Mapping the Market for Ideas in Europe, 1450–1650:
A Title Embeddings Approach

Noel D. Johnson
(George Mason University)

Alexander Taylor
(University of Evansville)

This version: June 19, 2026

njohnsoL@gmu.edu

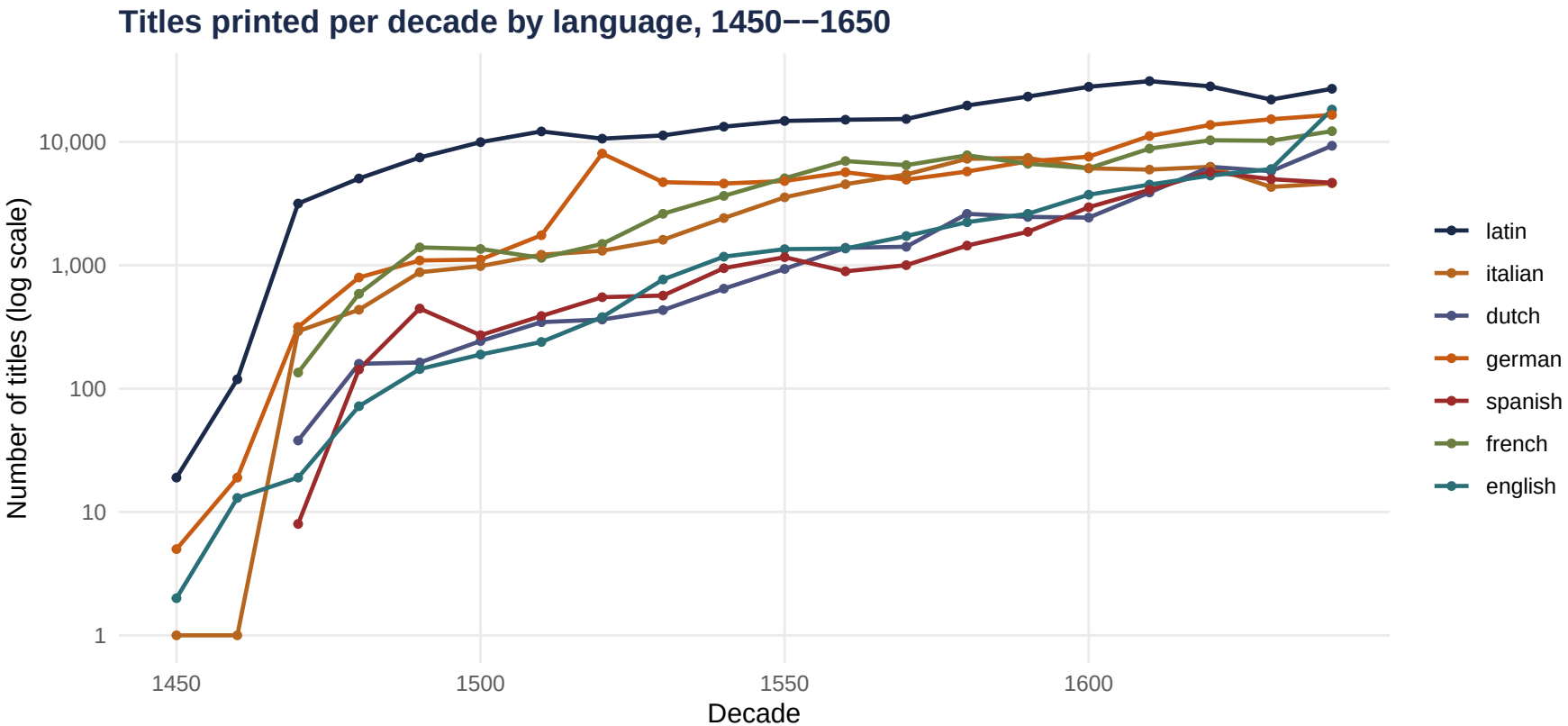
Books contain ideas

TYCHONIS BRAHE
MATHIM: EMINENT: DANI
OPERA OMNIA,
Sive
ASTRONOMIÆ INSTAVRATÆ
PROGYMNASMATA
In duas partes distributa,
QVORVM PRIMA DE RESTITVTIONE MOTVVM
Solis & Lunæ, Stellarumq; inerrantium tractat.
SECVNDA AVTEM DE MVNDI ÆTHEREI
Recentioribus Phænomenis agit.

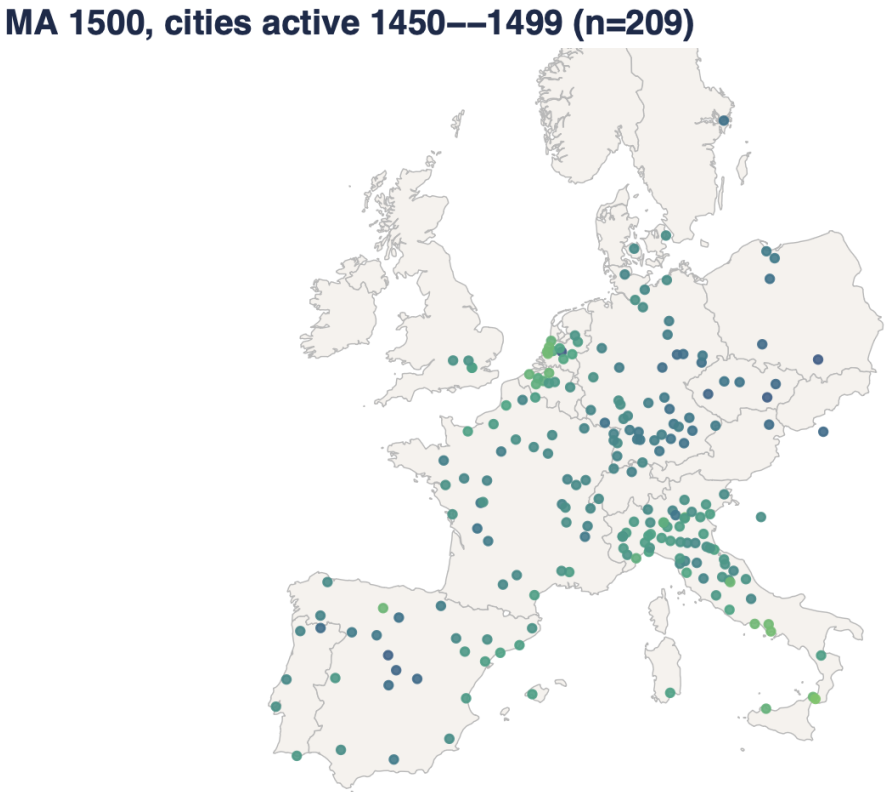
ANNO M.DC.XLVIII.

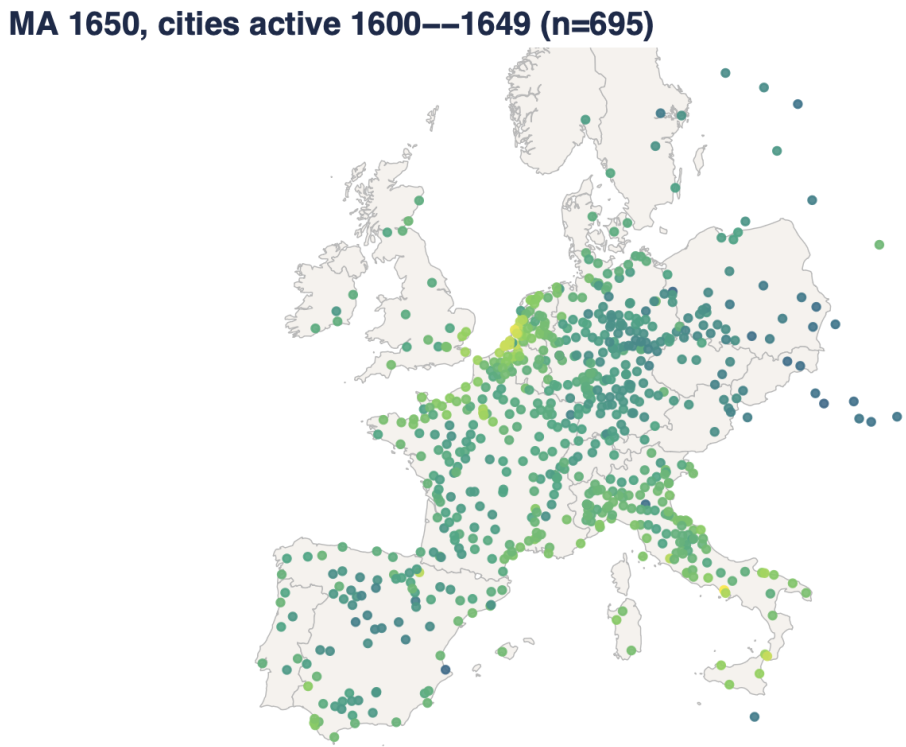
Editio ultima nunc cum Indicibus & Figuris prodit.
FRANCOFVRTI,
Impensis Ioannis Godofredi Schönvvetteri.

Europe printed a lot of books after 1450



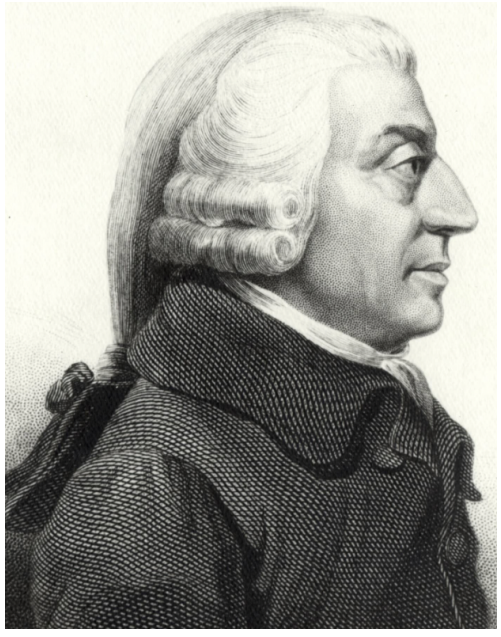
As the press spread, market access also increased unevenly





Antwerp, Amsterdam rise. Bruges declines.

How do increases in market potential affect the production of knowledge?



Adam Smith
1776
"division of labour"

Paul Krugman
1979 / 1980
"monopolistic competition"

Joel Mokyr
2002 / 2009
"useful knowledge"

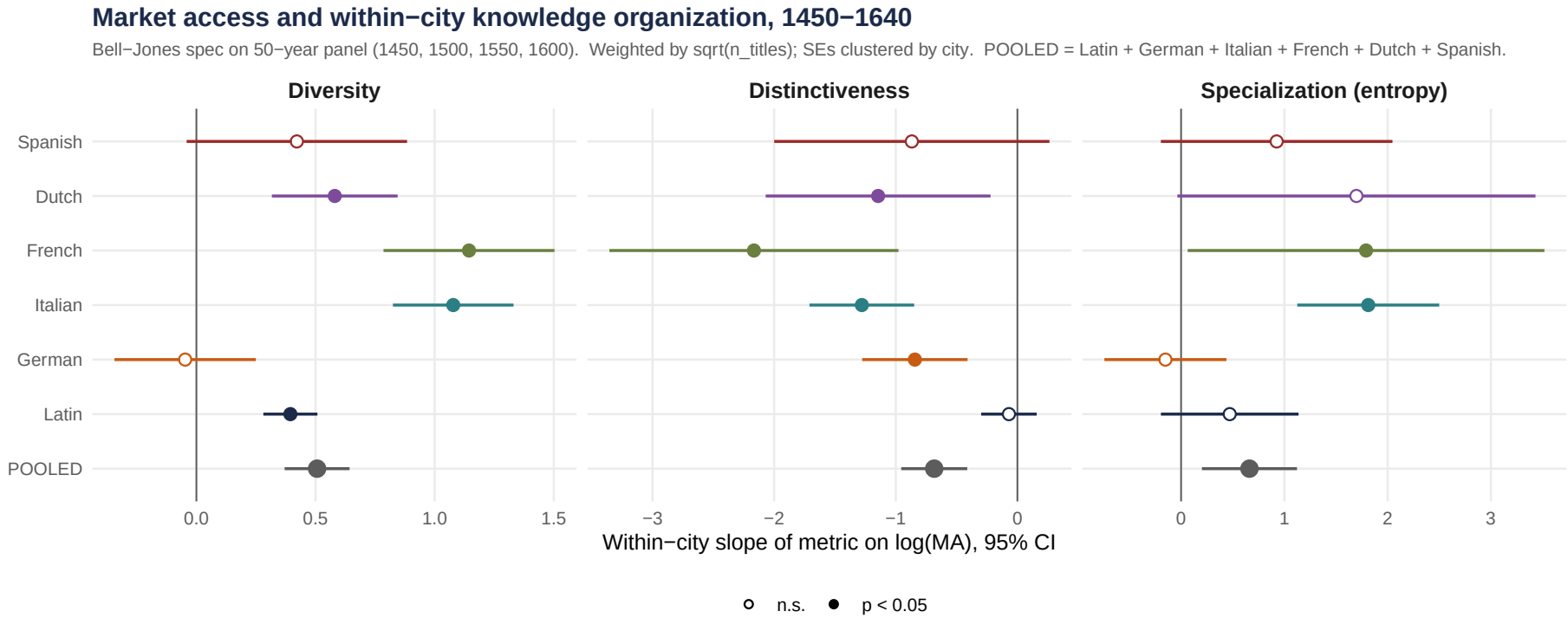
How do we tell these predictions apart?

(i) Measure the variety of printed material
$\rightarrow$ title embeddings

(ii) Measure trade potential
$\rightarrow$ market access (MA)

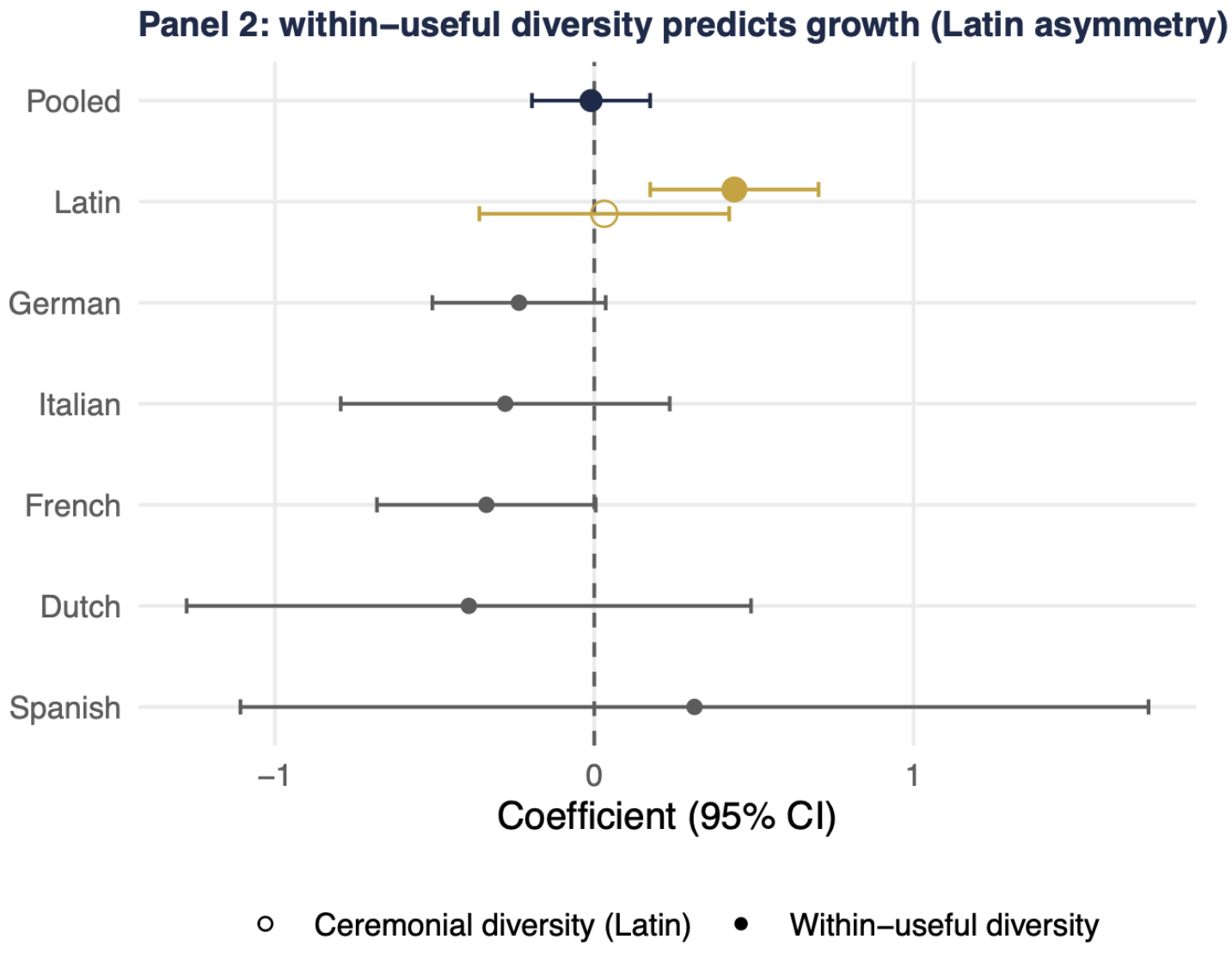
(iii) Link variation in (i) with variation in (ii)
$\rightarrow$ within-city panel exploiting variation in MA

We get results consistent with New Trade Theory (Krugman)



As market access increases: Diversity ↑ · Distinctiveness ↓ · Entropy ↑

We also find support for a Mokyr growth channel



Frankfurt am Main, 1580–1620

Typical Latin titles:

- *Florilegium Magnum seu Polyanthea*
- *Rerum Persicarum Historia*
- *Imperatorum Romanorum Numismatum Series*
- Paracelsus, *Operum Medico-Chimicorum*

- **HIGH MA**
- Homogenised toward the European mainstream
- Population grew sharply

Tübingen, 1550–1599

Typical print:

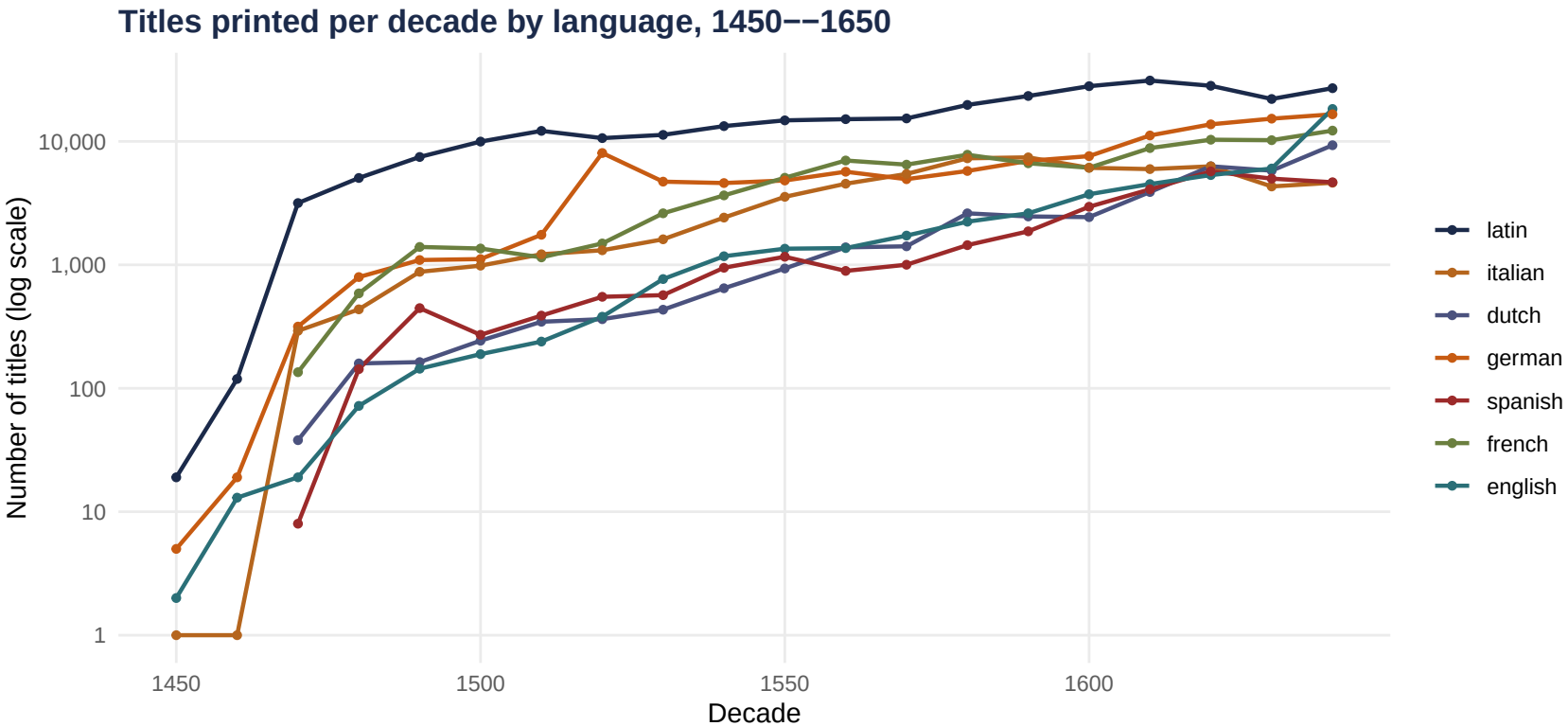
- Lutheran scholastic disputations from the Stift
- Hebrew and biblical philology
- University medical dissertations
- Astronomical-mathematical works (Mästlin, who taught Kepler)

- **LOW MA**
- Distinctive academic-confessional niche
- Population modest

Part II
Data and measurement

Title embeddings, market access, and measuring knowledge production

The USTC corpus: 800K editions across six languages



800,000 editions · 1,300 cities · 1450–1650

We have two types of big data problems

Language	Vocabulary (features)	Titles (documents)	Title length (words) median	mean
Latin	22,362	314,172	17	19.3
German	12,032	140,207	27	25.4
French	5,385	104,771	12	15.4
Italian	5,164	66,655	17	20.3
English	5,712	52,265	34	34.9
Dutch	2,848	44,782	14	15.0
Spanish	3,389	42,448	19	23.9

Title length tokenised on `short_title` via word boundaries. USTC 1450–1650.

Representing words by their context

Distributional hypothesis: words that occur in similar contexts tend to have similar meanings

J.R.Firth 1957

- "You shall know a word by the company it keeps"
- One of the most successful ideas of modern statistical NLP!

…government debt problems turning into	***banking***	*crises as happened in 2009…*
…saying that Europe needs unified	***banking***	*regulation to replace the hodgepodge…*
…India has just given its	***banking***	*system a shot in the arm…*

These context words will represent *banking*.

Distributional hypothesis

"tejuino"

C1: A bottle of ____ is on the table.

C2: Everybody likes ____.

C3: Don't have ____ before you drive.

C4: We make ____ out of corn.

Distributional hypothesis

C1: A bottle of ____ is on the table.

C2: Everybody likes ____.

C3: Don't have ____ before you drive.

C4: We make ____ out of corn.

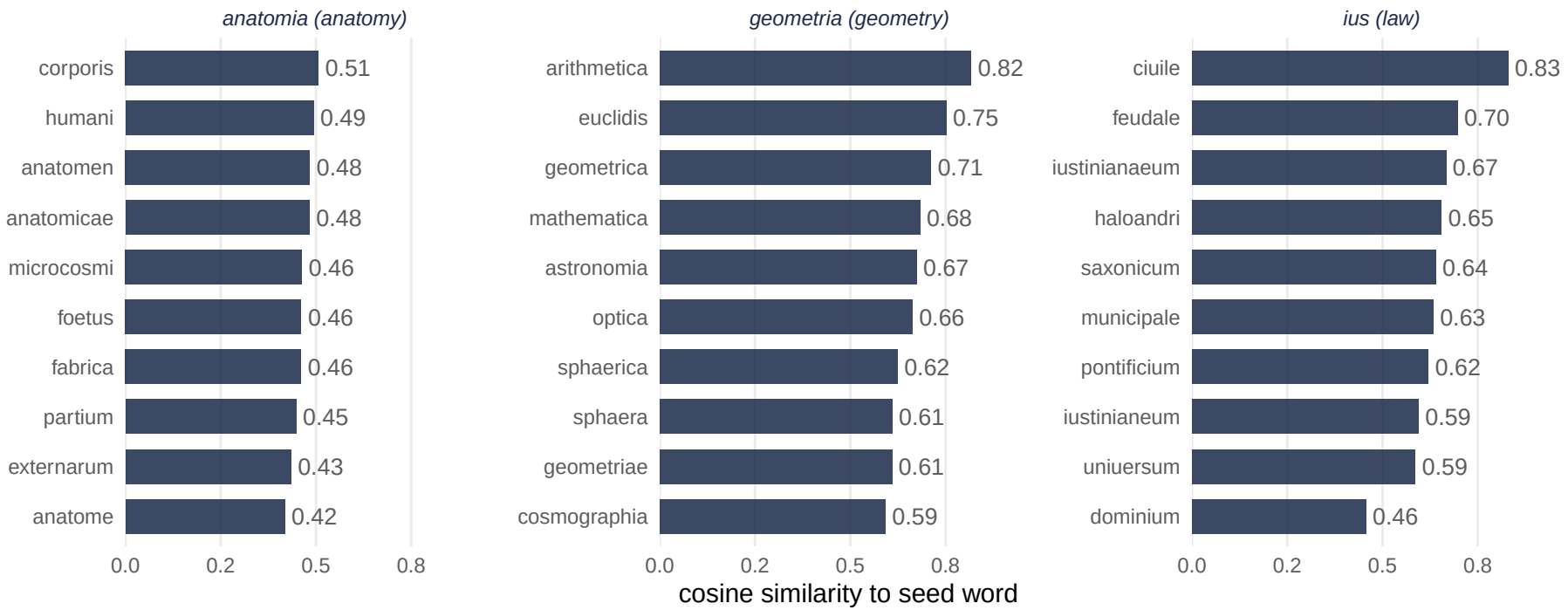
	C1	C2	C3	C4
tejuino	1	1	1	1
loud	0	0	0	0
motor-oil	1	0	0	0
tortillas	0	1	0	1
choices	0	1	0	0
wine	1	1	1	0

"words that occur in similar contexts tend to have similar meanings"

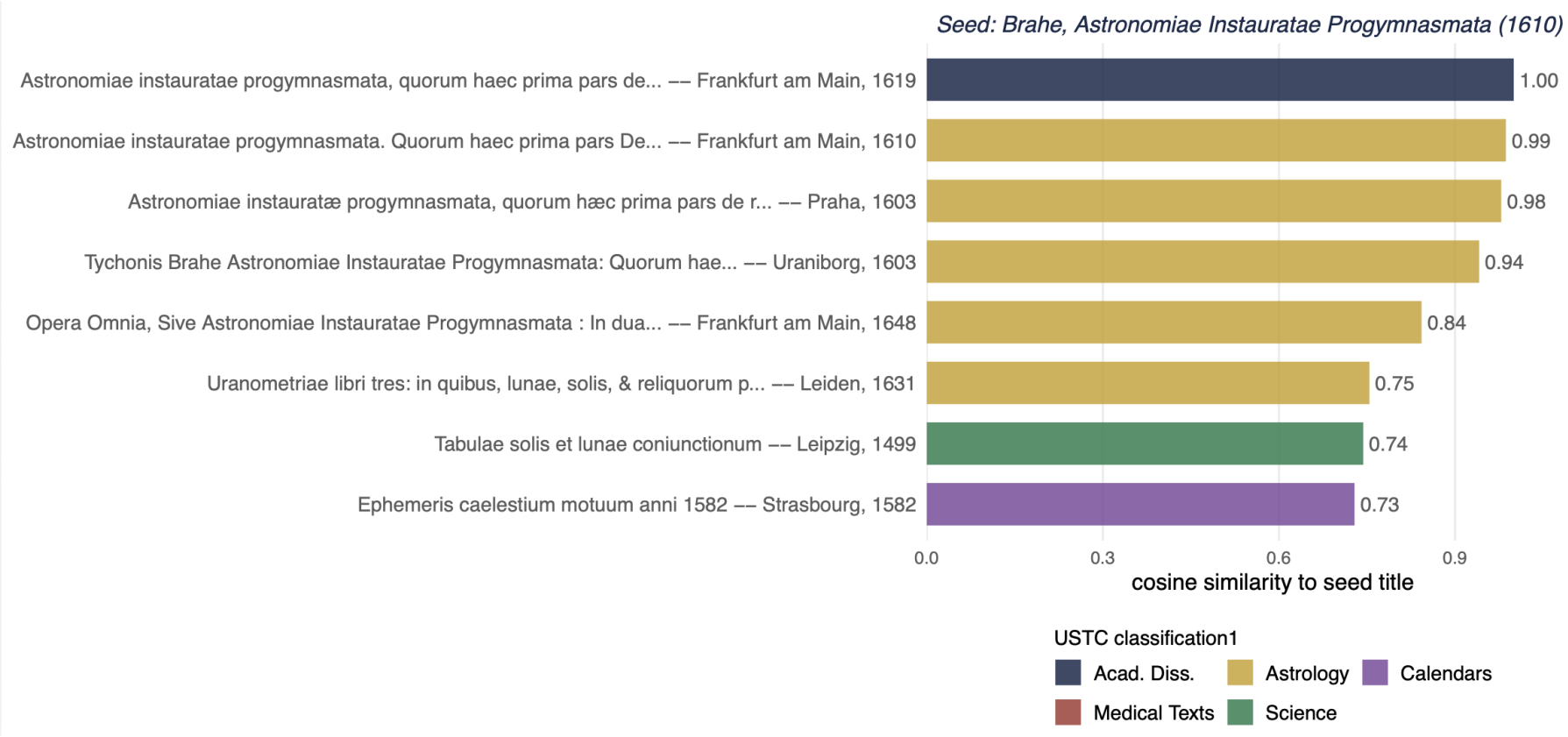
The embedding pipeline, end-to-end

- Five-word skip-gram windows across all USTC titles in a language
- Co-occurrence counts $\rightarrow$ positive pointwise mutual information (PPMI)
- Truncated singular-value decomposition $\rightarrow$ 200 dimensions per word
- SIF-weighted averaging of word vectors $\rightarrow$ title vectors (Arora et al. 2017)
- First principal component removed

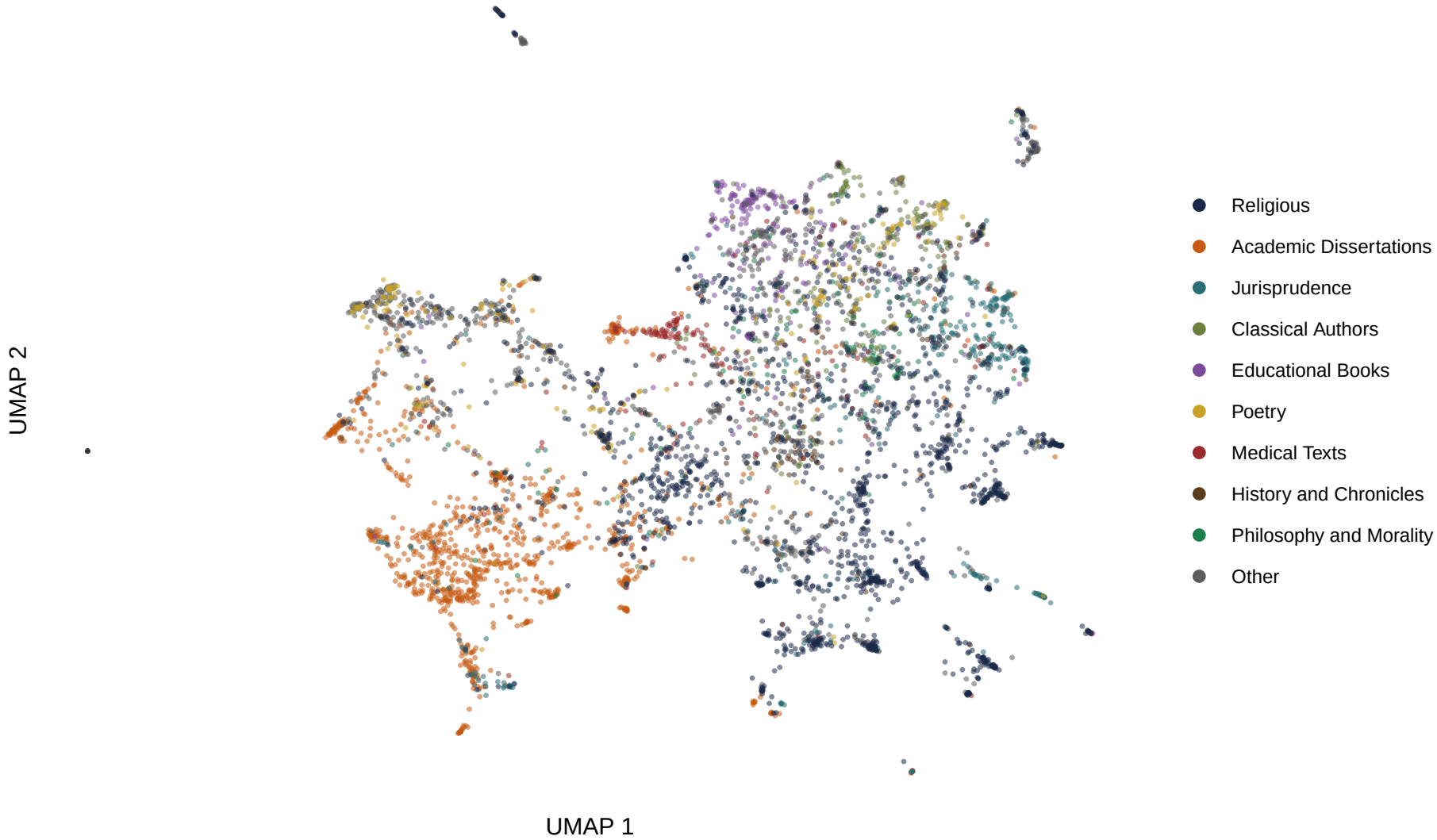
The embedding recovers intellectual structure (Latin)



Title-level nearest neighbours



Embedding space recovers intellectual structure



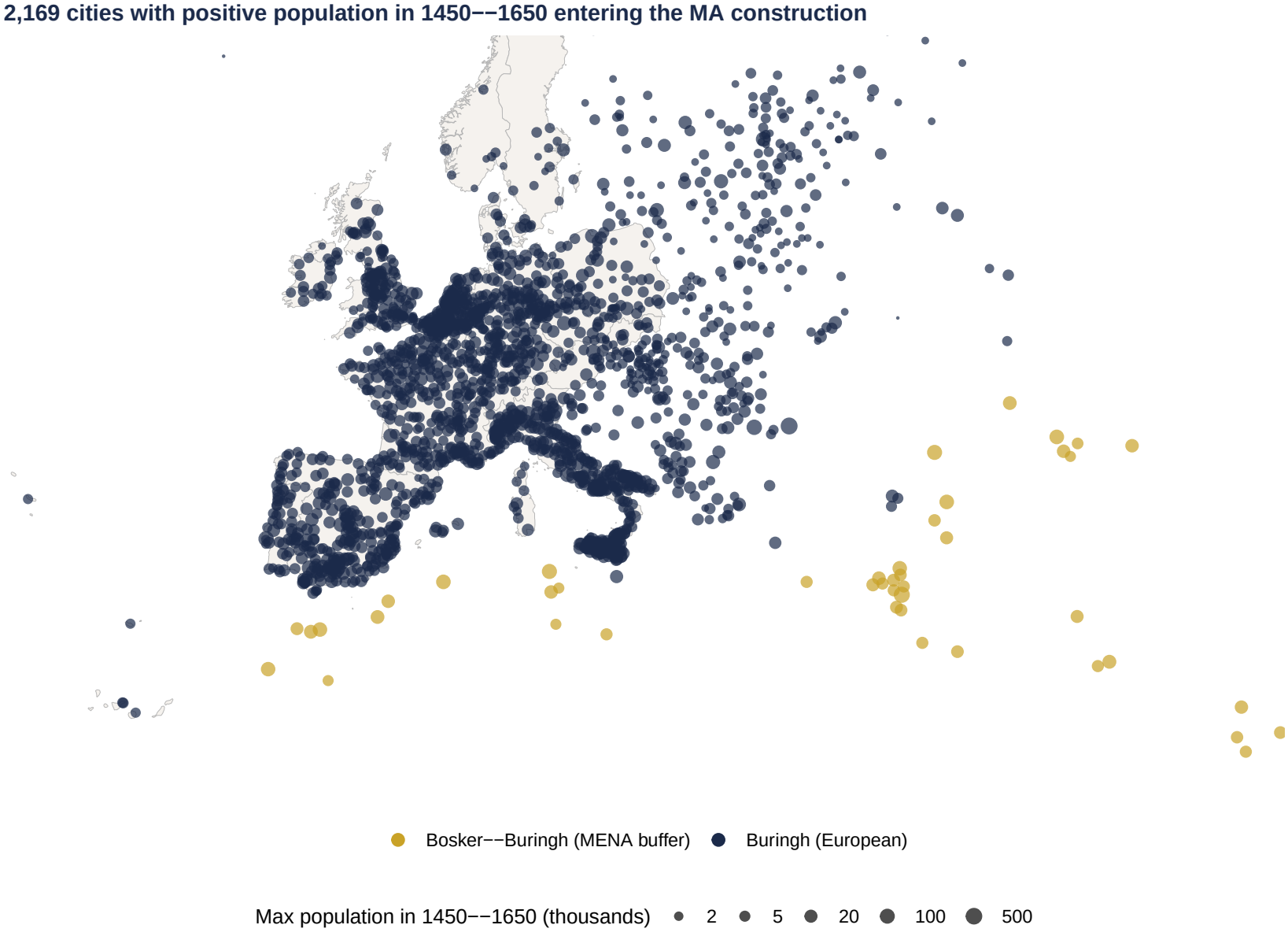
Market access: a sufficient statistic for gains from trade

- In modern quantitative trade theory (Eaton & Kortum 2002; Arkolakis, Costinot & Rodriguez-Clare 2012; Allen & Arkolakis 2014; Donaldson & Hornbeck 2016), market access summarises a city's gains from trade as

$$\mathrm{MA}_i = \sum_{j \neq i} \mathrm{pop}_j \, \tau_{ij}^{-\theta}$$

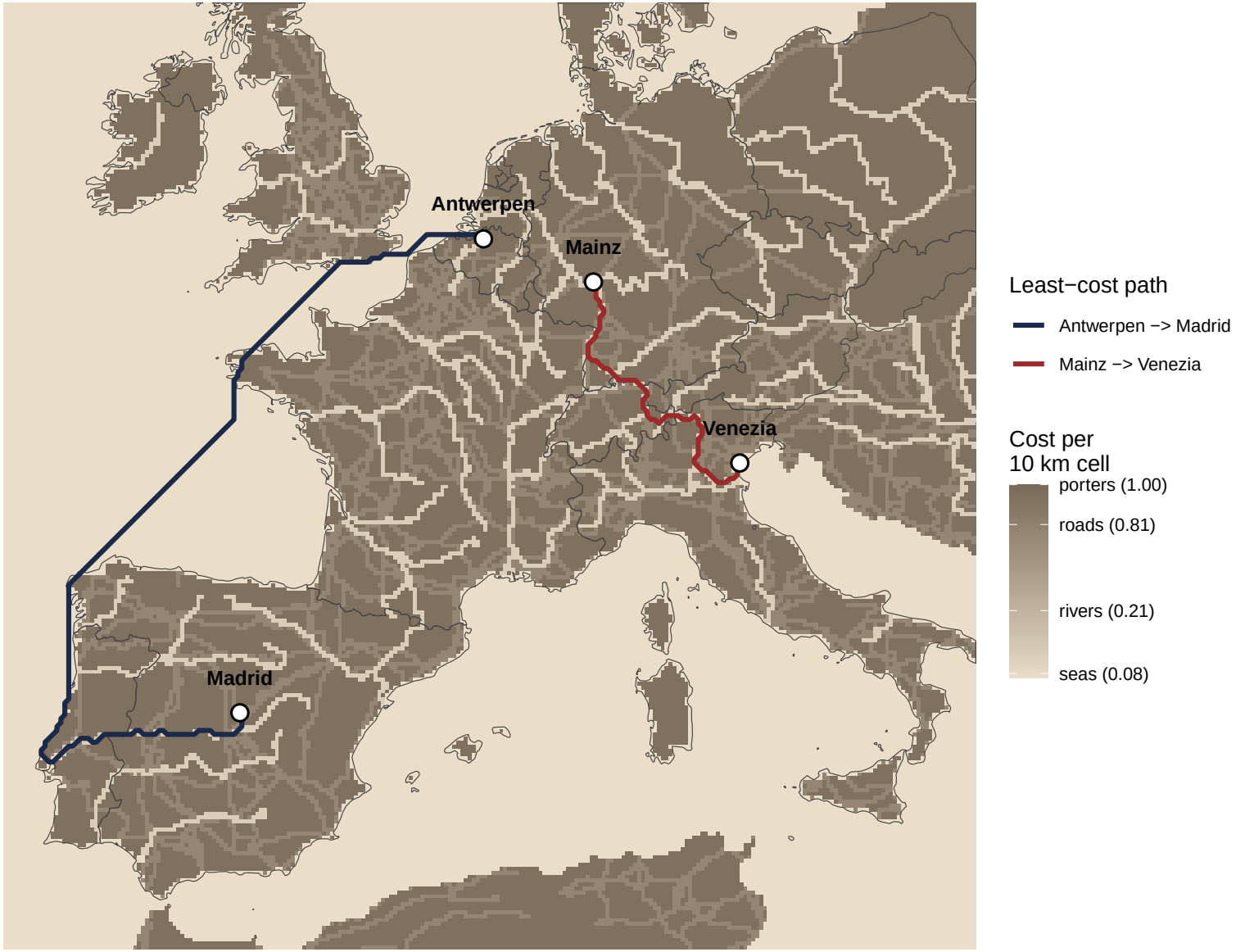
- The hard part is measuring bilateral transport cost τ_{ij}
- Shapefiles required: Roman roads, medieval routes, navigable rivers, sea links
- Cost-of-transport data from historical sources
- Overlay Europe map with 10×10 km grids
- Extract the least-cost travel through each grid cell based on the technology present
- Run Dijkstra's algorithm for each populated-city dyad ($\sim$2,300,000 pairs)
- Save cumulative least-cost travel cost and calculate MA for each city

Population data: 2,169 cities, 1450–1650



Buringh (2021) European Urban Population panel

Two example least-cost paths



The Economic Geography of European Print

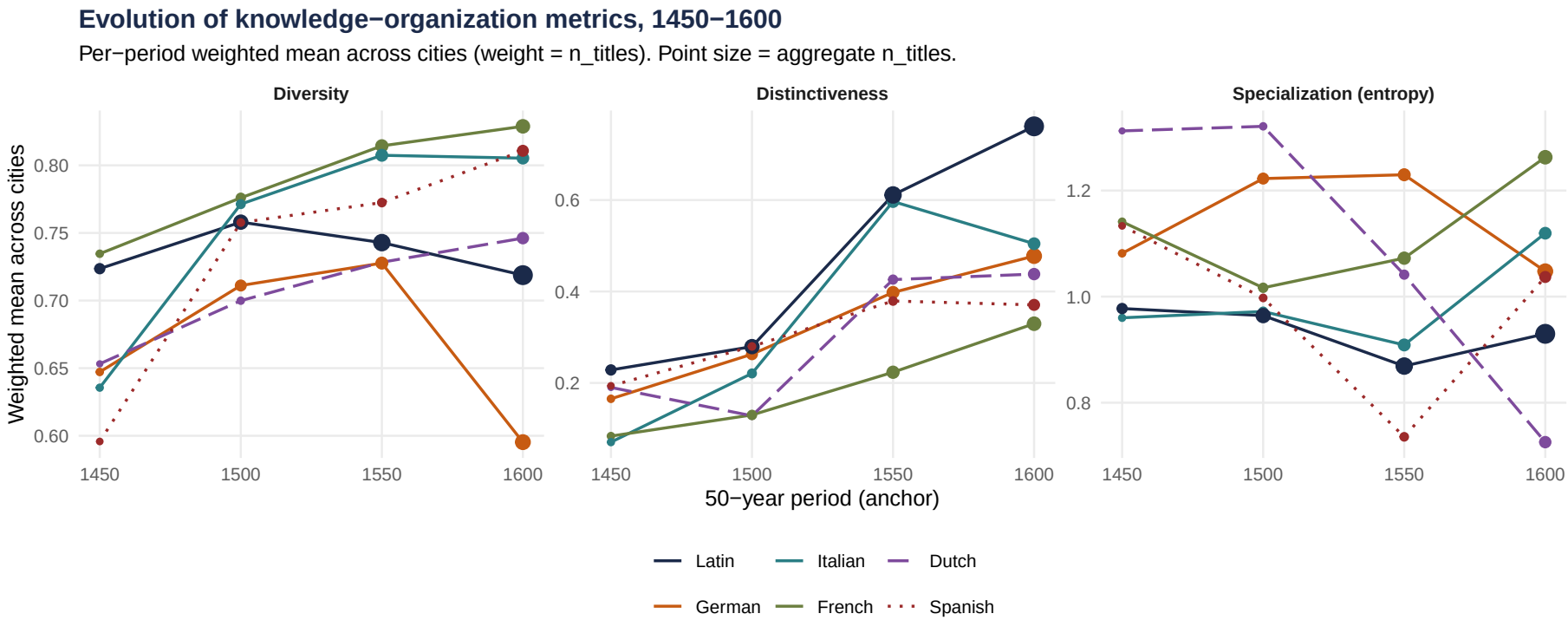
Language	N titles	N cities	Top-city share	Implied regime
Latin	301,216	873	Paris 9.0%	polycentric
German	117,223	459	Leipzig 10.4%	polycentric
Italian	65,124	259	Venezia 34.7%	semi-polycentric
French	94,447	366	Paris 53.6%	semi-polycentric (Paris-dominant)
Dutch	39,931	167	Amsterdam 29.1%	semi-polycentric
Spanish	32,752	234	Madrid 24.0%	semi-polycentric
English	51,502	109	London 85.1%	strongly monocentric

Six panel-credible languages used in the regression analysis; English excluded as monocentric.

We create three outcome variables from the title embeddings

- **Diversity**: the diversity of a city-period cell is the average cosine distance between its constituent titles and its own centroid.
- **Distinctiveness**: the distinctiveness of a city-period cell is the cosine distance between its centroid and the contemporaneous European centroid for the same language.
- **Entropy**: the Shannon entropy of the city-period titles' assignments to $K = 10$ topical clusters obtained from k-means on the full title-embedding corpus of language ℓ. Measures how evenly the city's titles are distributed across topical clusters.

We construct a 4-period, 50-year panel



Part III
Analysis and Results

Empirical Strategy

We estimate the Mundlak / Bell-Jones within-city specification:

$$y_{i\ell d} \;=\; \beta\, \widetilde{m}_{i\ell,d} \;+\; \delta\, \overline{m}_{i\ell} \;+\; \delta_d \;+\; \mu_\ell \;+\; \varepsilon_{i\ell d}$$

- $y_{i\ell d}$: composition metric (diversity / distinctiveness / entropy) for city i in language ℓ at decade d.
- $\widetilde{m}_{i\ell,d} \;=\; \log(\mathrm{MA})_{id} \;-\; \overline{\log \mathrm{MA}}_{i\ell}$ is the within-(city, language) deviation of $\log(\mathrm{MA})$ from its mean across periods.
- $\overline{m}_{i\ell}$ is that mean (the between-city component).
- δ_d, μ_ℓ are period and language fixed effects.
- β is the within-city longitudinal slope; δ is the between-city cross-sectional slope.

We identify off within-city variation

Language	$\log(\mathrm{MA})$	Diversity	Distinctiveness	Entropy
Dutch	5.8%	29.5%	32.6%	36.2%
French	4.5%	22.2%	22.1%	38.5%
German	7.2%	35.1%	32.6%	34.8%
Italian	8.3%	25.4%	38.2%	34.3%
Latin	8.6%	25.2%	52.6%	28.1%
Spanish	7.7%	26.4%	18.5%	24.6%

Within-city share of $\log(\mathrm{MA})$ *variance is thin (4.5–8.6%); composition metrics carry more.*

Theoretical assumptions

The Krugman + spatial-product-differentiation framework rests on five assumptions:

A1 Consumers value variety (CES aggregation across titles).

A2 Printing has increasing returns at the title level (fixed typesetting cost per title).

A3 Consumer preferences vary across cities in a positional way (Italian readers want different titles from Polish readers, in a sense that admits a distance metric in title space).

A4 The early modern book trade had bilateral transport costs with a proximity advantage (closer markets weigh more in a city's effective demand).

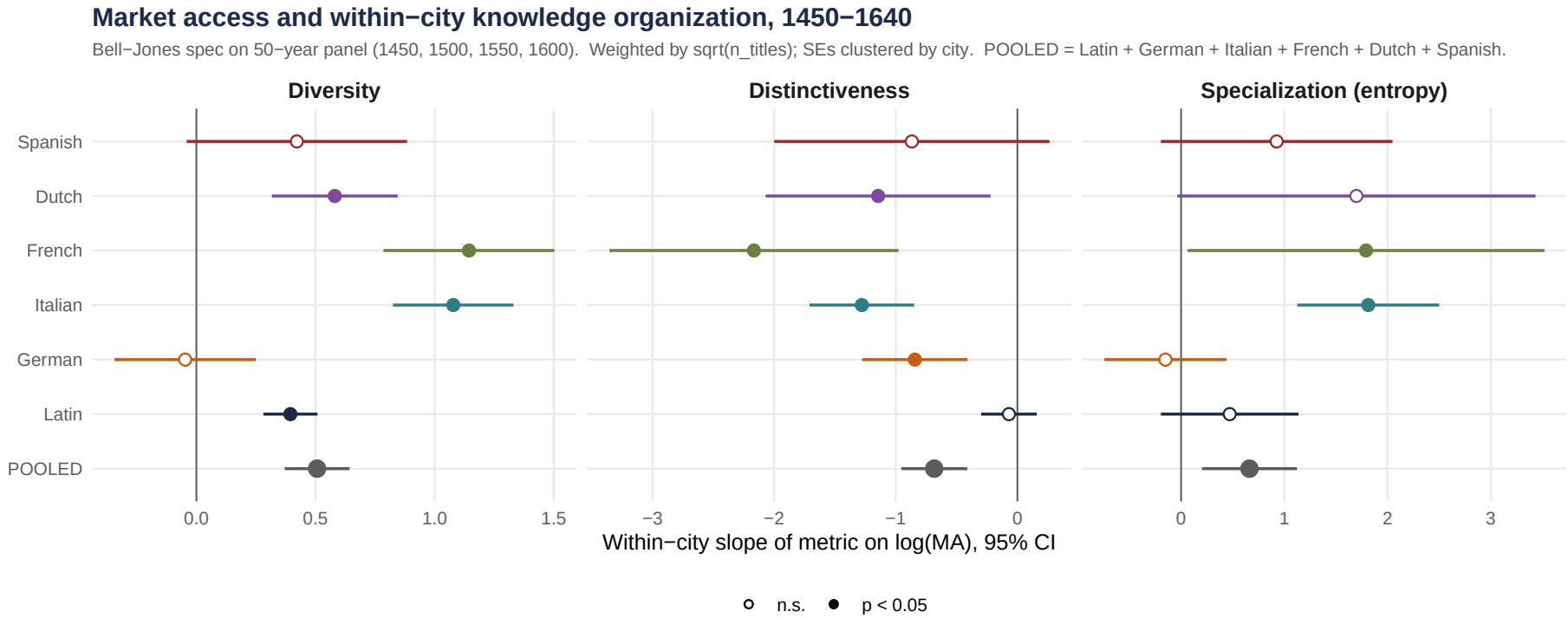
A5 Printers chose product positions strategically (they did not all print the same canon, and they responded to what other printers were producing).

Predictions from theory

Coefficient	Krugman + spatial diff.	Smith	Required assumptions
$\beta_{\text{diversity}}$	> 0	< 0	A1, A2
$\beta_{\text{distinctiveness}}$	< 0	> 0	A1–A5
$\beta_{\text{specialisation-entropy}}$	> 0	< 0	A1, A2, A3, A5

The two frameworks make opposing signed predictions on all three coefficients.

The impact of market access on within-city knowledge organisation



As market access increases: Diversity ↑ · Distinctiveness ↓ · Entropy ↑

What about city growth?

- We estimate Barro-style long-difference specifications to test whether any of our title-composition metrics predict subsequent 50-year city population growth.

$$\Delta_{50} \log(\text{pop})_{i,\, t \to t+50} = \beta\, y_{i,t} + \gamma \log(\text{MA})_{i,t} + \lambda \log(\text{pop})_{i,t} + \delta_t + \mu_\ell + \varepsilon_{i,t}$$

- $\Delta_{50} \log(\text{pop})_{i,\, t \to t+50}$: 50-year city population growth.
- $y_{i,t}$: a composition metric (diversity, distinctiveness, or entropy).
- $\log(\text{MA})_{i,t}$, $\log(\text{pop})_{i,t}$: market access and initial city size as controls.
- δ_t, μ_ℓ: period and language fixed effects.
- β is the coefficient of interest: does composition predict growth above and beyond MA?
- **NTT prior:** both growth and product/title variety are downstream of MA, so conditional on MA the composition metrics should carry no independent growth signal ($\beta = 0$).

Knowledge Composition and City Growth

Extension to the framework	Predicted sign
Mokyr (2002, 2009) useful-knowledge channel: cities printing a higher share of instrumental, scholarly, scientific content drive growth via Republic-of-Letters knowledge transfer.	$\hat{\beta}_{\text{share-useful}} > 0$ on the extensive margin; $\hat{\beta}_{\text{within-useful diversity}} > 0$ on the intensive margin
Endogenous growth via knowledge spillovers (Romer 1986; Grossman & Helpman 1991). Jacobs externalities (Glaeser et al. 1992): diverse knowledge stocks generate ideas, ideas drive growth.	$\hat{\beta}_{\text{diversity}} > 0$
Marshall externalities (Henderson 2003): own-industry specialisation generates within-industry productivity gains.	$\hat{\beta}_{\text{distinctiveness}} > 0$, $\hat{\beta}_{\text{diversity}} < 0$

Pooled Barro Growth Coefficients on each composition metric

Composition metric	$\hat{\beta}_{\text{metric}}$	$\hat{\gamma}_{\text{MA}}$	$\hat{\lambda}_{\text{pop}}$	n
Diversity	0.143^{*}	0.073^{**}	-0.009	1,882
	(0.082)	(0.029)	(0.015)	
Distinctiveness	-0.058	0.072^{**}	-0.006	1,882
	(0.065)	(0.029)	(0.014)	
Specialisation (entropy)	-0.036	0.070^{**}	-0.003	1,882
	(0.031)	(0.029)	(0.015)	
Share of useful content	0.123^{**}	0.068^{**}	-0.004	1,486
	(0.049)	(0.029)	(0.015)	

Table: Each row is a separate regression; standard errors clustered at the city in parentheses; observations weighted by $\sqrt{n_{\text{titles}}}$ (or $\sqrt{n_{\text{useful+ceremonial}}}$ for share-of-useful). Period and language fixed effects included; no city FE. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

How we measure Mokyr's "Useful Knowledge"

- **Useful:** Classical Authors, Medical Texts, Jurisprudence, Philosophy and Morality, Academic Dissertations, Periodicals, Educational Books, History and Chronicles, Political Tracts.
- **Ceremonial:** Religious, Funeral orations, Wedding pamphlets, Poetry, Drama.
- **Extensive margin (share of useful content):**

$$s_{i,t,\ell}^{\text{useful}} = \frac{n_{i,t,\ell}^{\text{useful}}}{n_{i,t,\ell}^{\text{useful}} + n_{i,t,\ell}^{\text{ceremonial}}}$$

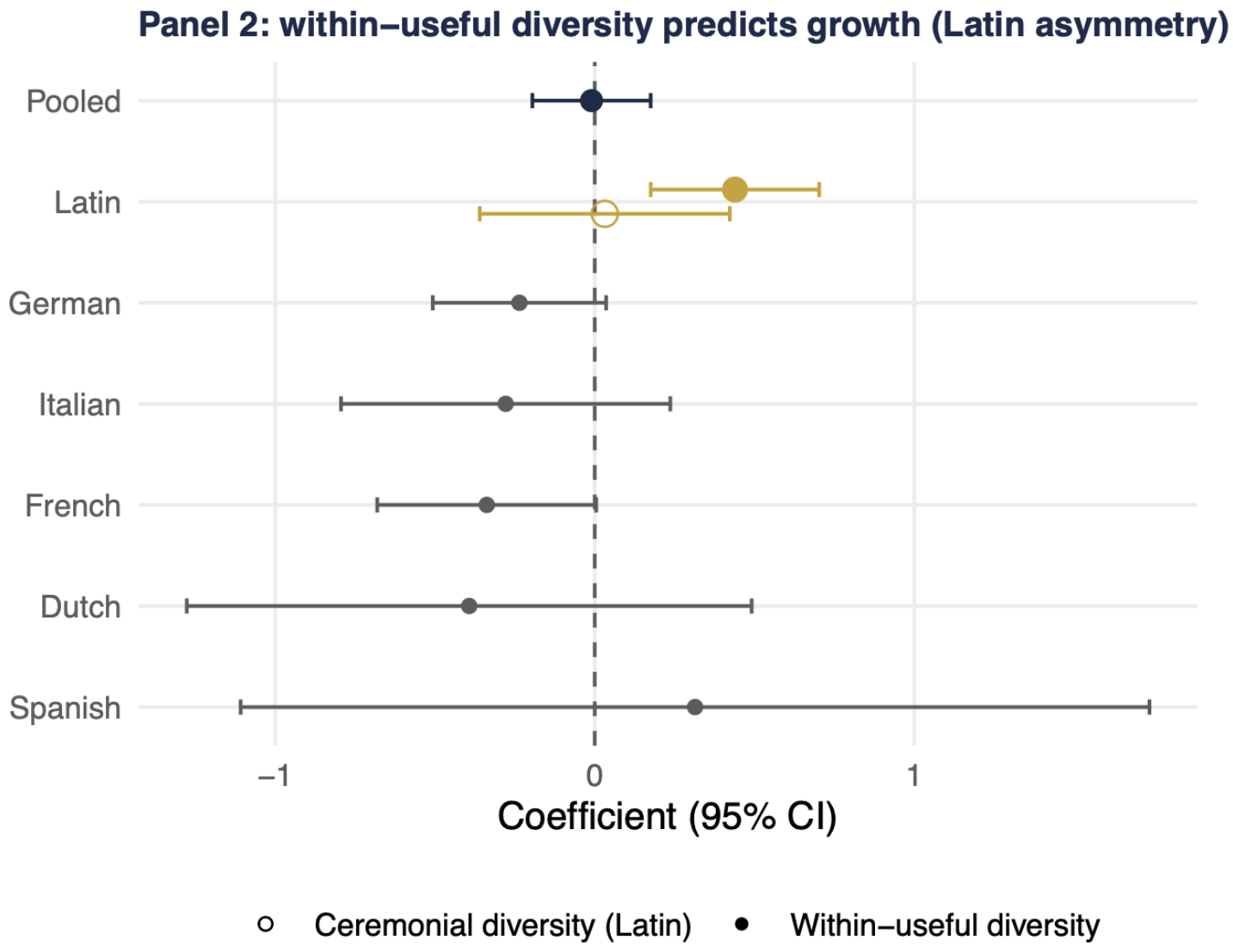
- **Intensive margin (within-useful diversity):** the diversity metric defined above, computed on the useful subset of titles only.

Share of useful and city growth (Barro long-difference)

	Pooled (1)	Latin (2)	German (3)	Italian (4)	French (5)	Dutch (6)	Spanish (7)
Share-of-useful$_{i,t}$	0.123**	0.084	0.247***	0.218**	0.167	0.066	0.270
	(0.049)	(0.076)	(0.081)	(0.105)	(0.134)	(0.235)	(0.207)
log(MA)$_{i,t}$	0.068**	0.084**	−0.027	0.059**	0.049	0.014	0.329***
	(0.029)	(0.034)	(0.033)	(0.025)	(0.051)	(0.090)	(0.106)
log(pop)$_{i,t}$	−0.004	−0.005	−0.071***	0.012	0.022	0.040	0.028
	(0.015)	(0.016)	(0.015)	(0.012)	(0.030)	(0.050)	(0.031)
Period FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language FE	Yes	No	No	No	No	No	No
Observations	1,486	666	286	148	185	93	108
Cities (clusters)	421	327	148	73	105	52	60

Dep. var. is $\Delta_{50} \log(\text{pop})_{i,t \to t+50}$. SE clustered by city; weighted by $\sqrt{n_{u+c}}$. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Within-useful diversity also predicts growth



Latin within-useful diversity coefficient $\hat{\beta}_{\text{within-useful}} = 0.439^{***}$ ($p = 0.001$).

Takeaways from the paper

- We document a joint MA $\rightarrow$ print composition $\rightarrow$ growth relationship across six European languages between 1450 and 1650, with a Mokyr-style middle node.
- Within-city increases in market access produce print output that is more topically *diverse*, *less distinct* from the contemporaneous European centroid, and *more even* across topical clusters — all three signs reverse the Smithian specialisation prior and support NTT.
- The useful-knowledge slice of composition — share of prescriptive, instrumental, scholarly subjects, and within-useful diversity in the trans-European Latin medium — predicts subsequent 50-year city growth conditional on $\log(\mathrm{MA})$ and initial $\log(\mathrm{pop})$.
- Cumulative European intellectual integration required, as one necessary condition, that the highest-market-access producers tilt toward a shared main-stream rather than entrenching local idiosyncrasies. Our measurement captures the slope of that tilt directly.

How we built the Latin UMAP figure

- Start with the 200-dimensional Latin title embeddings (one vector per title)
- Stratified subsample of $\sim$ 5,000 titles, drawn proportionally from the top USTC `classification1` categories
- Run `uwot::umap` with
  - `n_neighbors = 15` (local neighbourhood size)
  - `min_dist = 0.10` (controls clustering tightness)
  - `metric = "cosine"` (matches the embedding metric used everywhere else)
- Project the 200-d vectors to 2 dimensions
- Plot with `ggplot2`; colour each point by its USTC `classification1`
- Random seed pinned for reproducibility

Source: `scripts/18_embedding_explainer_figures.R`.

Tokenization & pre-processing: the universal pipeline

Applied identically to all seven panel languages (Latin, German, French, Italian, Dutch, Spanish, English); language-specific resources noted in parentheses.

- **Pre-tokenisation cleaning** on `short_title`:
  - Lower-case the title
  - Apply normalisation rules (language-specific orthography)
  - Strip imprint phrases (*e.g.* "anno domini", "chez")
  - Strip multipart phrases (*e.g.* "vol.", "tomus")
  - Strip non-letter, non-whitespace characters
- **Tokenisation**: `tidytext::unnest_tokens(..., token = "words")` — regex word-boundary splitter, no dictionary
- **Post-tokenisation filtering**:
  - Drop pure-digit tokens
  - Drop tokens with length ≤ 2
  - Anti-join a language-specific stopword list
  - Drop words appearing < 20 times in that language's corpus (`pmi_min_n`)

Source: scripts/lib/01_text_cleaning.R, scripts/config/<lang>.R.

Language-specific configuration

Language	Stopword source	Normalisation rules
Latin	local file `stopwords_la.txt`	`v`→`u`, `j`→`i`
German	`stopwords::stopwords("de")`	`v`→`u`, `j`→`i`
French	local `stopwords_fr.txt` / `quanteda("fr")`	`œ`→`oe`, `æ`→`ae`
Italian	local `stopwords_it.txt` / `quanteda("it")`	`v`→`u`, `j`→`i`
Dutch	local `stopwords_nl.txt` / `quanteda("nl")`	(none)
Spanish	local `stopwords_es.txt` / `quanteda("es")`	(none)
English	`tidytext::get_stopwords("en")`	`v`→`u`, `j`→`i`

Imprint phrases (*e.g.* "Parisiis", "Lugduni", "Venetiis" for Latin; "imprimé par", "chez" for French) and multipart phrases (*e.g.* "vol.", "tomus", "secunda") are curated per language in `scripts/config/<lang>.R`. Latin's `v`→`u`, `j`→`i` normalisation reflects classical orthography; French Œ/Æ ligatures decomposed to digraphs.

The Brahe frontispiece, in English

Latin (Frankfurt 1648):

TYCHONIS BRAHE
MATHIM: EMINENT: DANI
OPERA OMNIA,
Sive
ASTRONOMIAE INSTAURATAE
PROGYMNASMATA
In duas partes distributa,
QUORUM PRIMA DE RESTITUTIONE MOTUUM
Solis & Lunae, Stellarumqúe inerrantium tractat.
SECUNDA AUTEM DE MUNDI AETHEREI
Recentioribus Phaenomenis agit.

ANNO M.DC.XLVIII.

Editio ultima nunc cum Indicibus & Figuris prodit.
FRANCOFURTI,
Impensis Ioannis Godofredi Schönvvetteri.

English (working translation):

The Complete Works of TYCHO BRAHE,
Most Eminent Danish Mathematician,
or
Preliminary Exercises in the Restored Astronomy,
divided into two parts,
of which the FIRST treats the Restoration of the Motions of the Sun, the Moon, and the Fixed Stars; **the SECOND** treats the more recent Phenomena of the Ethereal World.

In the year 1648.

The latest edition, now issued with Indices and Figure
At Frankfurt,
At the expense of Johann Gottfried Schönwetter.

Notes. *MATHIM: EMINENT: DANI* contracts *Mathematici Eminentissimi Dani*. The trumpeting winged figure is *Fama* (Fame). *Mundi Aetherei recentioribus phaenomenis* refers to Brahe's most famous astronomical work — the comet of 1577 and the supernova of 1572 — which he showed were celestial rather than atmospheric, overturning the Aristotelian doctrine that the heavens above the moon were unchanging. Originally compiled 1588–1602; completed and published by Kepler in 1602 (Prague). This 1648 Frankfurt edition is the posthumous *opera omnia* with apparatus, prepared by Schönwetter.