7, donde se puede ver que el tema 3, coincide plenamente con Finanzas. Vemos como en el área de Finanzas se han publicado el 7,8 % de todos los artículos aparecidos en las revistas *Top 5* y como tan sólo el 15,4 % de los autores que han publicado en esta área son mujeres.

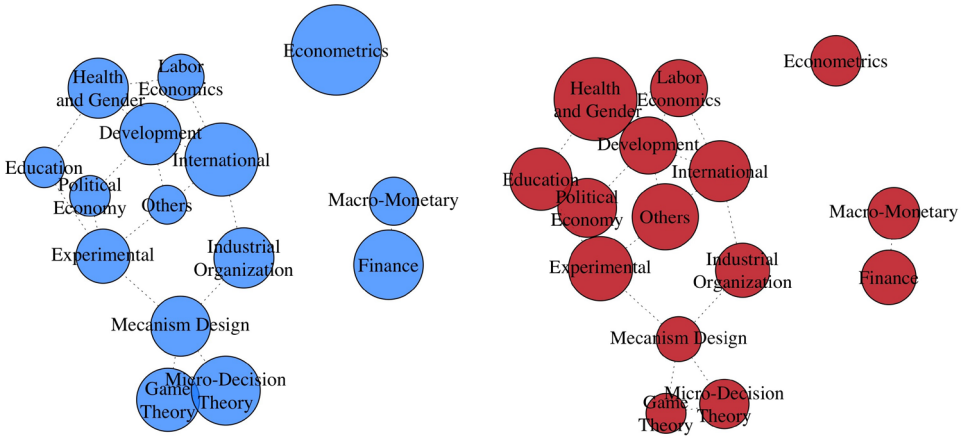
Cuando hacemos un análisis de las distancias entre temas o áreas de investigación vemos como "Finanzas" está bastante aislado del resto de temas y tan sólo tiene conexión con "Macroeconomía Financiera", mientras que hay otros campos que están mucho más conectados. En la figura 8 vemos también como esta geografía o distancias entre áreas de investigación es similar entre hombres y mujeres, aunque en la distribución entre áreas por género (*i.e.* el tamaño de los círculos), las mujeres trabajan en términos proporcionales menos que los hombres en el área de "Finanzas". También, vemos cómo los artículos publicados por mujeres están sobrerrepresentados en el área de "Economía de la Salud y Género".

Figura 8.

Conectividad entre los temas para K = 15

(a) Conexión entre temas y fracción de documentos/ resúmenes en cada *topic* (o área de investigación) (θ_d distribution)

(b) Conexión entre los temas y los documentos/ resúmenes de autoras en cada *topic* (o área de investigación)

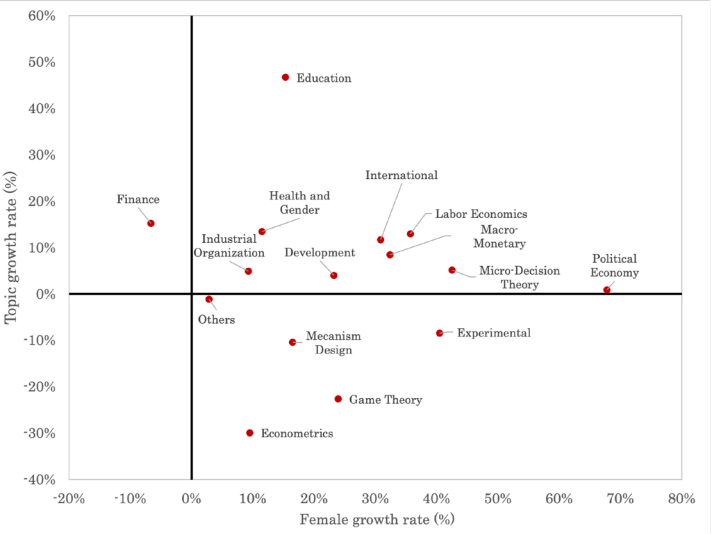


La figura 9 analiza conjuntamente la evolución de la prevalencia de los temas (o campos de investigación) y la proporción de mujeres autoras. Para construir la figura, hemos calculado la tasa de crecimiento de la prevalencia de los temas y la proporción de autoras utilizando los últimos siete años (2013-2019) y los primeros siete años (2002-2008). En primer lugar, podemos observar que la proporción de mujeres ha aumentado en todos los temas excepto en Finanzas que ha caído un 6,6 %. Este dato es sorprendente, pues a pesar del aumento de

las mujeres académicas y que las publicaciones en el área de Finanzas en revistas *Top 5* han aumentado un 18 %, la presencia de mujeres autoras en ese campo sea el único en el cual han disminuido.

Figura 9.

Tasas de crecimiento de la prevalencia del área de investigación y de la proporción de mujeres en cada área



En cuanto a la prevalencia, sólo cuatro áreas de investigación han disminuido su peso en términos artículos publicados en las revistas *Top 5*: i) diseño de mecanismos (-10,3 %); ii) econometría (-29 %); iii) teoría de juegos (-22,5 %), y iv) economía experimental (-8,4 %). Salvo en finanzas, en todas las demás áreas de investigación el porcentaje de mujeres autoras han subido. Las mayores subidas se han producido en economía política (+67,7 %), teoría de la decisión (+42,5 %), macroeconomía y monetaria (+32,3 %), economía experimental (+40 %) o economía laboral (+35 %). En todos ellos las mujeres estaban claramente infrarrepresentadas.

Por otro lado, los temas en los que menos ha crecido el porcentaje de mujeres, además de finanzas, que ha disminuido, han sido en economía de la salud y género que ha aumentado un +11,4 %, en econometría con un aumento del +9,4 %, y en organización industrial (IO) que ha subido un +9,2.

Por último, no parece existir una relación clara entre la tasa de crecimiento de la prevalencia de temas y el aumento de la representación femenina. Esto creemos que es sorprendente. Sería esperable que aquellas áreas cuyas publicaciones ganan peso en las mejores revistas del

mundo, sean también en las que la presencia de mujeres autoras aumente. No disponemos de datos sobre la edad de los autores, pero como la proporción de mujeres va en aumento, cabe esperar que la proporción de mujeres entre los nuevos autores que han publicado en revistas *Top 5* sea relativamente grande. Por lo tanto, es más probable que los recién llegados trabajen en temas "candentes" que en temas en declive. La combinación de ambos efectos debería conducir a una correlación positiva entre el aumento de la prevalencia de un tema y el aumento de la representación femenina, algo que no observamos claramente en los datos.

6. CONCLUSIONES

En resumen, utilizando técnicas de aprendizaje automático no supervisado sin ninguna intervención en la determinación de los temas de investigación y la asignación de los artículos a los temas, se ha mostrado evidencia empírica de que existen diferencias horizontales entre hombres y mujeres en la elección de temas de investigación. Se ha llevado a cabo un análisis de textos utilizando para ello, todos los artículos publicados en revistas T5 entre los años 2002 y 2019. Se trata de una muestra muy relevante, dado el impacto de las revistas T5 en las carreras profesionales de los investigadores en las mejores universidades del mundo. Este resultado es importante porque los recientes resultados teóricos de Conde-Ruiz y Ganuza (2017); Conde-Ruiz *et al.* (202) y Siniscalchi y Veronesi (2020) muestran que las diferencias "horizontales" de género en la elección del tema de investigación pueden conducir a una discriminación permanente por género en la carreras profesionales académicas de economía y finanzas. Brecha de género que los datos muestran que existe y que al revés que en otros sectores, no ha mejorado significativamente en la última década.

La evidencia empírica de este capítulo esta extraída de Conde-Ruiz *et al.* (2022), artículo que profundiza además en otras dimensiones relacionadas con las diferencias de género en la investigación en economía. Por ejemplo, al escribir un artículo, un autor puede contribuir a un solo tema latente o a varios, los autores que han publicado varios trabajos pueden haber escrito artículos similares o pueden haber sido más diversos: ¿son estos patrones de diversificación diferentes para los hombres y las mujeres? Conde-Ruiz *et al.* (2022) analiza esta cuestión utilizando el índice Herfindahl-Hirschman (HHI) que se utiliza para medir la concentración en un mercado, como una medida de dispersión entre los temas. El principal resultado es que las mujeres son más diversas (HHI más bajo) cuando publican uno o dos artículos, pero menos (HHI más alto) cuando publican un mayor número de artículos en el *Top 5*. Otra perspectiva es la de Hengel (2020) que utiliza algoritmos de legibilidad para medir la calidad de la escritura de los resúmenes de los artículos[7]. Conde-Ruiz *et al.* (2022) aplican estos algoritmos de calidad de la escritura a su base de artículos T5 y concluyen que los temas más femeninos están mejor escritos que las de los temas más masculinos. Sin embargo, es difícil desentrañar el papel de la prevalencia de autoras frente a la redacción dentro de un tema.

[7] Como Hengel (2020) discute en detalle, la legibilidad de los resúmenes está fuertemente correlacionada positivamente con la legibilidad de otras secciones de un artículo.

Referencias

Arrow, K. J. (1973). The Theory of Discrimination. En O. Ashenfelter y A. Rees (eds), *Discrimination in Labor Markets.* Princeton University Press.

Azmat, G. y Ferrer, R. (2017). Gender Gaps in Performance: Evidence from Young Lawyers. *Journal of Political Economy, 125*(5), pp. 221–242.

Bertrand, M., Goldin, C. y Katz, L. (2010). Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors. *American Economic Journal: Applied Economics*, 2, pp. 228–255.

Bohren, J. A., Imas, A. y Rosenberg, M. (2019). The Dynamics of Discrimination, Theory and Evidence. *American Economic Review*, 109, pp. 3395–3436.

Bordalo, P., Coffman, K. y Gennaioli, N. (2019). Beliefs about Gender. *American Economic Review*, 109, pp. 739–773.

Conde-Ruiz, J. I., Ganuza, J. J., García, M. y Puch, L. A. (2022). Gender Distribution across Topics in Top 5 Economics Journals: A Machine Learning Approach, *SERIEs: Journal of the Spanish Economic Association*, 13, pp. 269–308.

Conde-Ruiz, J. I., Ganuza, P. y Profeta, J- J. (2017). Statistical Discrimination and the Efficiency of Quotas. *Fedea Working Papers*.

Conde-Ruiz, J. I., Ganuza, P. y Profeta, J- J. (2022). Statistical discrimination and committees. *European Economic Review*, 141, 103994.

Hansen, S., McMahon, M. y Prat, A. (2017). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics, 133*(2), pp. 801–870.

Heckman, J. y Sidharth, M. (2020). Publishing and Promotion in Economics: The Tyranny of the Top Five. *Journal of Economic Literature, 58*(2), pp. 419–470.

Hengel, E. (2020). Publishing while Female. Are women held to higher standards? Evidence from peer review. *Cambridge Working Papers in Economics*, 1753, Faculty of Economics, University of Cambridge.

Matsa, D. A. y Miller, A. (2011).Chipping Away at the Glass Ceiling: Gender Spillovers in Corporate Leadership. *American Economic Review, 101*(3), pp. 635–639.

Niederle, M. y Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives, 24*(2), pp. 129–144.

Reuben, E., Sapienza, P. y Zingales, L. (2014). How Stereotypes Impair Women's Careers in Science. *Proceedings of the National Academy of Sciences*, 111, pp. 4403–4408.

Roberts, M. E., Stewart, B. M. y Tingley, D. (2019). STM: An R Package for Structural Topic Models. *Journal of Statistical Software, Articles, 91*(2), pp. 1–40.

Siniscalchi, M. y Veronesi,P. (2020). Self-image Bias and Lost Talent. December 2020, (28308).

Tang, C., Ross, K., Saxena, N. y Chenn, R. (2011). What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook. pp. 344–356.

Tang, C., Ross, K., Saxena, N., Chenn, R., Campa, P. y Bagues, M. (2017). Can Gender Quotas in Candidate Lists Empower Women? *Evidence from a Regression Discontinuity Design*, (12149).

Tang, C., Ross, K., Saxena, N., Chenn, R., Jordan, M., Blei, D. y Ng, A. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, March 2003, 3 (null), pp. 993-1022.