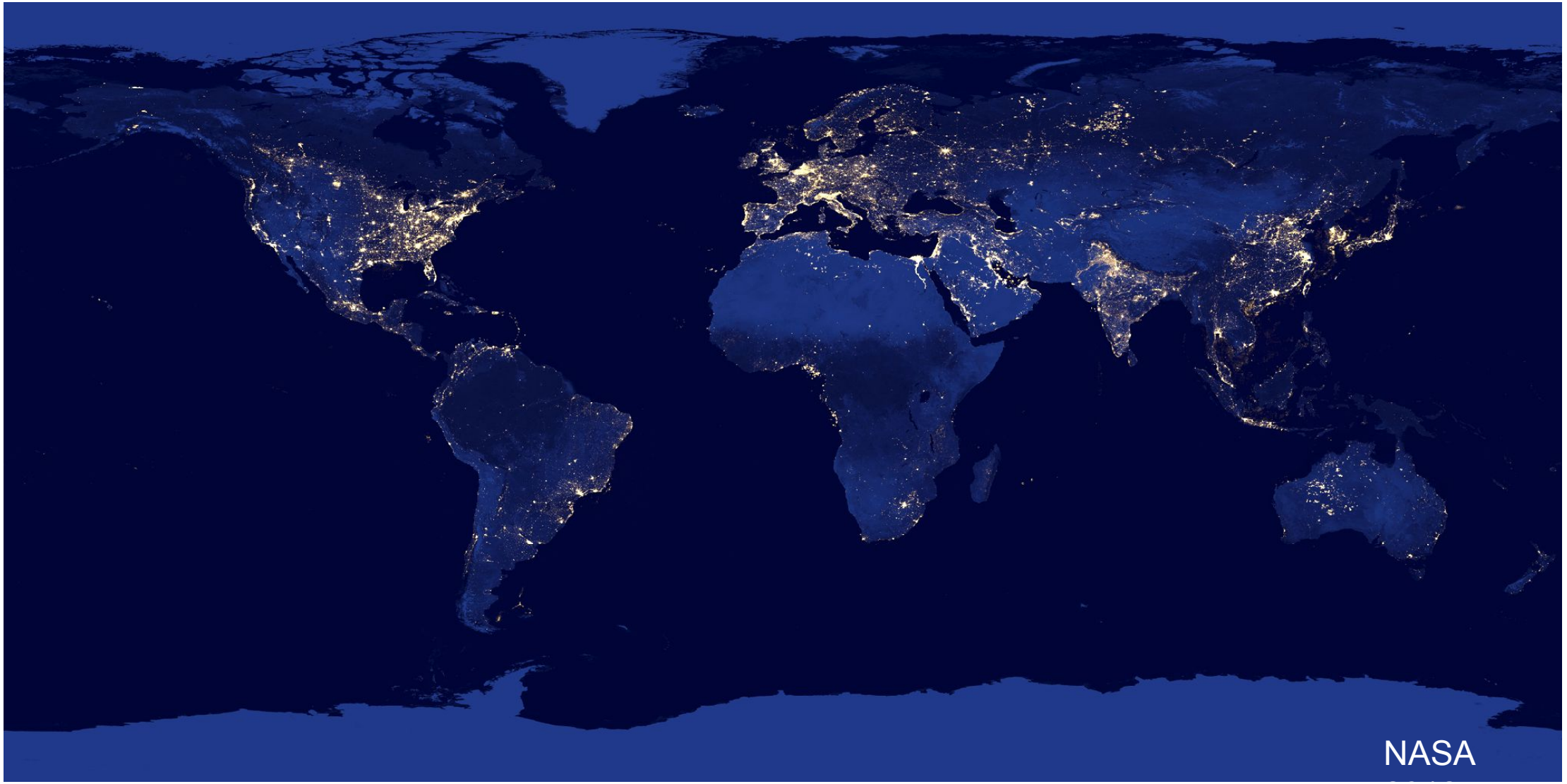
The Extent of the Market: A Method with an Application to the Wealth of Nations

Roy Elis, Stephen Haber, and Jordan Horrillo

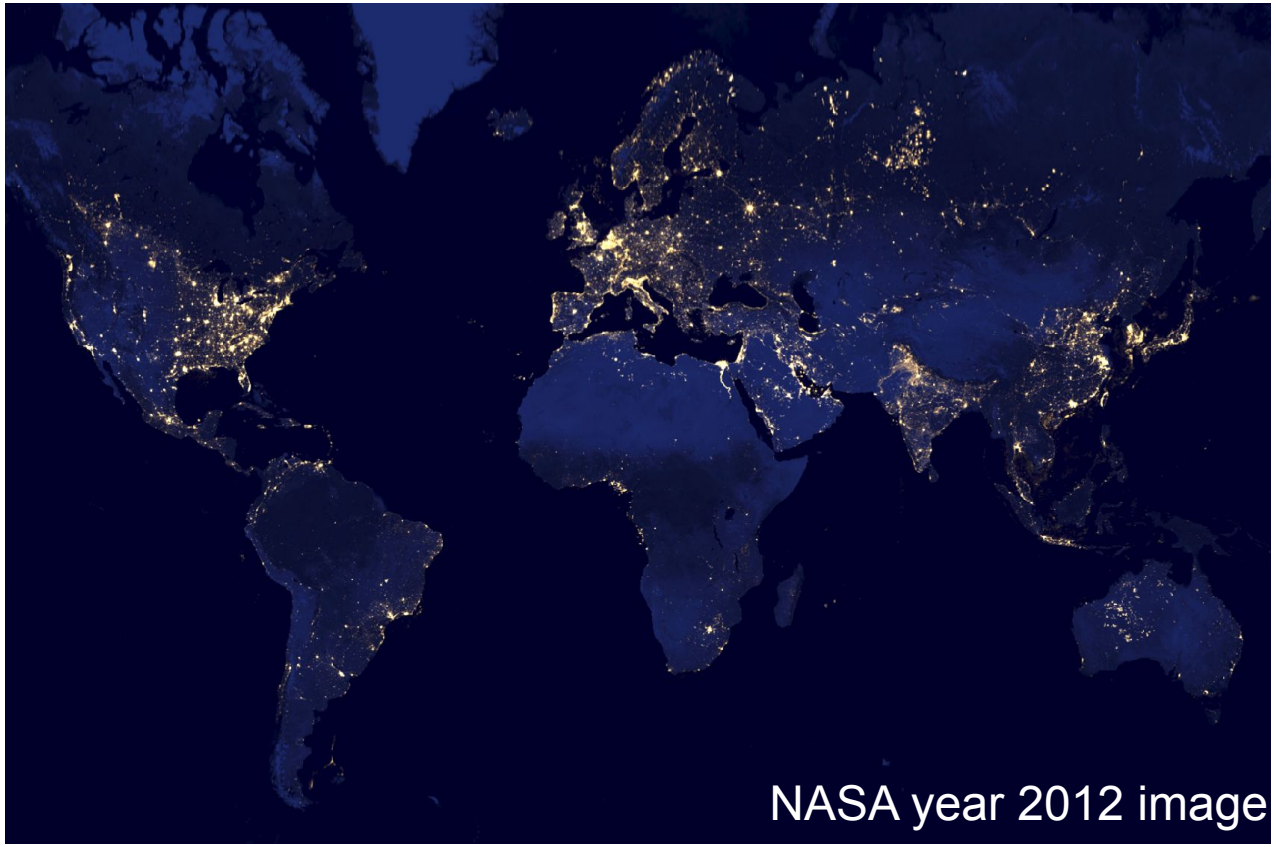
Presented at Hoover Economic Policy Working Group

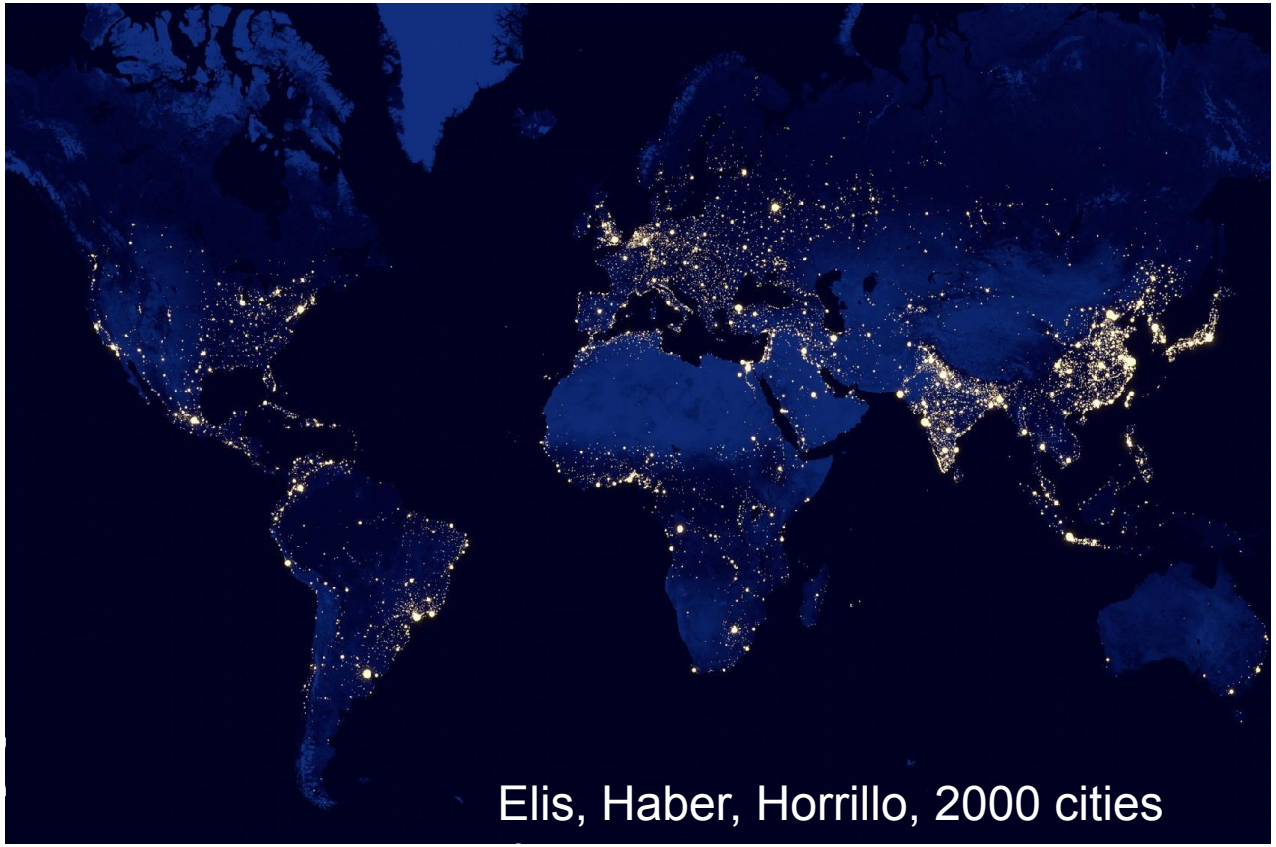
October 7, 2026

“Night Lights” have given rise to an immense literature on the distribution of economic activity

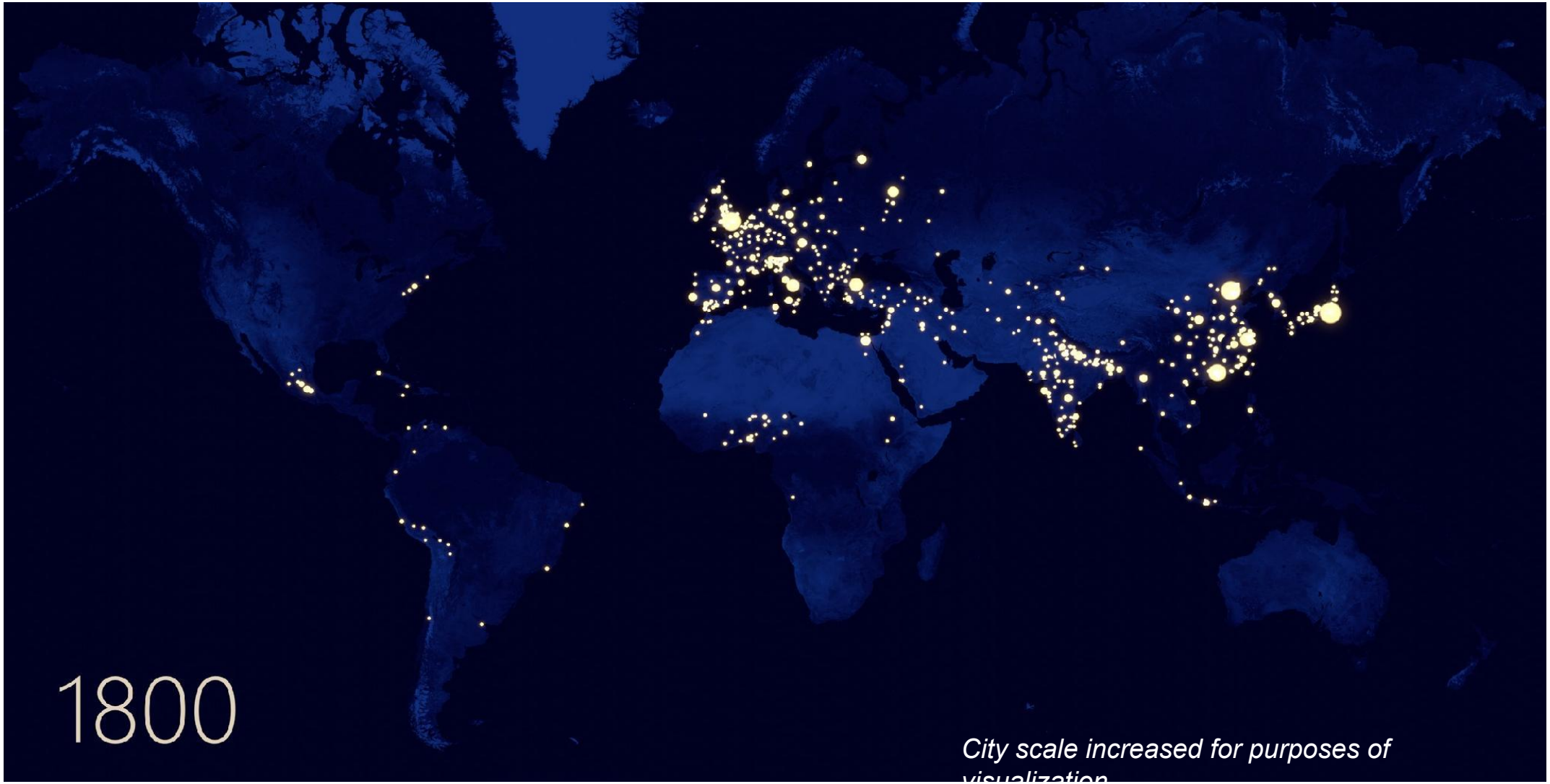


The “night lights” are, to a first approximation, detecting cities

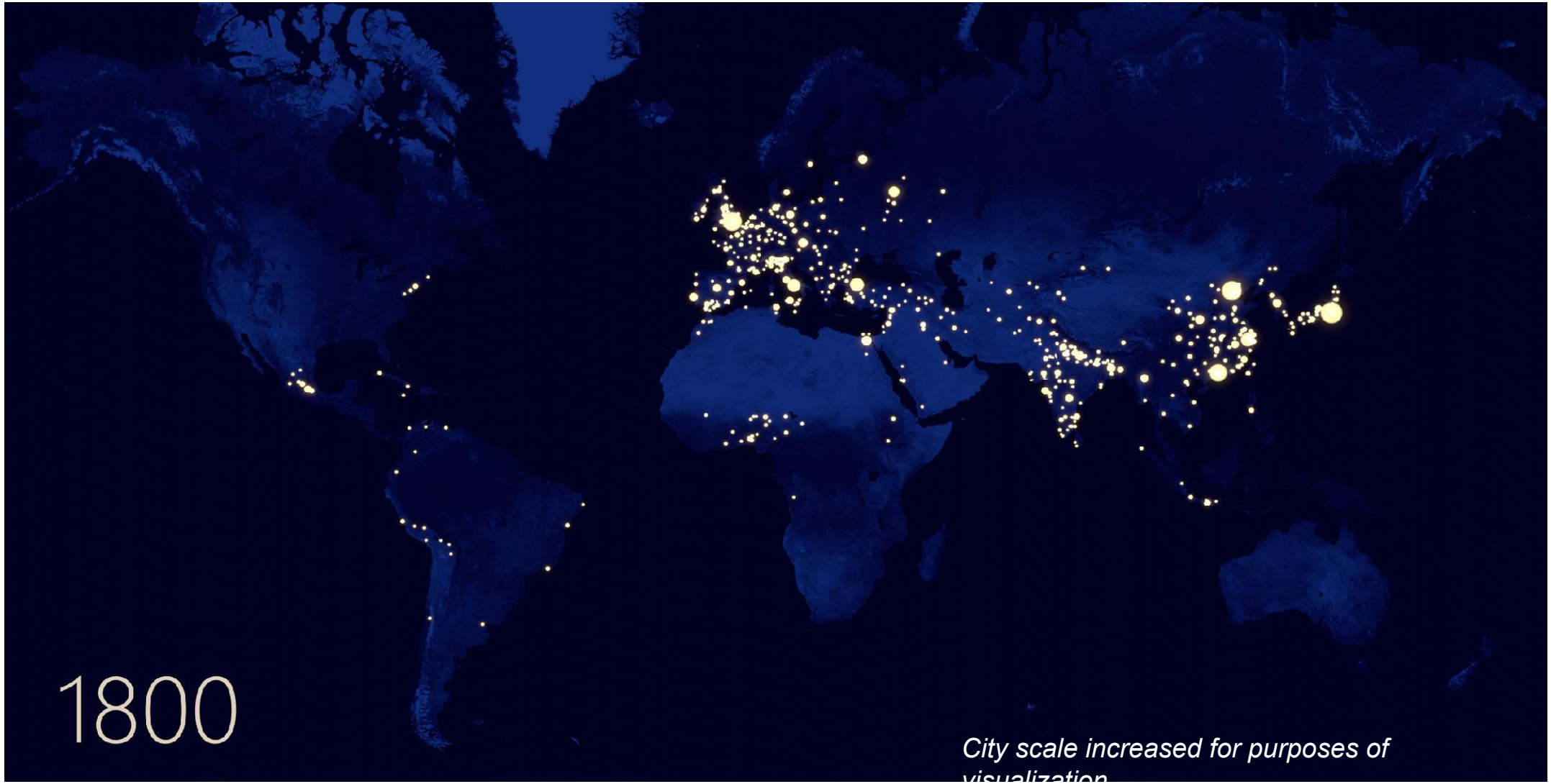




Our approximation of the distribution of economic activity based on city sizes and locations in 1800 …but, what accounts for this distribution?



In Eurasia and Africa, we are seeing the equilibrium outcome of four millennia of technological diffusion, trade, and migration



What we do in this paper

- We build a geo-spatial model that predicts where clusters of economic development in Eurasia and Africa (based on city sizes and locations) should have emerged circa 1800
- We then compare the predictions of the model against the historical record of city emergence circa 1800

The assumptions of our (parsimonious) model

1. Humans must eat or they die.

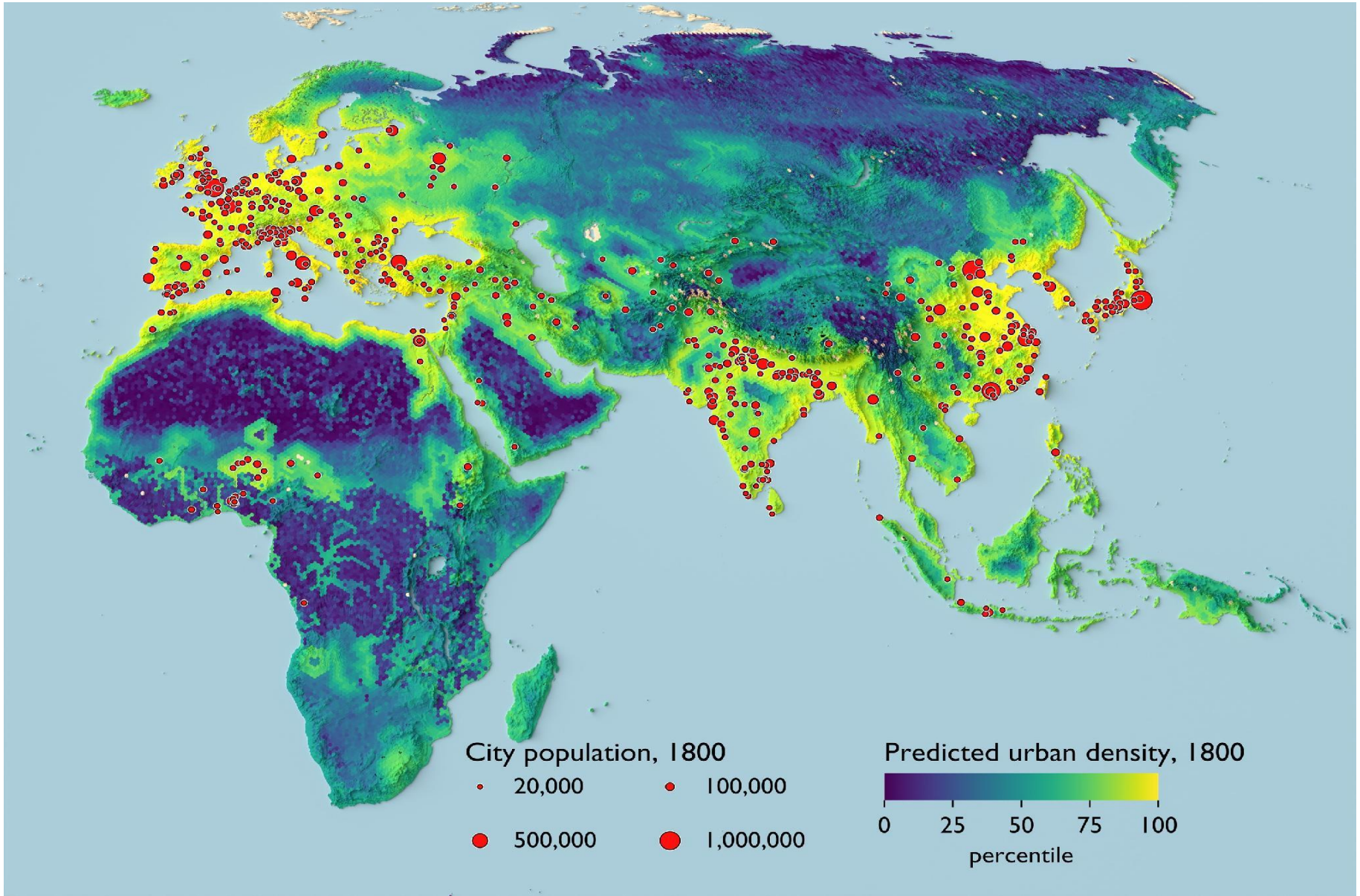
2. To support collocated populations of non-farmer specialists, humans had to be able to grow and move staple foods to cities.

3. The features of the Earth's surface in the 18th century, conditional on the state of technology in the 18th century, determined how far staple foods could be moved.

4. Clusters of cities emerged from:
   a) trade in staple foods from rural areas to cities;
   b) trade in manufactures and services from cities to rural areas;
   c) trade in manufactures and services across cities.

5. Clusters of cities indicate economic development through markets

What we discover substantively



What we discover methodologically

1. Market extents can be estimated as technology conditioned cost surfaces, allowing them to vary based on the features of the Earth's surface at a point in time, a state of technology at that point in time, and an energy budget that has verisimilitude for that point in time.
2. Market integration can be estimated by the degree to which market extents overlap one another.
3. Our model for estimating market extents and market integration has applications beyond economic development circa 1800. *The model can be tuned to other states of technology.*

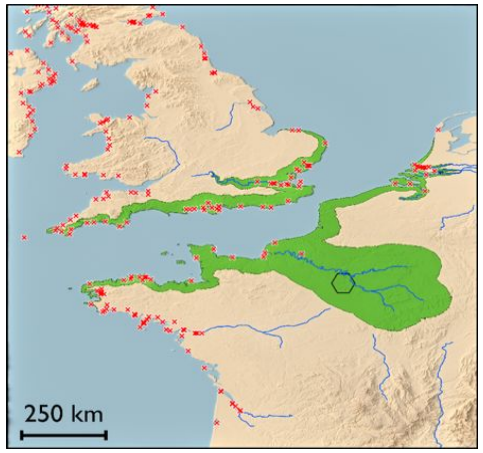
The market as the unit of analysis

We draw 41,435 markets from initial points 60 km apart. Market extents are a function of the distance that one metric ton of bulk goods could be moved from an initial point given:

- the energy costs of overcoming gravity, friction, and drag given 18th century technologies and the features of the Earth's surface as they existed in the 18th century;
- the most efficient combination of three 18th century transport technologies
- An energy budget of 40 megajoules.

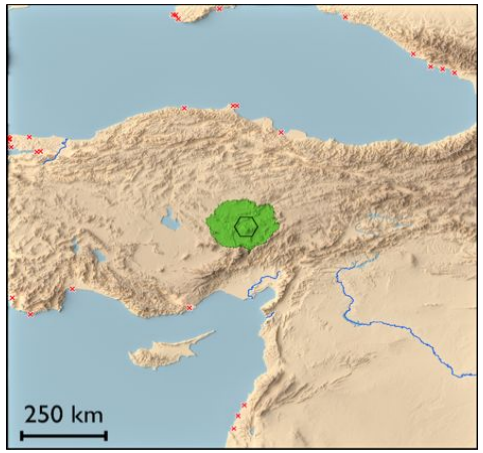
Each market has four characteristics:

- Size (in km^2)
- Natural harbors per km^2
- Staple food output per km^2
- Urban population per km^2



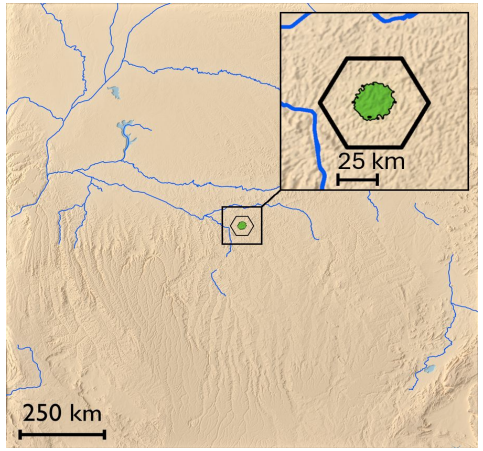
Paris, France

Area (1,000 km²)	121.7
Agricultural potential (Mkcal/km²/yr)	490.0
Harbors (count)	84



Kayseri, Turkey

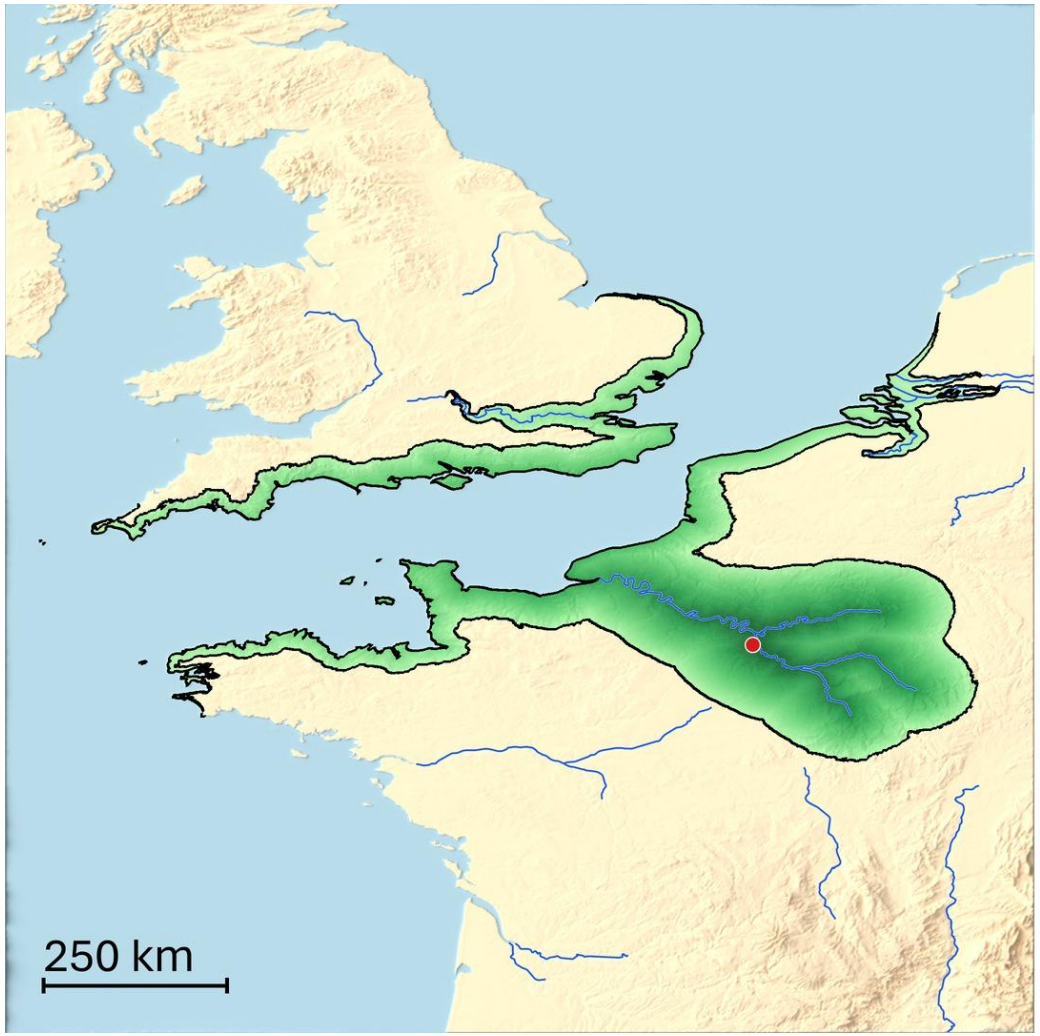
Area (1,000 km²)	19.5
Agricultural potential (Mkcal/km²/yr)	362.2
Harbors (count)	0



Nshenge, DRC

Area (1,000 km²)	0.5
Agricultural potential (Mkcal/km²/yr)	308.3
Harbors (count)	0

How a market's extent is drawn



One metric ton of grain leaves Paris

40 MJ

Energy budget
0.0 MJ left

0 MJ

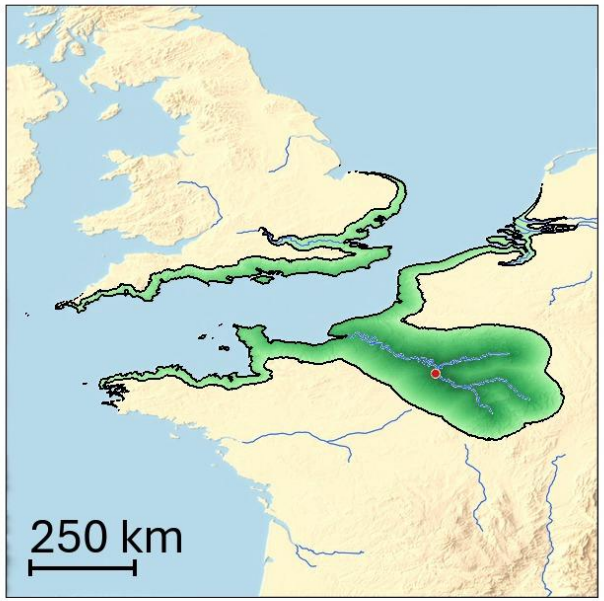
Cheapest mix of wagon, porter and boat over 18th-century terrain and navigable water

Boundary: where the 40 MJ runs out

40 MJ = 81 km by wagon on flat ground

Cost surface: gravity, friction and drag; cheapest mix of wagon, porter and boat. Darker green = more of the 40 MJ left on

Three markets on the same 40 MJ budget



Paris

121,681 km² of land

Reached by boat and wagon

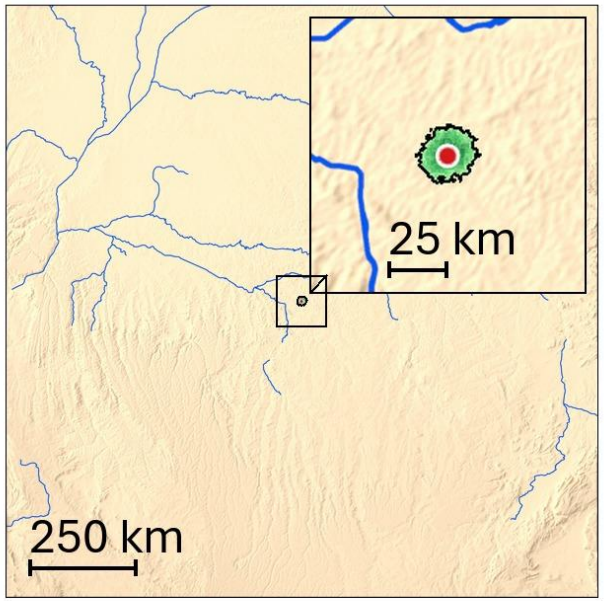


Kayseri

19,521 km² of land

1/6 of Paris

Reached by wagon



Nshenge

473 km² of land

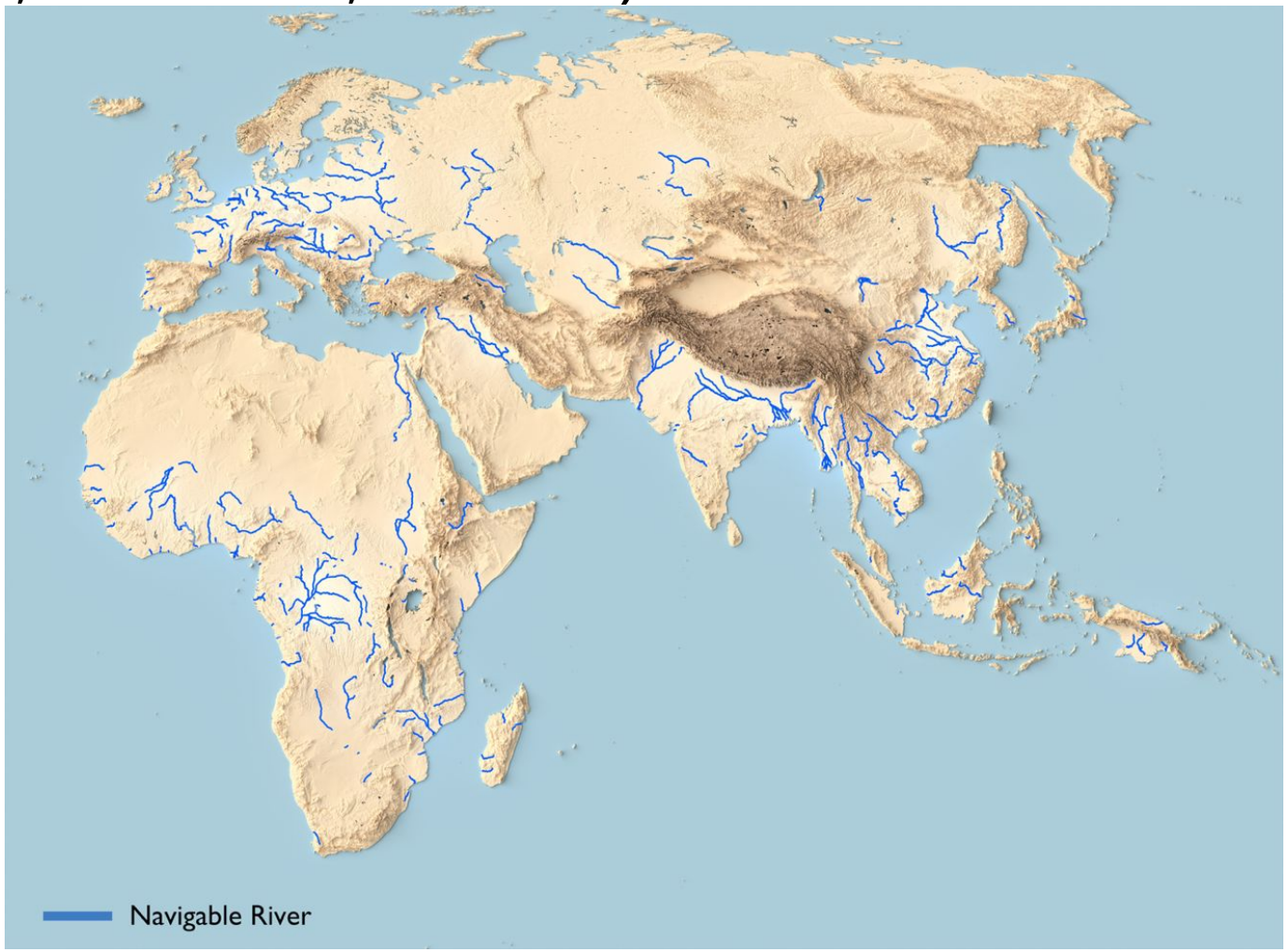
1/257 of Paris

Reached by porter

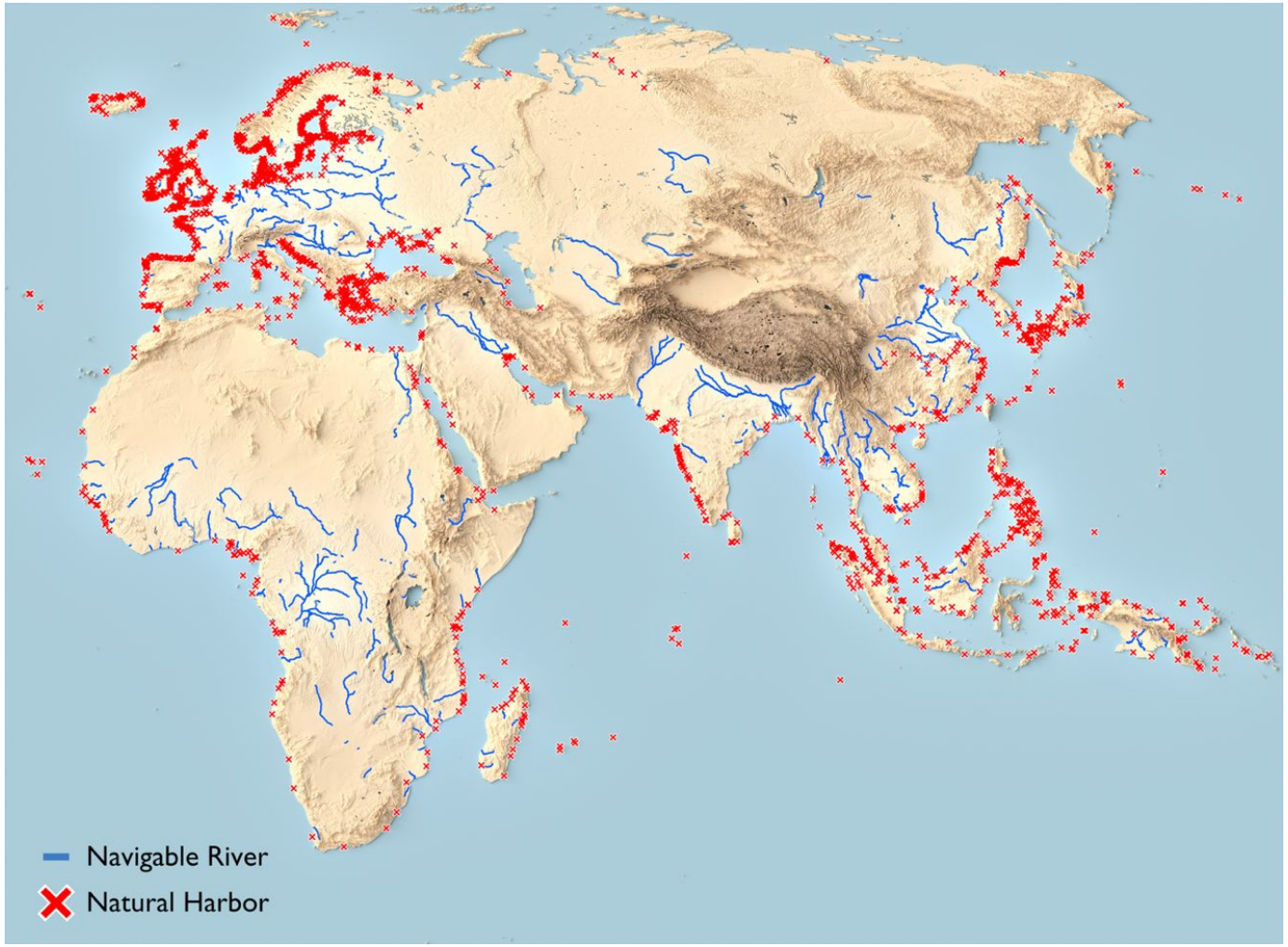
Energy left on arrival: 40 MJ (dark) to 0 MJ at the boundary (light)

Each panel is 1,400 km across. Areas are the land area of each market. Nshenge is enlarged in the inset.

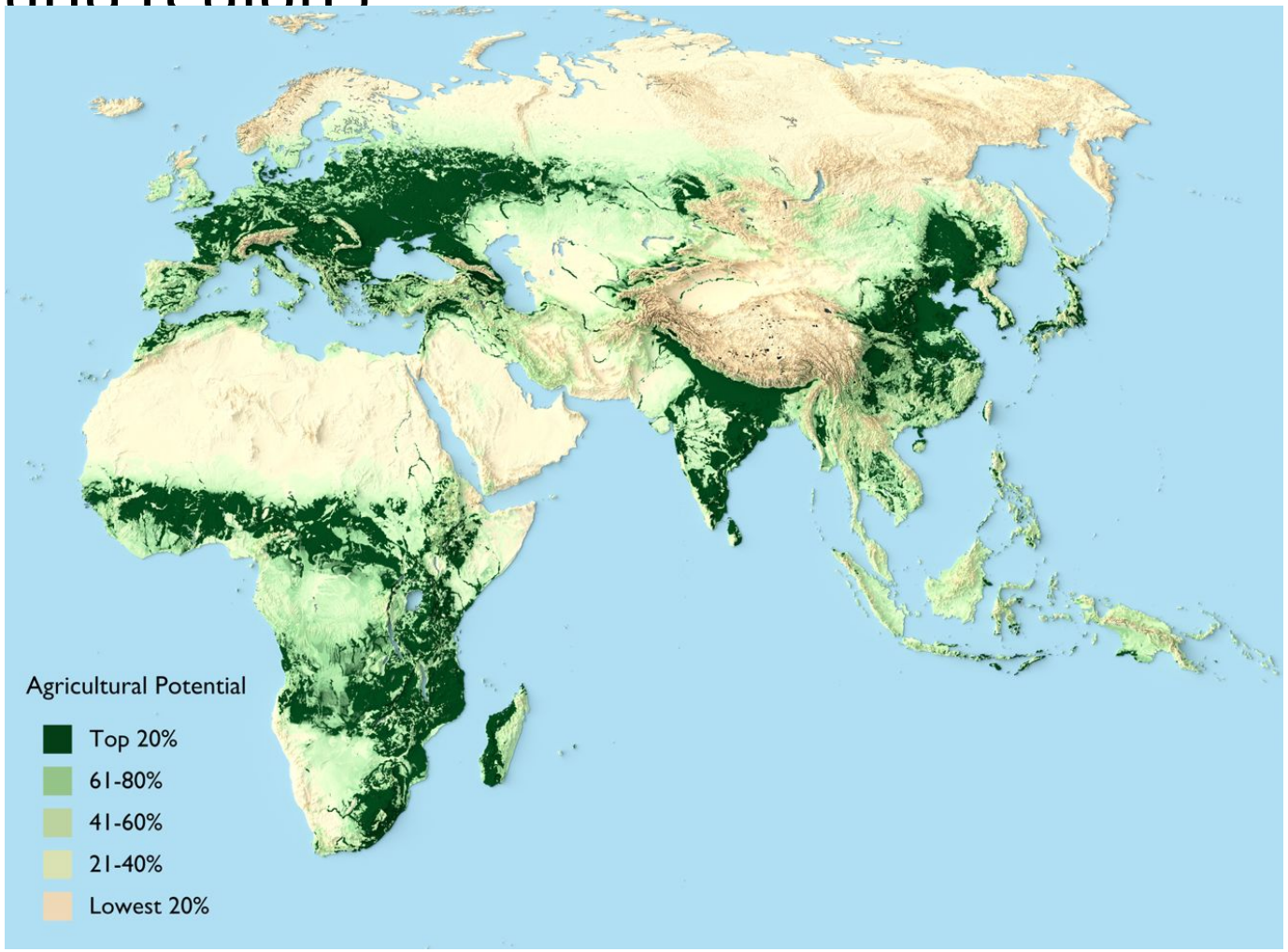
To draw markets, we take terrain slopes as they exist today, but generate maps of navigable water conditional on 18th century technologies and 18th century obstructions (rapids, waterfalls, sand bars, swamps, lakeshores, & sea ice)



We estimate natural harbors per U.S. Navy (1953), Deasy (1942), and Encyclopedia Britannica (1911)



We estimate potential staple crop calorie production using GAEZ 4.0 and USDA (2016) —but make an improvement to standard approaches by estimating the impact of traditional flood irrigation in arid regions



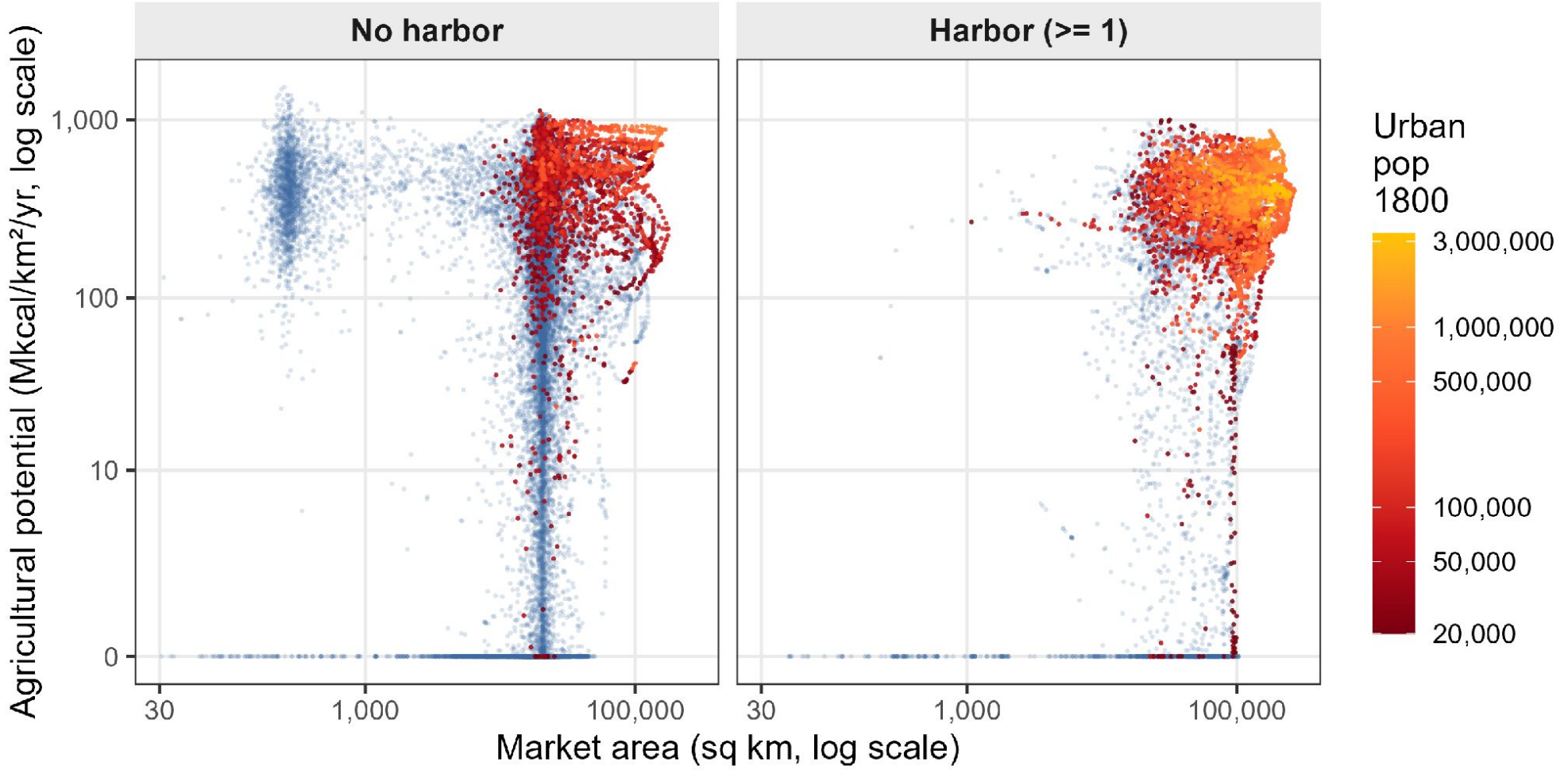
We estimate urban populations

- We build a dataset that identifies, geolocates, and provides population estimates in 1700 and 1800 for any city that meets a threshold size of 20,000 inhabitants.
- We construct the dataset by integrating information from online datasets compiled by other scholars, monographs by historians and archaeologists, specialist encyclopedias, historical dictionaries, and colonial gazetteers.
- Many 18th century city centers were located on the banks of rivers, lakes, and oceans. The present-day geo-coordinates for cities (typically fixed at their city hall) may not accurately capture their historical proximity to navigable water. We therefore “snap” the geo-coordinates of a present-day city to the banks of the nearest historically navigable body of water, if it is within 10 kilometers of that body of water.

Summary statistics of markets

Variable	N	Mean	SD	Min	P25	Median	P75	Max
Market Area (km^2)	41,435	34,002.5	36,268.2	30.9	19,002	20,799.5	29,117.3	267,230.8
Agricultural Potential (Mkcal/km^2/yr)	41,435	205.2	230.4	0	0	132.1	355.3	1,525.1
Harbors (count)	41,435	7.7	31.4	0	0	0	0	359
Harbors per 1,000 km^2	41,435	0.11	0.86	0	0	0	0	79.34
Cities in 1700 (count)	41,435	0.5	2.1	0	0	0	0	29
Urban Population in 1700	41,435	38,300.4	179,378.2	0	0	0	0	2,331,000
Urban Population per km^2, 1700	41,435	0.44	1.91	0	0	0	0	41.98
Cities in 1800 (count)	41,435	0.8	3.1	0	0	0	0	45
Urban Population in 1800	41,435	58,010	246,583.7	0	0	0	0	3,322,000
Urban Population per km^2, 1800	41,435	0.67	2.49	0	0	0	0	57.41

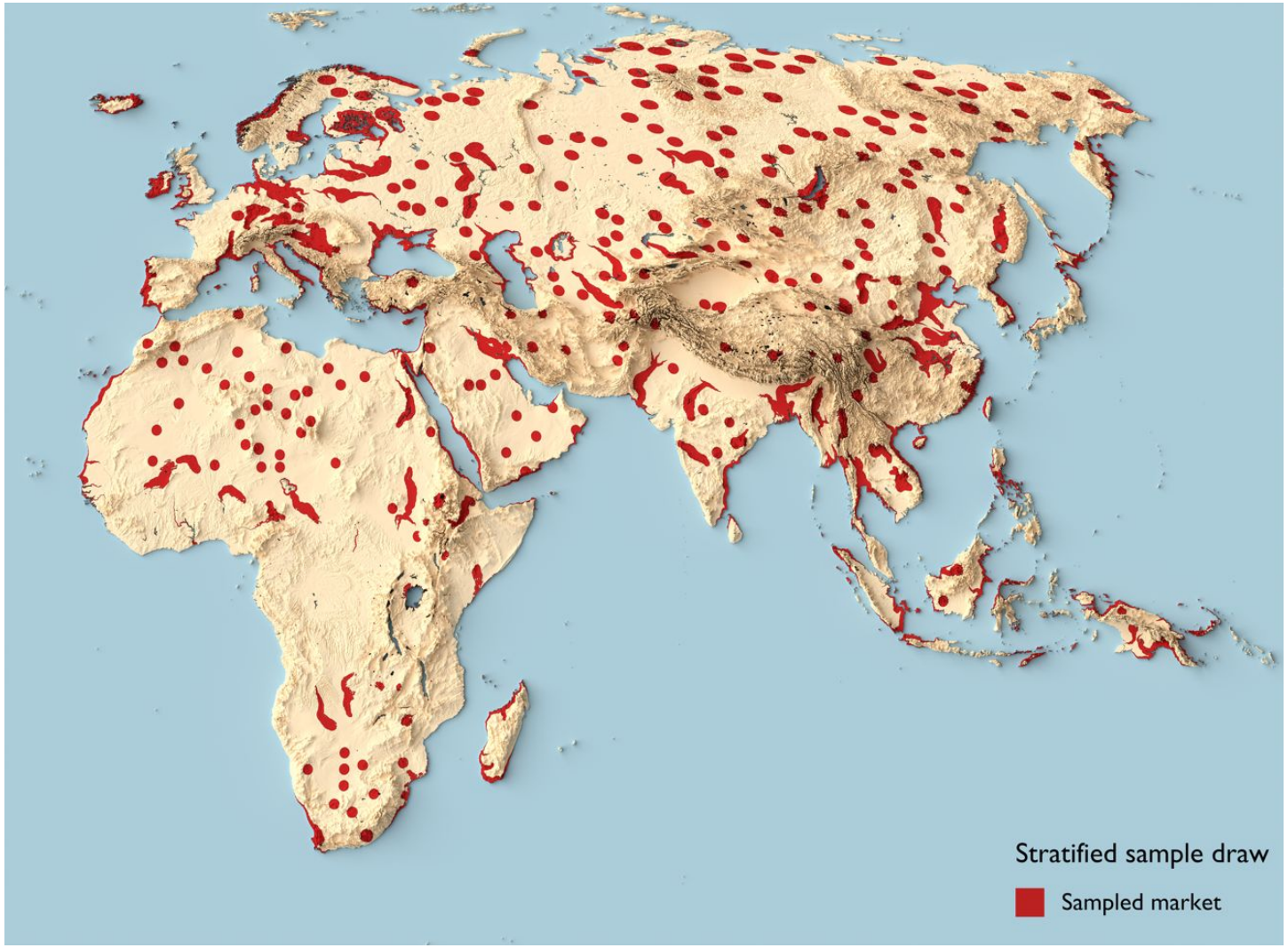
The data suggest that urban populations emerge through three-way, non-linear interactions



N = 41,435 Eurasian and African markets (area >= 30 km²).
Blue = no urban population in 1800; red to gold = increasing urban population
(color scaled by log(1+population); the largest 1800 cities are brightest).

We estimate random forest regressions on stratified samples of non-overlapping markets

- Markets are split into two strata at the 85th percentile of market area; large markets are placed first
- No two markets in a draw overlap
- 6,000 draws of 400 markets: 4,800 train the forest (one tree each), 1,200 are held out
- Fit is the holdout R^2, averaged across the 1,200 held-out draws

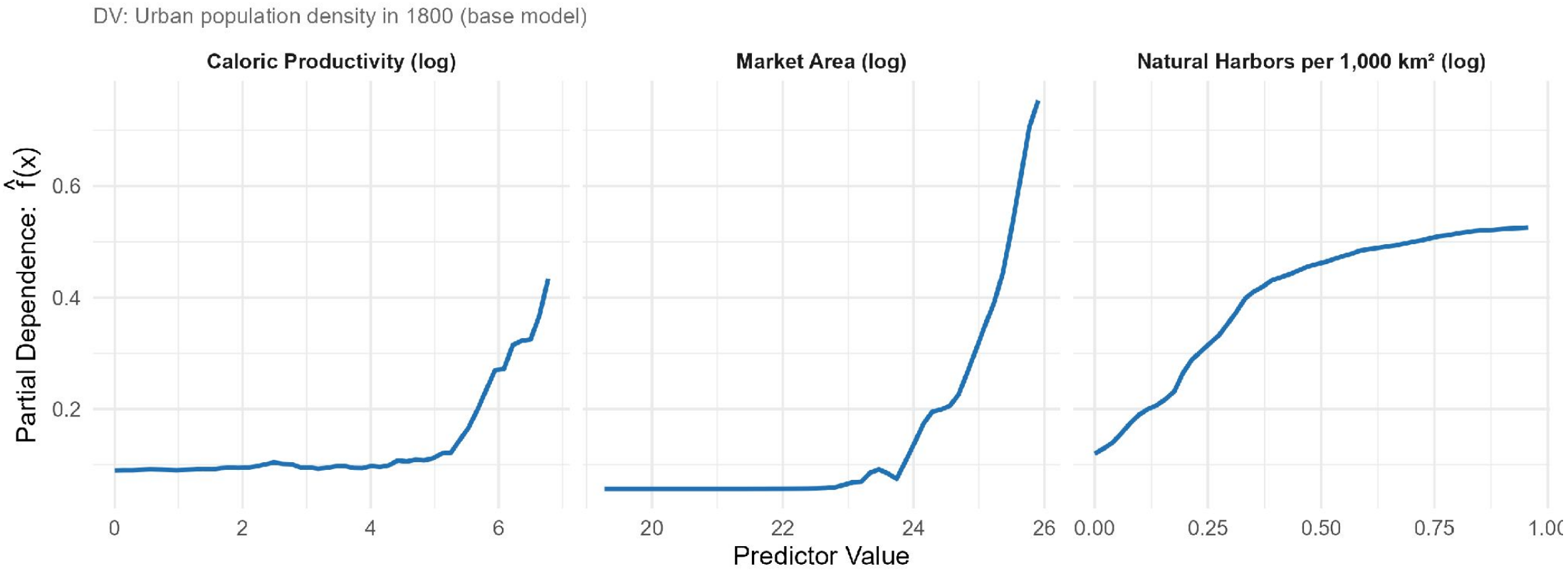


Base Random Forest Results: Market Area, Caloric Density, Density of Natural Harbors

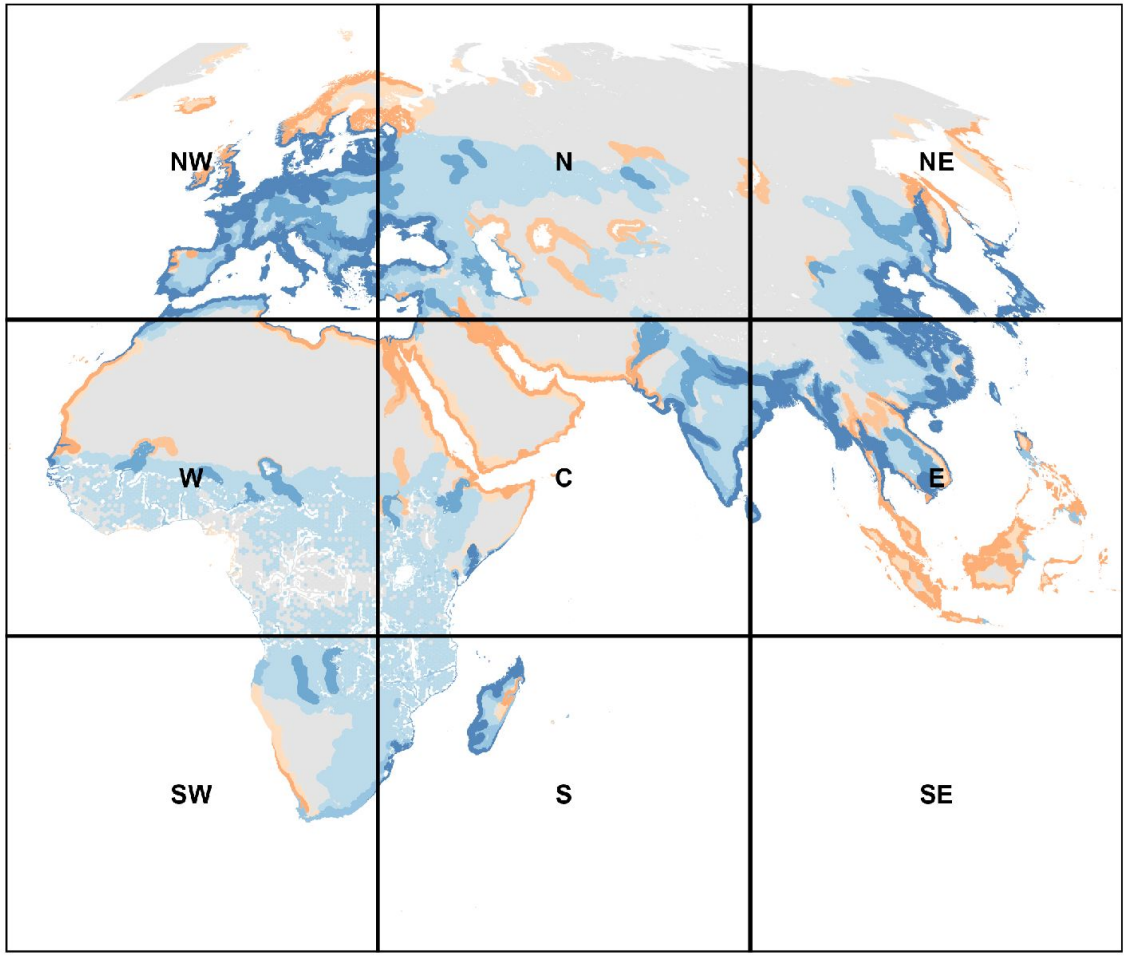
Panel A: Ensemble Specification		
Dependent variable	$\ln(1 + \text{urban population per km}^2 \text{ of market area, 1800})$	
Predictors	ln area; ln caloric density; ln harbor density	
Potential markets	41,435	
Training / holdout draws	4,800 / 1,200 (400 markets per draw)	
Hyperparameters	`mtry`=3, min. node size=15, max depth=7	
Panel B: Ensemble Performance	R^2	RMSE
Cross-validated, training draws	0.477 (0.004)	0.340 (0.002)
Training draws, in sample	0.477 (0.042)	0.339 (0.026)
Held-out draws	0.477 (0.042)	0.340 (0.026)
OLS benchmark, held-out draws	0.237 (0.034)	0.410 (0.023)
Permutation null (100 shuffles of DV)	0.010 (max 0.043)	0.467
Panel C: Variable Importance (share of impurity reduction)		
Market area (log)	45.5%	
Caloric density (log)	30.2%	
Harbor density (log)	24.3%	
Panel D: Spatial (LOBO) Cross-Validation		
Holdout R^2, range across nine dropped blocks	0.385 – 0.509	
Holdout R^2, Northwest block dropped	0.385 (0.049)	
Jackknife 95% CI for holdout R^2	[0.24, 0.71]	

Mean (SD across draws; cross-validated: across 5 folds). Training draws, in sample: the 4,800-tree forest on its own training draws. OLS: one fit per training draw, coefficients averaged. Permutation: 100 shuffles, 4,800 trees, 1,200 holdout draws each. Panel D (as the paper): each block of a 3×3 grid is dropped from training and holdout draws; R^2 on the remaining holdout markets. The market's own population is never a predictor.

Partial Dependence Plots, Base Model



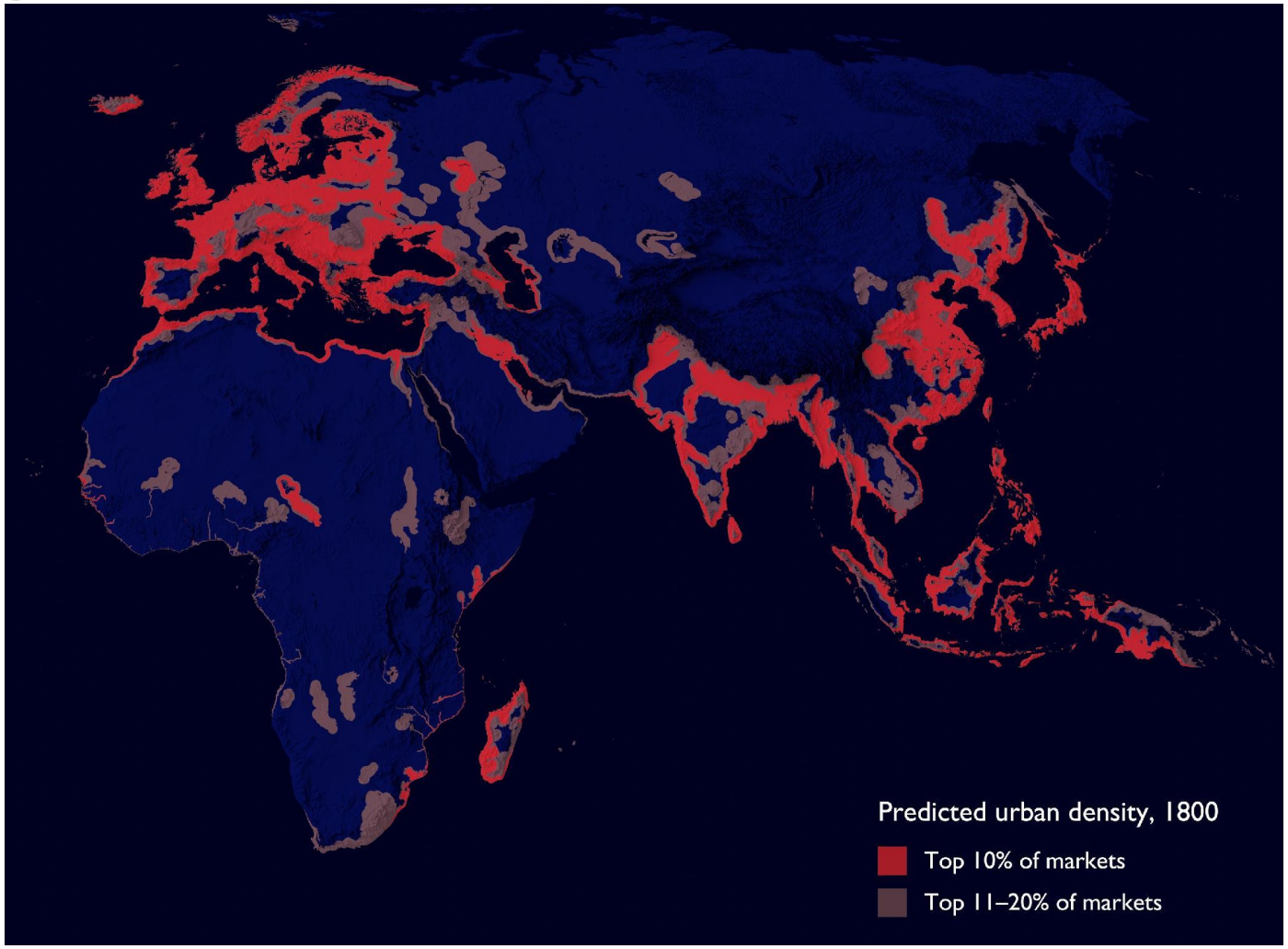
Are our results sensitive to dropping regions?



Block dropped	Markets dropped	Held-out R^2 (SD)
Northwest	3,314	0.385 (0.049)
North	12,936	0.509 (0.045)
Northeast	6,393	0.442 (0.046)
West	5,203	0.474 (0.043)
Center	5,770	0.475 (0.045)
East	3,712	0.445 (0.047)
Southwest	1,113	0.487 (0.044)
South	1,286	0.504 (0.043)
Southeast	14	0.481 (0.043)
None (all markets)	–	0.477 (0.042)

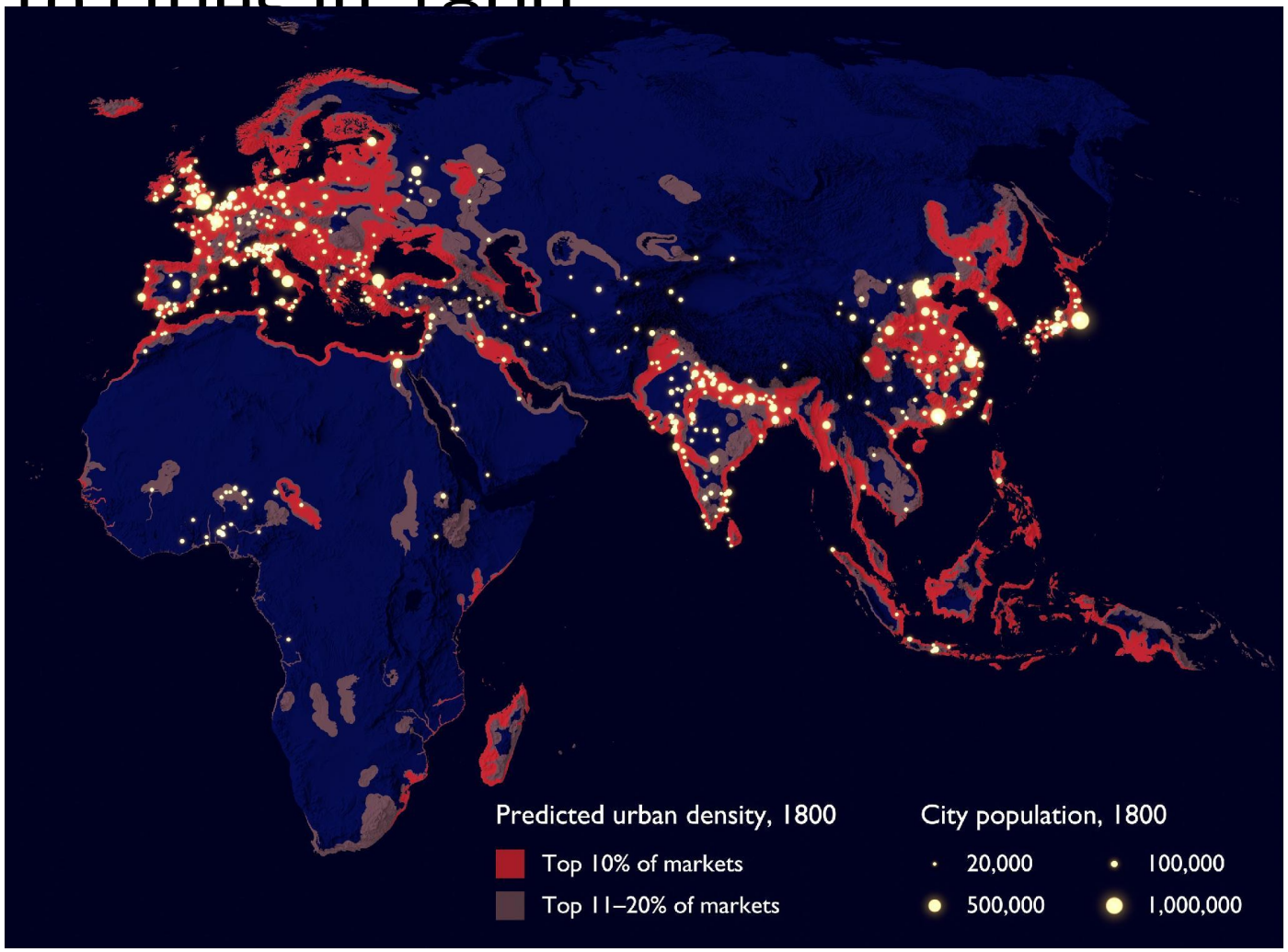
Base model: market area, caloric density, harbor density. As the paper's Table 4, Panel D: 3×3 grid over market centroids.
Each block is dropped from the training and holdout draws; R^2 (SD across the 1,200 holdout draws) is scored on the remaining holdout markets.

Top Predicted Regions of Economic Development: Out-of-Bag Predictions from Base Random Forest



Model: random forest, geography only (market area, caloric density, harbor density); out-of-sample (out-of-bag) predictions; holdout $R^2 = 0.48$

Top Predicted Regions of Economic Development: Out-of-Bag Predictions from Base Random Forest compared to cities in 1800



Model: random forest, geography only (market area, caloric density, harbor density); out-of-sample (out-of-bag) predictions; holdout $R^2 = 0.48$

Base model, out-of-bag predictions

	Top 10%	Top 20%
% of Old World land	15.5%	27.4%
% of 1800 urban pop.	81.1%	89.2%
Location quotient	5.2	3.3

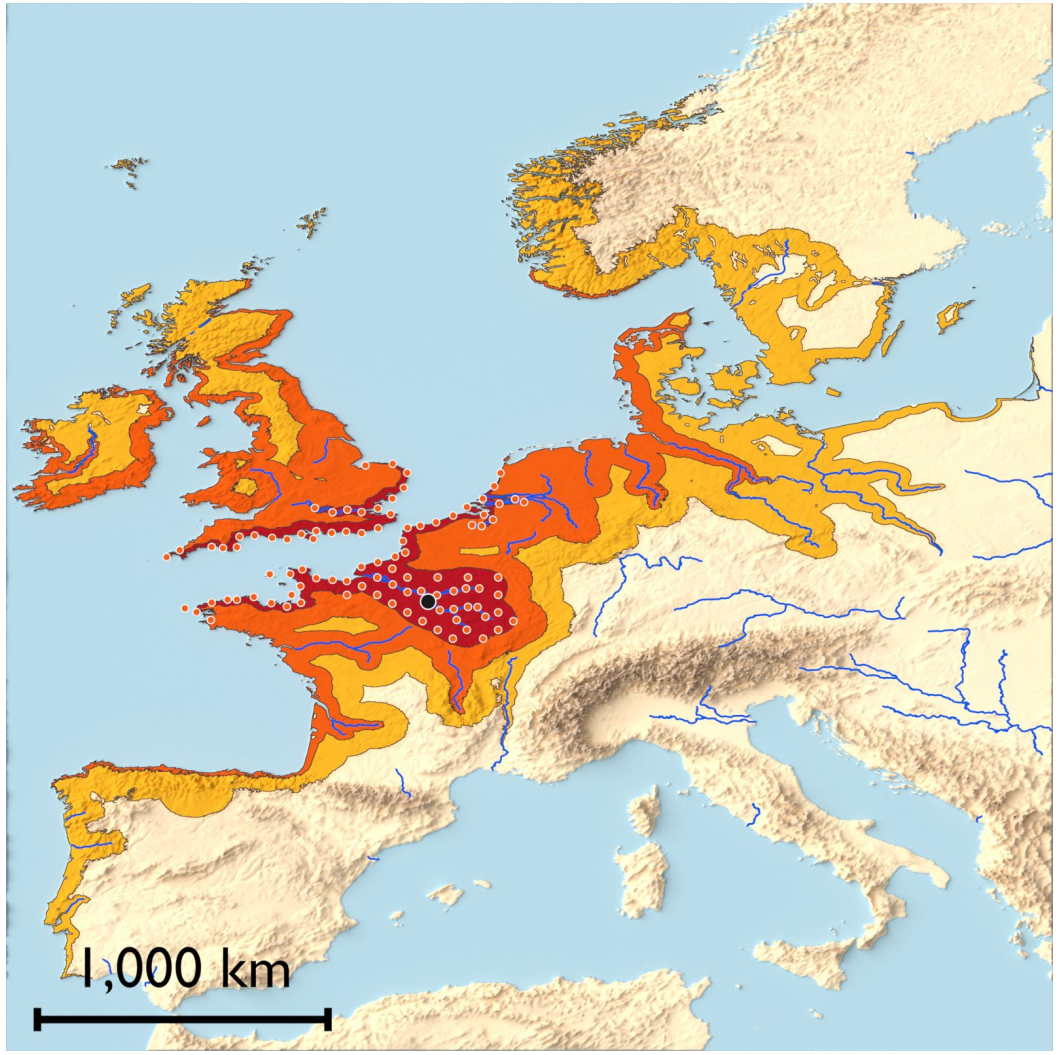
Implications

- Economists focus on markets as the fundamental unit of economic behavior but have not operationalized the concept.
- We have developed a way to operationalize the concept of a market.
- The concept of markets as the unit of analysis can be applied widely, using different states of technology.

Extensions

- Demonstrate that markets outperform grid-cells in predicting the location and clustering of economic development.
- Take into account spatial lags/network effects beyond those already captured.

Building the Market and Network of Paris



1. **Paris's market**: land reachable from its origin (**black dot**) within a 40 MJ energy budget.

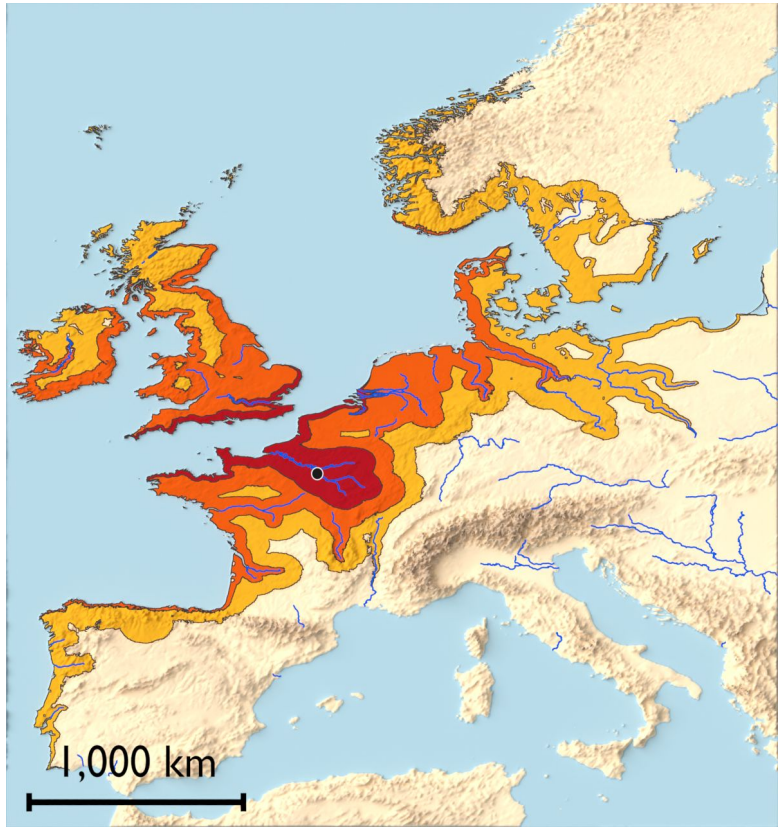
2. Each **orange dot** is the origin of a market that Paris's market reaches (103 neighboring markets).

3. **Ring 1**: land Paris can reach through one neighbor, beyond its own market.

4. **Ring 2**: land Paris can reach through two neighbors, beyond ring 1 (397 more markets).

Markets are neighboring when one contains the other's origin, counting water routes.

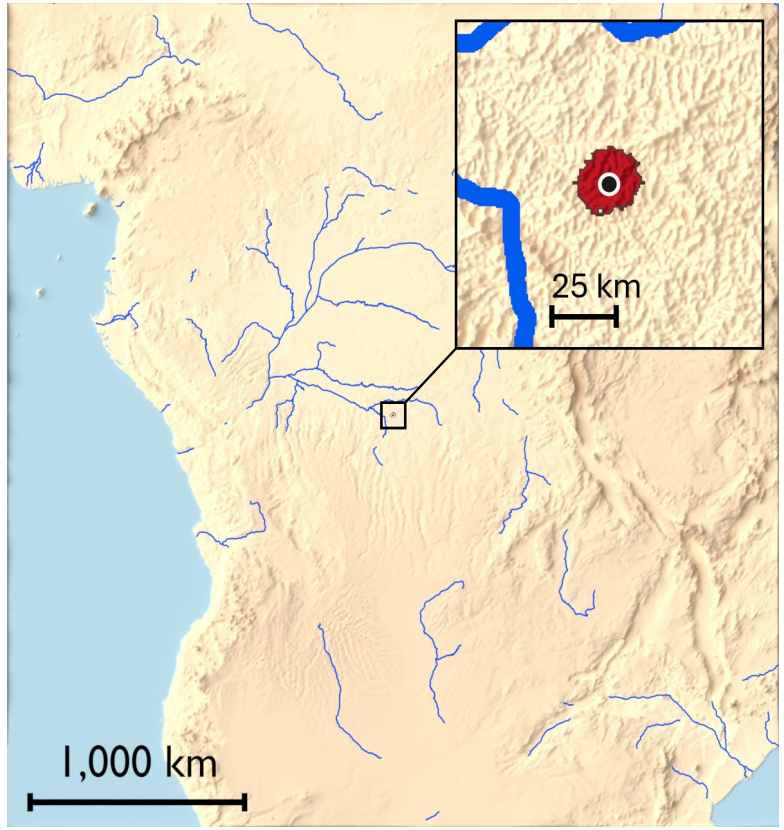
Our approach allows us to estimate urban population densities using the characteristics of linked markets as spatial lags



Paris: 103 / 397 markets



Kayseri: 8 / 23 markets



Nshenge: no linked markets

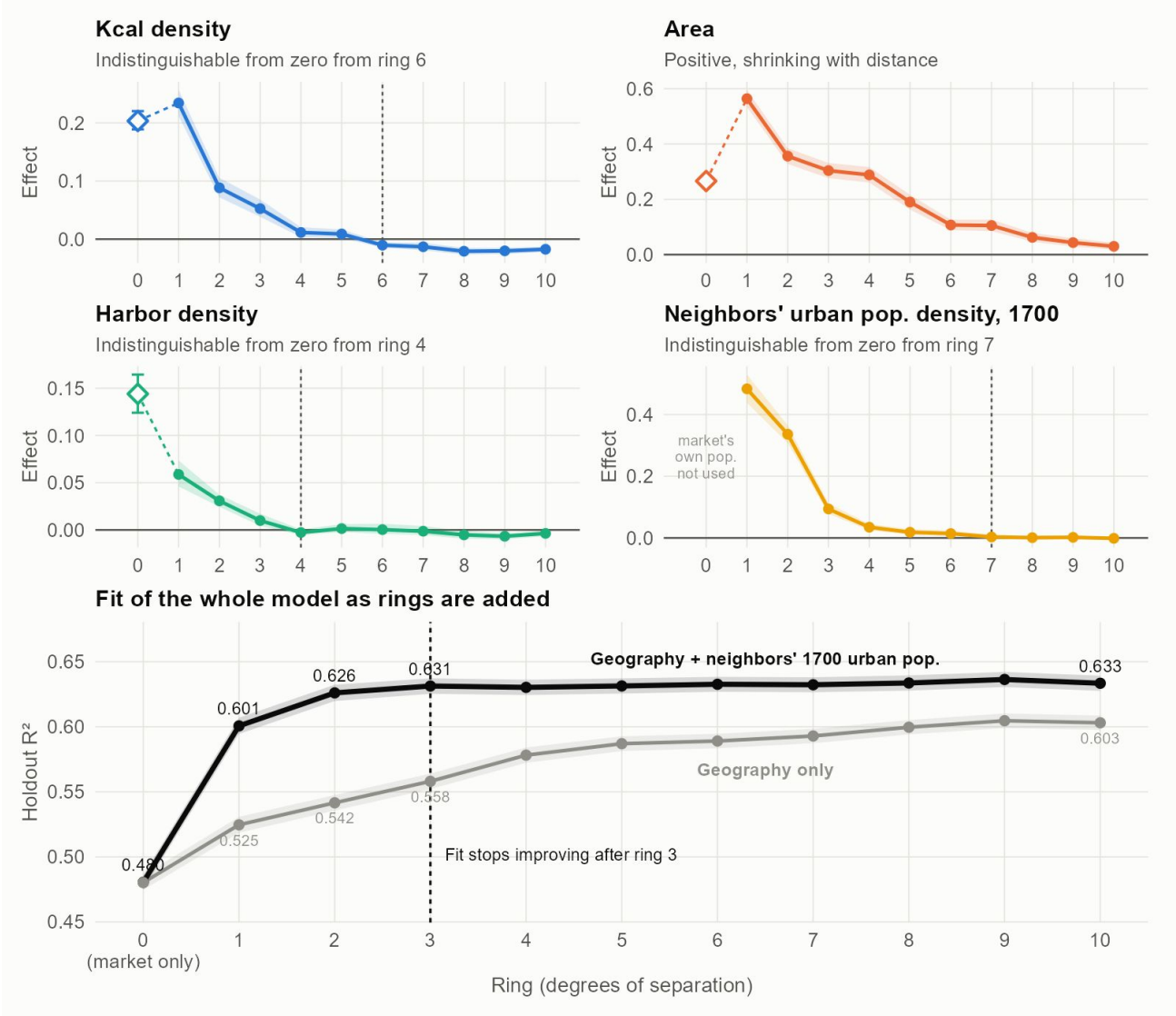
Red = the market; orange = first-degree lag (linked markets); yellow = second-degree lag; black dot = the market's origin point

Random Forest Results with Spatial and Time Lags

Model (each row adds to the one above)	Predictors	Training draws R^2	Training draws RMSE	Held-out draws R^2 (SD)	Held-out draws RMSE	Gain vs. previous step [95% CI]
Market only	3	0.477	0.339	0.477 (0.042)	0.340	–
+ First-degree lag: geography	6	0.521	0.324	0.521 (0.043)	0.325	+0.044 [+0.042, +0.045]
+ First-degree lag: urban density, 1700	7	0.598	0.297	0.599 (0.048)	0.297	+0.078 [+0.076, +0.080]
+ Second-degree lag: geography	10	0.601	0.296	0.603 (0.047)	0.296	+0.003 [+0.003, +0.004]
+ Second-degree lag: urban density, 1700	11	0.627	0.286	0.628 (0.046)	0.286	+0.026 [+0.025, +0.026]
+ Third-degree lag: geography	14	0.626	0.287	0.627 (0.046)	0.287	−0.001 [−0.001, −0.001]
+ Third-degree lag: urban density, 1700	15	0.629	0.285	0.631 (0.044)	0.285	+0.004 [+0.003, +0.004]

Outcome: $\ln(1 + \text{urban population per km}^2)$, 1800. Training = the forest on its own 4,800 training draws; held-out = the same 1,200 draws of 400 markets for every step.
Gain = paired mean difference in held-out R^2 from the previous step. First-degree lag = the markets linked to the market (a spatial lag); second-degree = the markets linked to those; and so on. Geography = their area, kcal density, harbor density; urban density, 1700 = their urban population per km^2 in 1700.

Successive spatial lags predict less of a market's urban population in 1800



Neighboring markets matter, but only nearby ones. Model fit stops improving after ring 3.

Model: A random forest predicts each market's urban population per km² in 1800 from its own kcal density, area and harbor density, plus the same variables for its neighbors, added one ring at a time (ring 1 = directly connected markets, ring 2 = their neighbors, ...).

Geography curves: Geography only. Each ring's effect is measured with the closer rings already in the model; ring 0 (diamond) = the market itself.

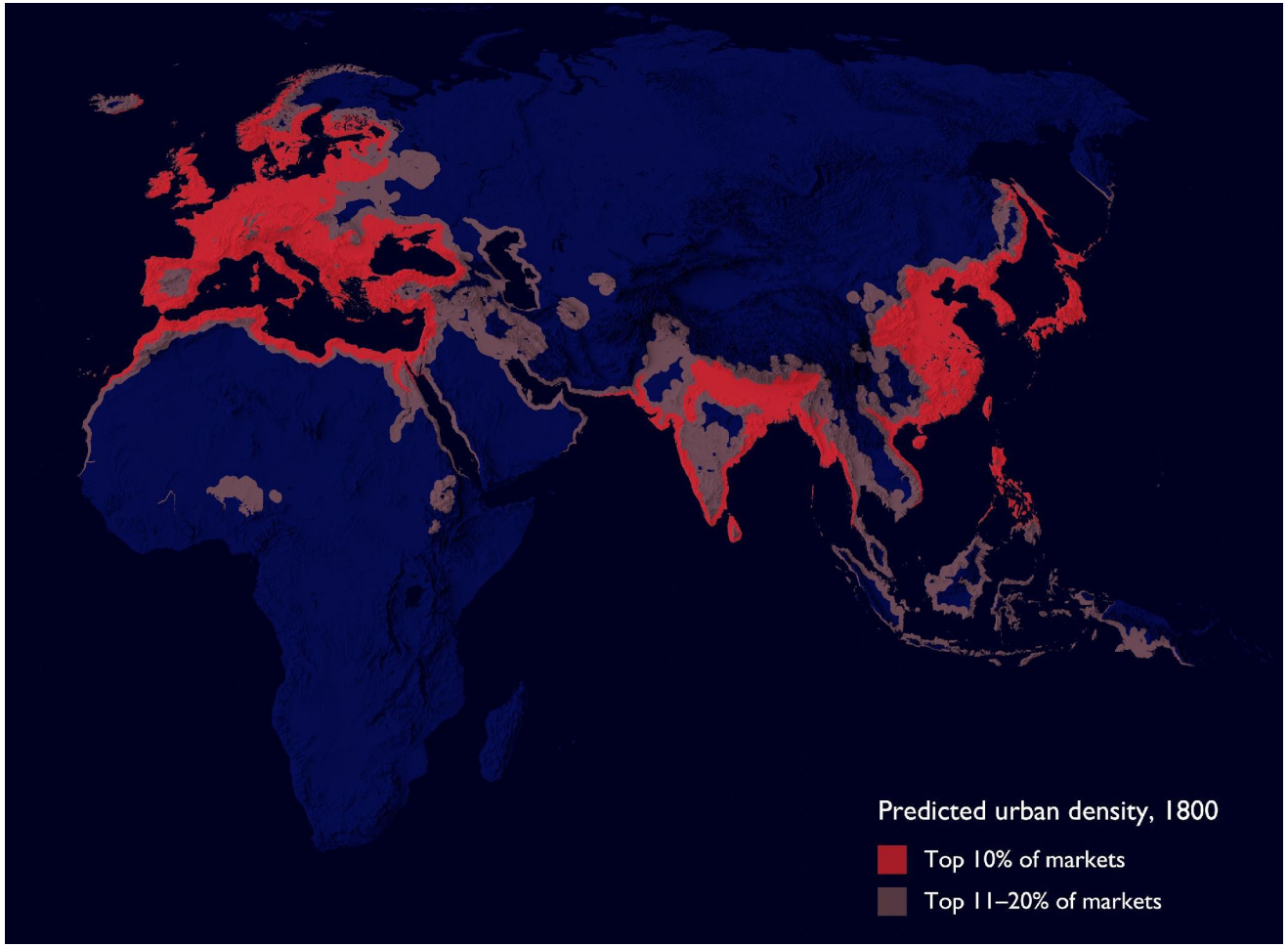
Urban population curve: Neighbors' urban population per km² in 1700, added ring by ring on top of all geography. The market's own population is never used.

Effect: Change in predicted log urban density as the ring's variable moves from its 10th to 90th percentile; bands = 95% intervals.

Fit: Holdout R² rises from 0.48 (market only) to 0.63 by ring 3, then stays flat.

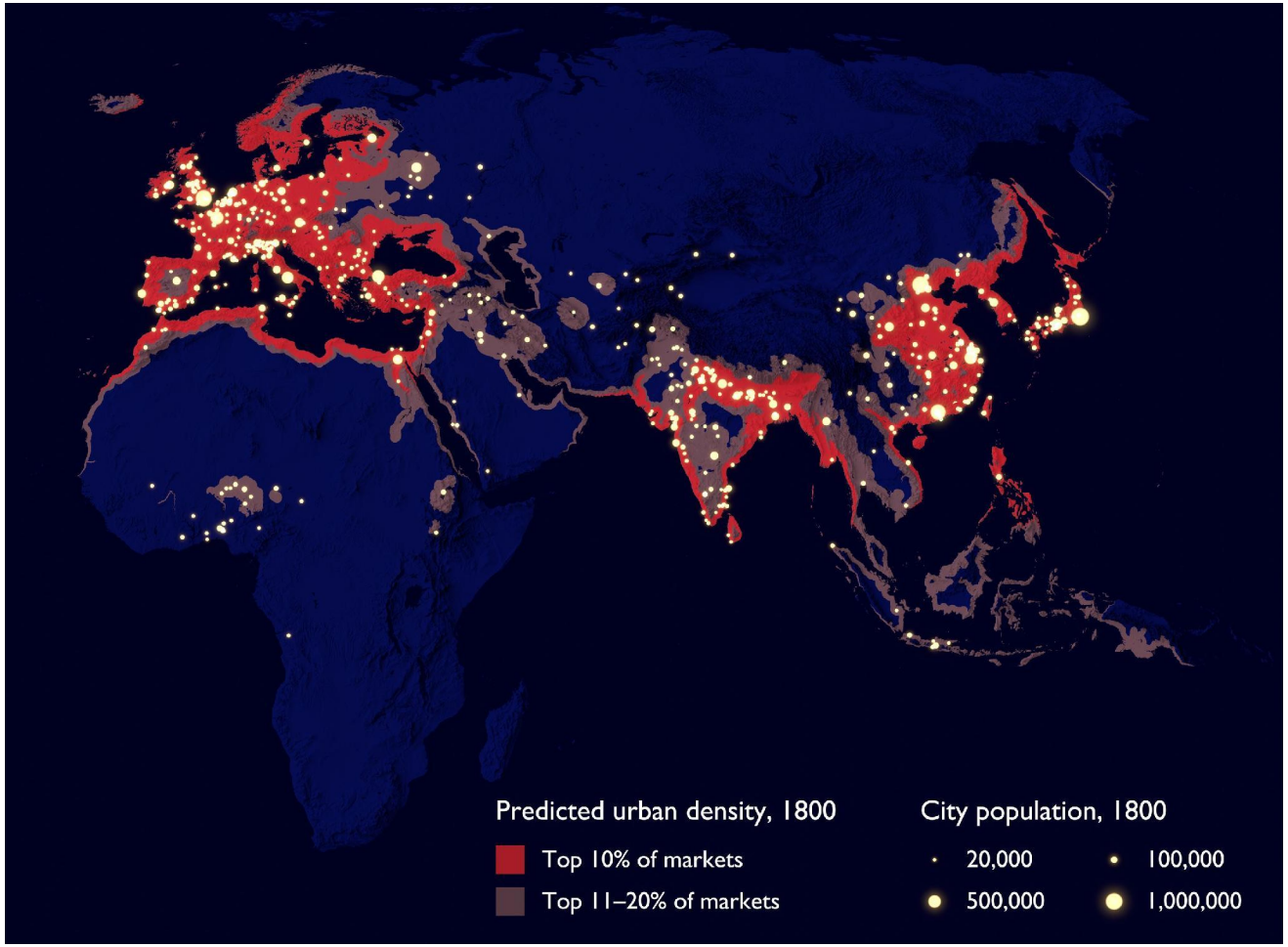
Sample: 41,435 Old World markets

Predicted regions of economic development: Random Forest model with first- to third-degree lags



Model: random forest, market + first- to third-degree lags (their geography and 1700 urban population density); out-of-sample (out-of-bag) predictions; holdout $R^2 = 0.63$

Predicted regions of economic development: Random Forest model with first- to third-degree lags compared to cities in 1800



Model: random forest, market + first- to third-degree lags (their geography and 1700 urban population density); out-of-sample (out-of-bag) predictions; holdout $R^2 = 0.63$

Model with lags, out-of-bag predictions

	Top 10%	Top 20%
% of Old World land	13.3%	26.3%
% of 1800 urban pop.	81.9%	92.1%
Location quotient	6.2	3.5

Full random forest results: base model and model with lags

	Market only		+ 1st–3rd-degree lags	
Panel A: Ensemble Specification				
Dependent variable	ln(1 + urban population per km^2 of market area, 1800)			
Predictors	ln area; ln caloric density; ln harbor density		market only + 1st–3rd-degree lags: geography and urban density 1700 (15)	
Potential markets	41,435			
Training / holdout draws	4,800 / 1,200 (400 markets per draw)			
Hyperparameters	`mtry`=3, min. node size=15, max depth=7		`mtry`=11, min. node size=15, max depth=7	
Panel B: Ensemble Performance	R^2	RMSE	R^2	RMSE
Cross-validated, training draws	0.477 (0.004)	0.340 (0.002)	0.629 (0.003)	0.286 (0.002)
Training draws, in sample	0.477 (0.042)	0.339 (0.026)	0.629 (0.044)	0.285 (0.024)
Held-out draws	0.477 (0.042)	0.340 (0.026)	0.631 (0.044)	0.285 (0.024)
OLS benchmark, held-out draws	0.237 (0.034)	0.410 (0.023)	0.531 (0.052)	0.321 (0.024)
Permutation null (100 shuffles of DV)	0.010 (max 0.043)	0.467	0.009 (max 0.032)	0.468
Panel C: Variable Importance (share of impurity reduction)				
Market geography (3 variables)	100.0%		9.8%	
Market area (log)	45.5%		5.4%	
Caloric density (log)	30.2%		2.9%	
Harbor density (log)	24.3%		1.4%	
1st–3rd-degree lags: geography (9 variables)	–		30.2%	
1st–3rd-degree lags: urban density 1700 (3)	–		60.1%	
Panel D: Spatial (LOBO) Cross-Validation				
Holdout R^2, range across nine dropped blocks	0.385 – 0.509		0.515 – 0.658	
Holdout R^2, Northwest block dropped	0.385 (0.049)		0.515 (0.053)	
Jackknife 95% CI for holdout R^2	[0.24, 0.71]		[0.38, 0.88]	

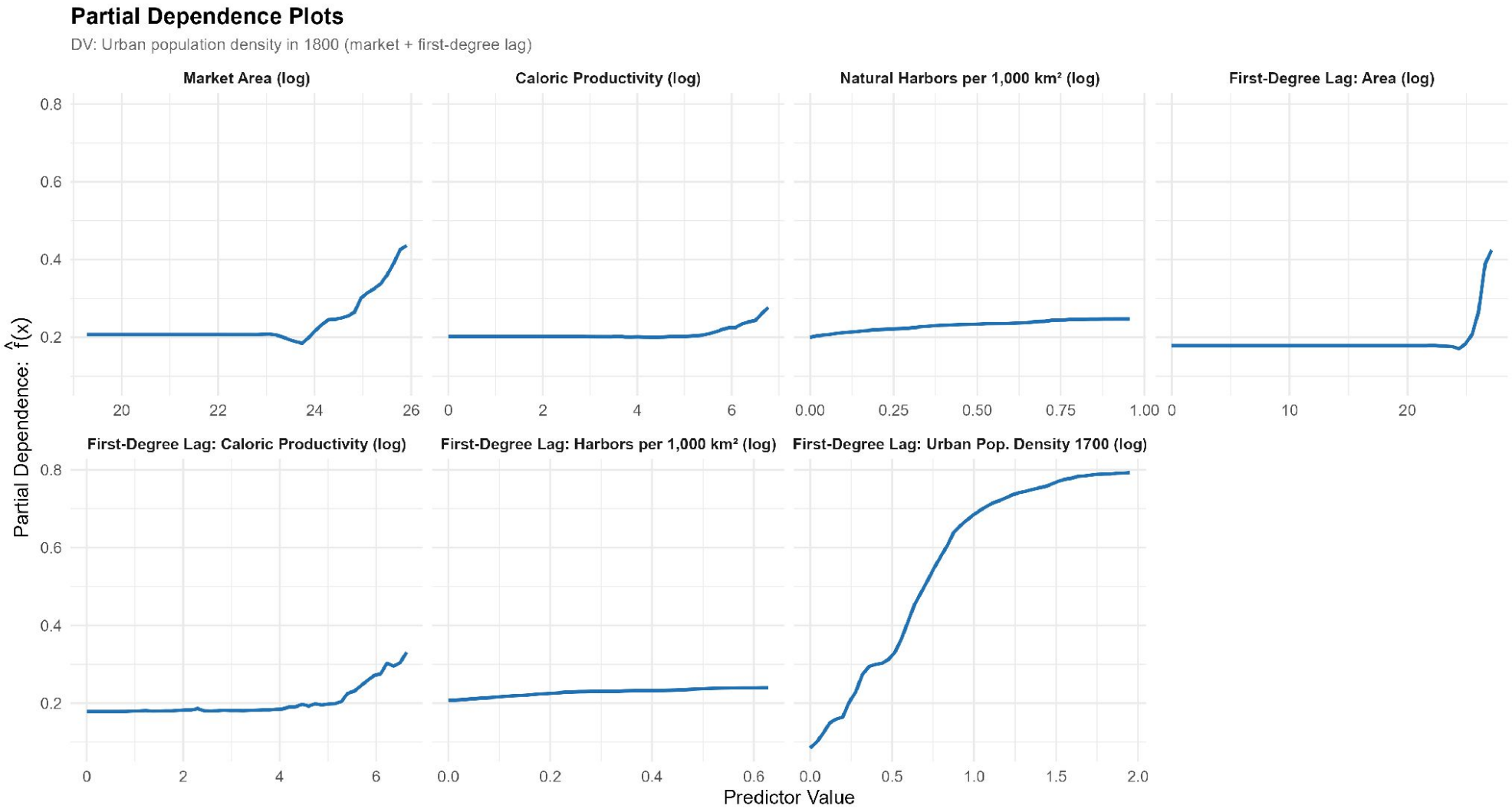
Mean (SD across draws; cross-validated: across 5 folds). Training draws, in sample: the 4,800-tree forest on its own training draws. OLS: one fit per training draw, coefficients averaged. Permutation: 100 shuffles, 4,800 trees, 1,200 holdout draws each.
Panel D (as the paper): each block of a 3×3 grid is dropped from training and holdout draws; R^2 on the remaining holdout markets. The market's own population is never a predictor.

Are our results sensitive to dropping regions? Base model and model with spatial lags

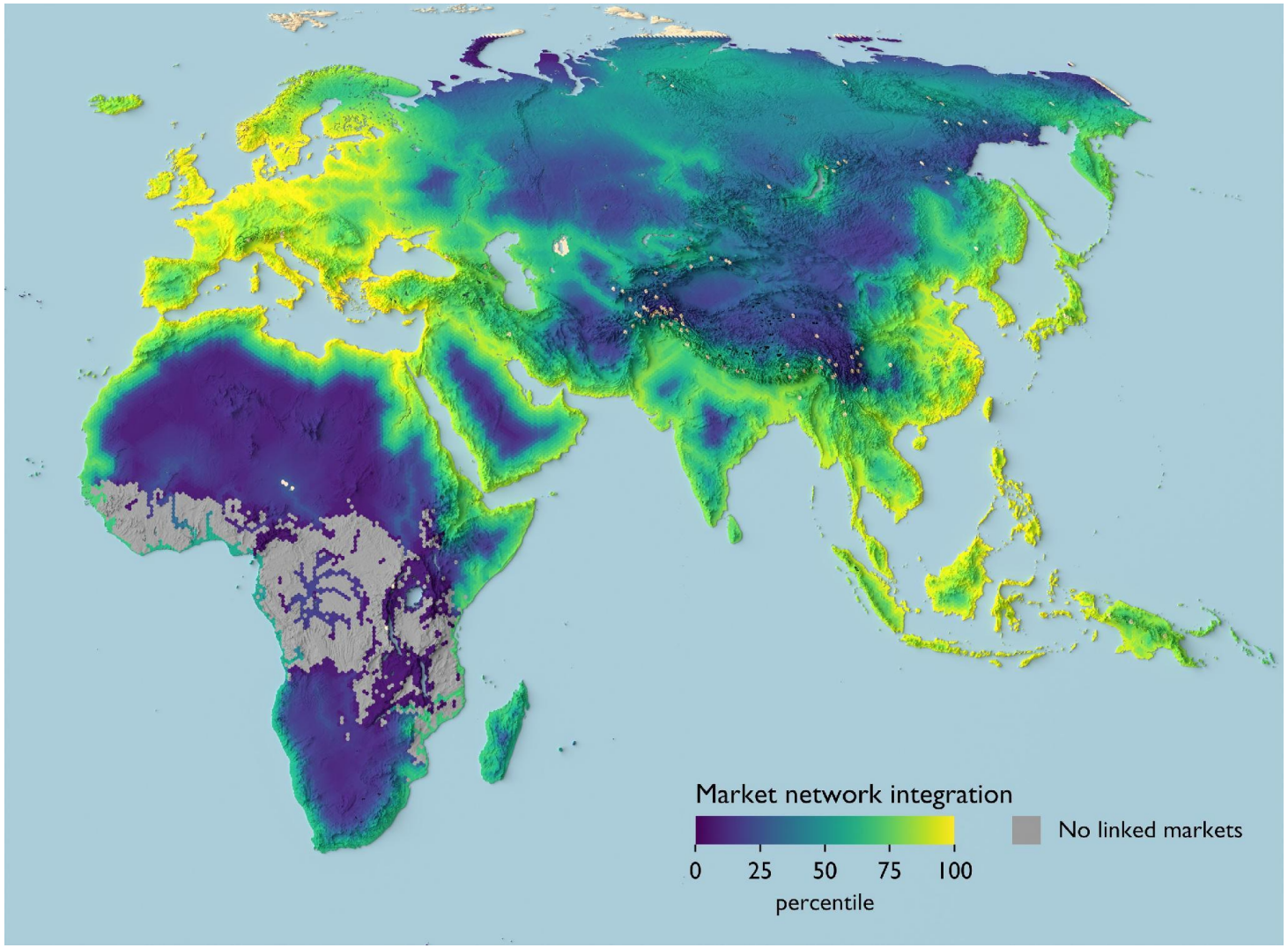
		Held-out R^2 (SD)	
Block dropped	Markets dropped	Market only	+ 1st–3rd-degree lags
Northwest	3,314	0.385 (0.049)	0.515 (0.053)
North	12,936	0.509 (0.045)	0.658 (0.044)
Northeast	6,393	0.442 (0.046)	0.613 (0.046)
West	5,203	0.474 (0.043)	0.633 (0.046)
Center	5,770	0.475 (0.045)	0.631 (0.048)
East	3,712	0.445 (0.047)	0.607 (0.051)
Southwest	1,113	0.487 (0.044)	0.629 (0.045)
South	1,286	0.504 (0.043)	0.630 (0.044)
Southeast	14	0.481 (0.043)	0.631 (0.044)
None (all markets)	0	0.477 (0.042)	0.631 (0.044)

As the paper's Table 4, Panel D: 3×3 grid over market centroids. The block is dropped from training and holdout draws; R^2 (SD across the 1,200 holdout draws) is scored on the remaining holdout markets, outside the block.

Partial Dependence Plots, Model with First-Degree Lag (7 Predictors)



Market Network Integration: How Connected Each Market Is to All Others



Extra Slides


1500