Digitized data

DSIER [/dɪˈzaɪər/] — Summer 2023

Julian Hinz

Bielefeld University

Today’s plan

- “Non-computable information”
- Lloyd’s shipping list: The Wind of Change: Maritime Technology, Trade, and Economic Development, Pascali (2017)
- Plantation records: “The Development Effects of the Extractive Colonial Economy: The Dutch Cultivation System in Java”, Dell and Olken (2020)
- Clay tablets: “Trade, Merchants, and the Lost Cities of the Bronze Age”, Barjamovic et al. (2019)

NON-COMPUTABLE INFORMATION

Non-computable information

- Standard digitization methods often fail to capture historical documents effectively

  → especially for less frequently used languages, scripts and settings

- Data may also be trapped in various types of images
- Text data contains a significant amount of non-computable information

Economics and data

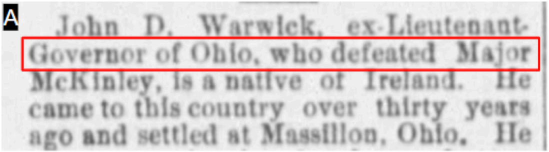
- Key economic questions necessitate disaggregated data: Misallocation, inequality, social mobility, welfare effects of trade

- Long-term digital disaggregated data uncommon

  → existing data predominantly originating from high resource contexts

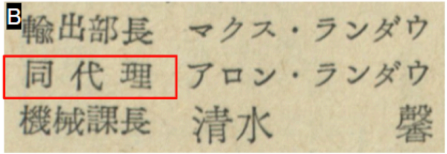
- Growing academic interest, also due to much better computing power and methods

Digitizing data



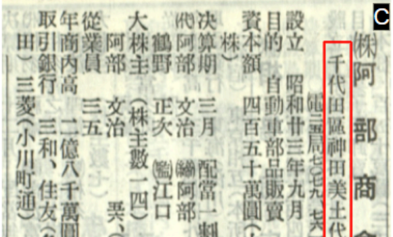
A

John D. Warwick, ex-Lieutenant-Governor of Ohio, who defeated Major McKinley, is a native of Ireland. He came to this country over thirty years ago and settled at Massillon, Ohio. He



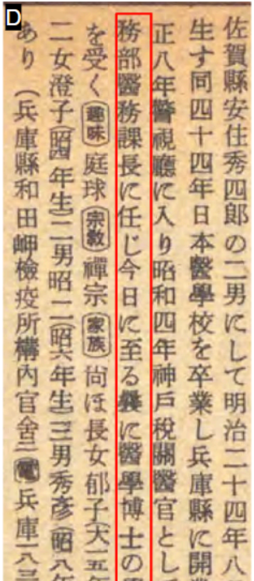
B

輸出部長　マクス・ランダウ
同代理　アロン・ランダウ
機械課長　清水　馨



C

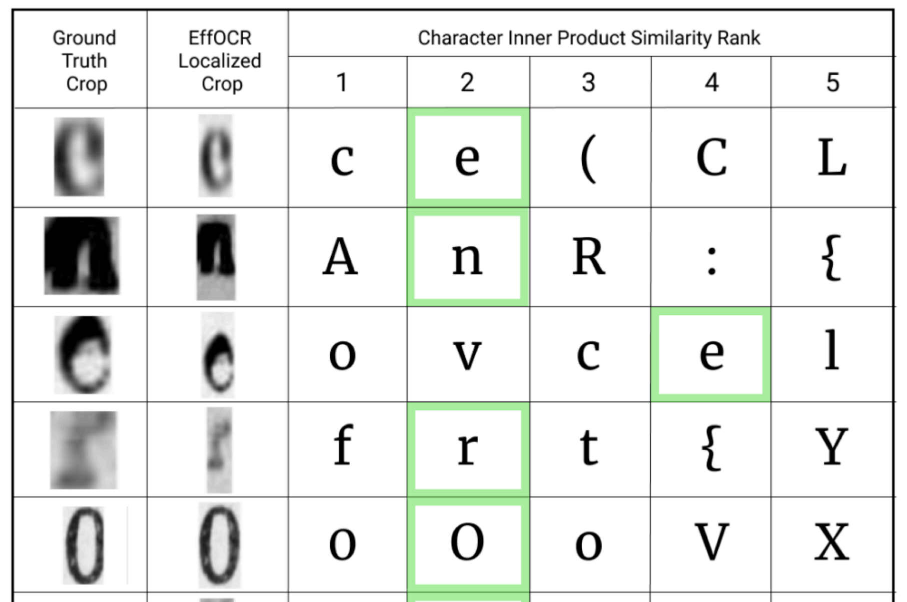
㈱阿部商
千代田區神田美土代
電三局七九七
設立　昭和廿三年九月
目的　自動車部品販賣
資本額　四百五十萬圓（
株）
決算期　三月　配當一割
㈹阿部文治　㈱阿部
鶴野正次　㈼江口
大株主（株主數一四）
阿部文治　巽、（
從業員　三五
年商内高　二億八千萬圓
取引銀行　三和、住友（
田）三菱（小川町通）



D

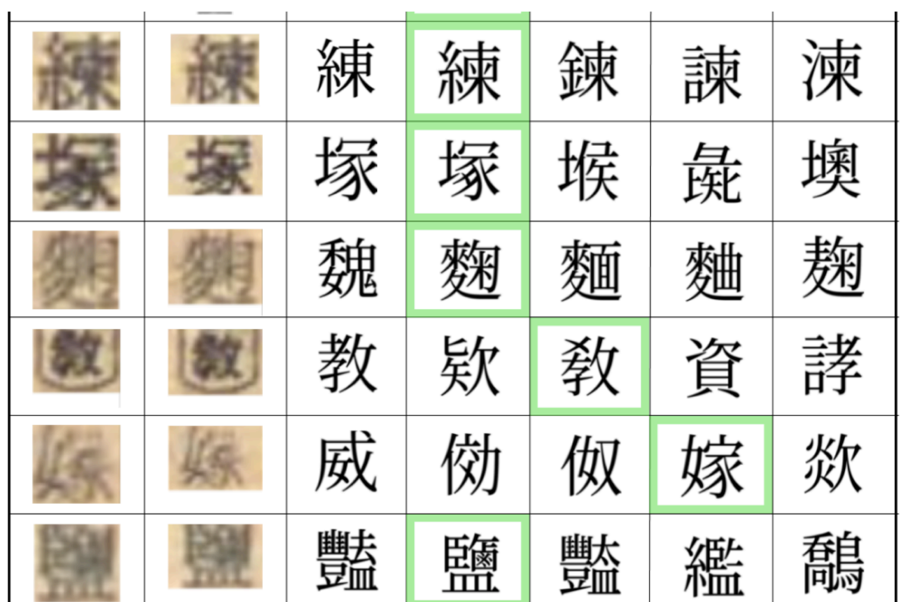
佐賀縣安住秀四郎の二男にして明治二十四年八
生す同四十四年日本醫學校を卒業し兵庫縣に開
正八年警視廳に入り昭和四年神戸税關醫官とし
務部醫務課長に任じ今日に至る傍ら醫學博士の
を受く趣味庭球宗教禪宗家族尚ほ長女郁子（大二五
二女澄子（昭四年生）二男昭一（昭六年生）三男秀彦（昭八
あり（兵庫縣和田岬檢疫所構内官舍三）電兵庫一八三

OCR



Ground Truth Crop	EffOCR Localized Crop	Character Inner Product Similarity Rank				
		1	2	3	4	5
		c	**e**	(	C	L
		A	**n**	R	:	{
		o	v	c	**e**	l
		f	**r**	t	{	Y
		o	**O**	o	V	X

OCR



		練	練	錬	諫	湅
		塚	塚	堠	㲋	墺
		魏	麹	麺	麯	麴
		教	欵	敎	資	誟
		威	㑁	㑀	嫁	欻
		豔	鹽	豓	繿	鷸

Accuracy

- OCR accuracy measured using character error rate (CER)

  → Levenshtein distance between recognized string and “ground truth”, normalized by length of “ground truth”

  → minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other

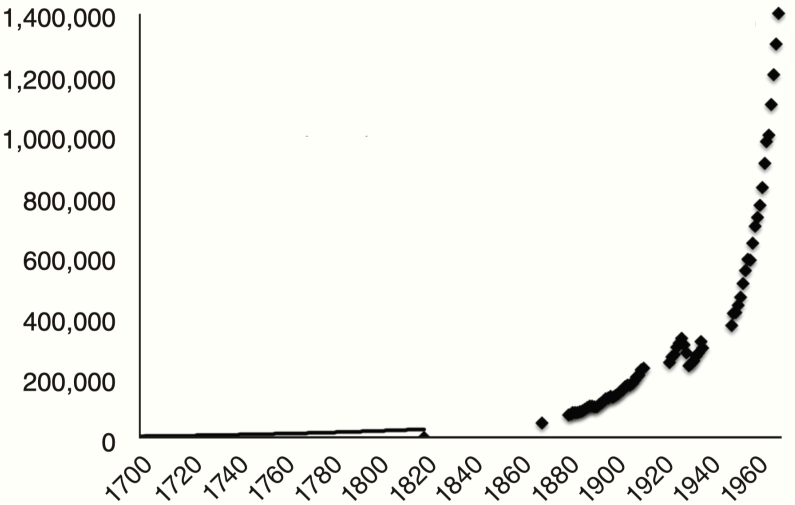
- CER of 0.5: mispredicting approximately half of characters

Software solutions

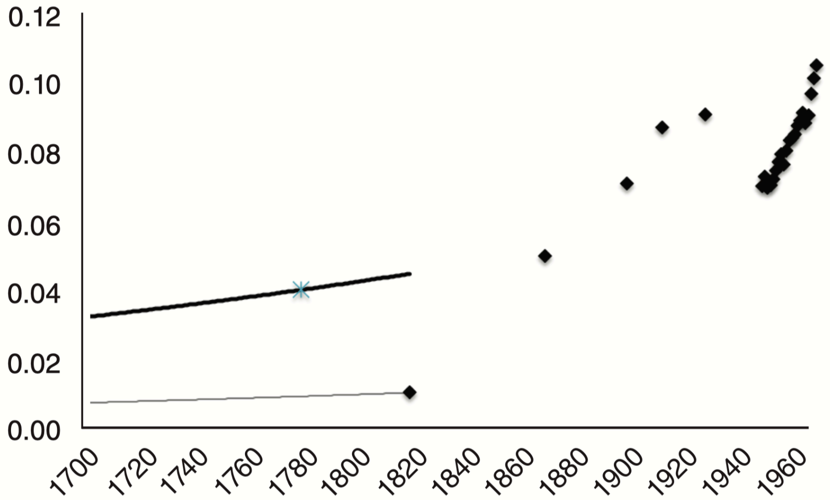
- Google Cloud Vision, Amazon Textract
- Baidu OCR (for Asian languages)
- Tesseract (bi-directional LSTM)
- Active research: EasyOCR (Shi et al., 2016), TrOCR (Li et al., 2021), PaddleOCR (Du et al., 2022), ...

WIND OF CHANGE – LLOYD’S LIST

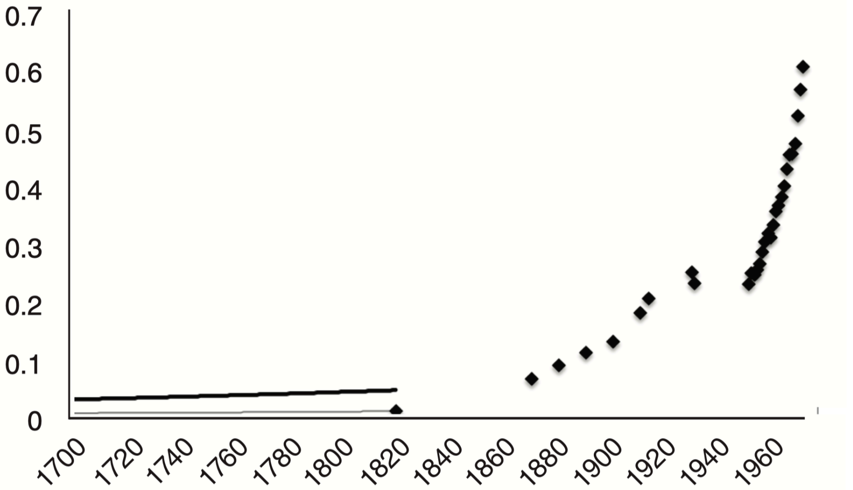
Panel A. World exports at constant prices (million 1990 dollars)



Panel B. World export-to-GDP ratio



Panel C. World export-to-population ratio



Idea

- 1870–1913 first era of trade globalization
- How did the increase in trade affect economic development?
- Causal mechanism: steamship vs. sailing
- asymmetric change in trade distances among countries
- steamship reduced shipping costs and time heterogeneously across countries and trade routes

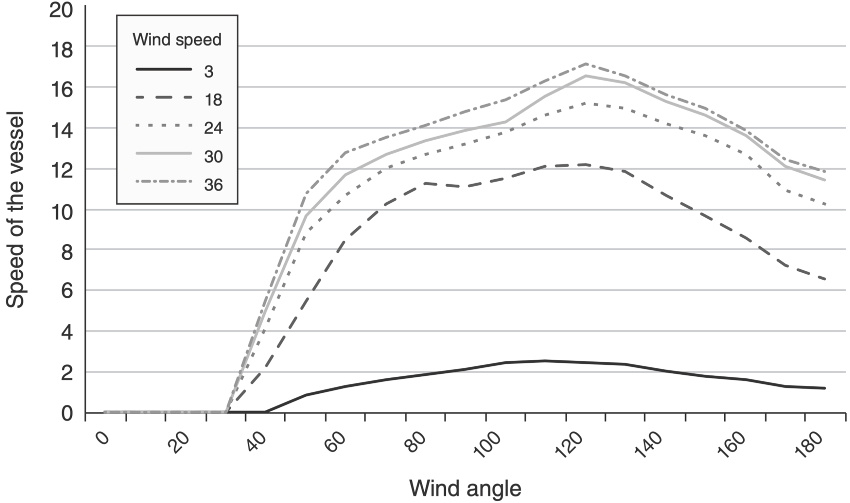
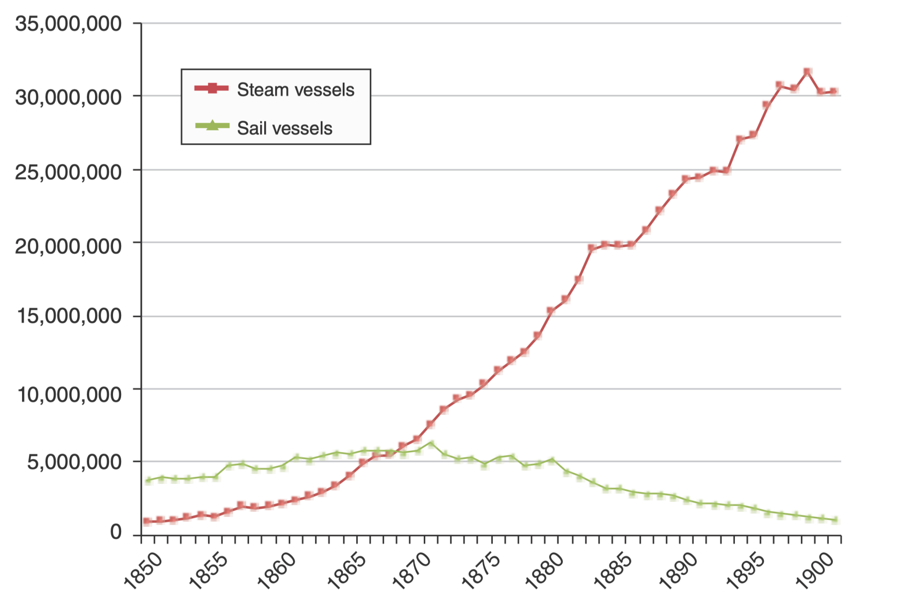


Figure 2. Polar Diagram of a Sailing Vessel: The Clipper in 1860

Digitized data

- Three novel datasets from 1850 to 1900
- First dataset: shipping times across 16,000 country pairs
- Second dataset: 23,000 bilateral trade observations, 1,000 distinct country pairs

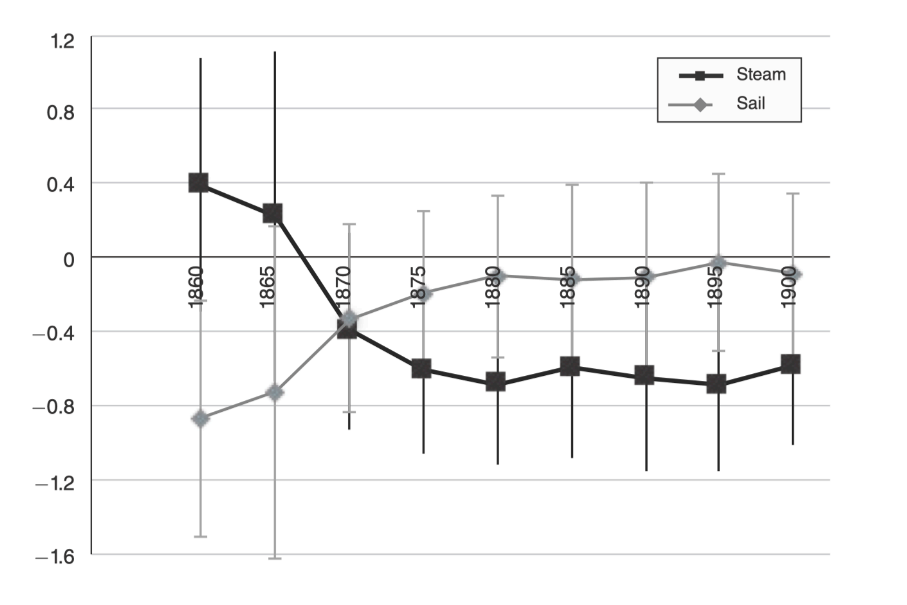
  → Sectoral-level export data for 37 countries
- Third dataset: freight rates across 291 shipping routes


35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
Steam vessels
Sail vessels
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900

Effect on trade and GDP

- Impact of steamship on world trade volumes

  → Reduction in geographical isolation measured by average shipping time
- Country-level regressions estimate impact of change in isolation

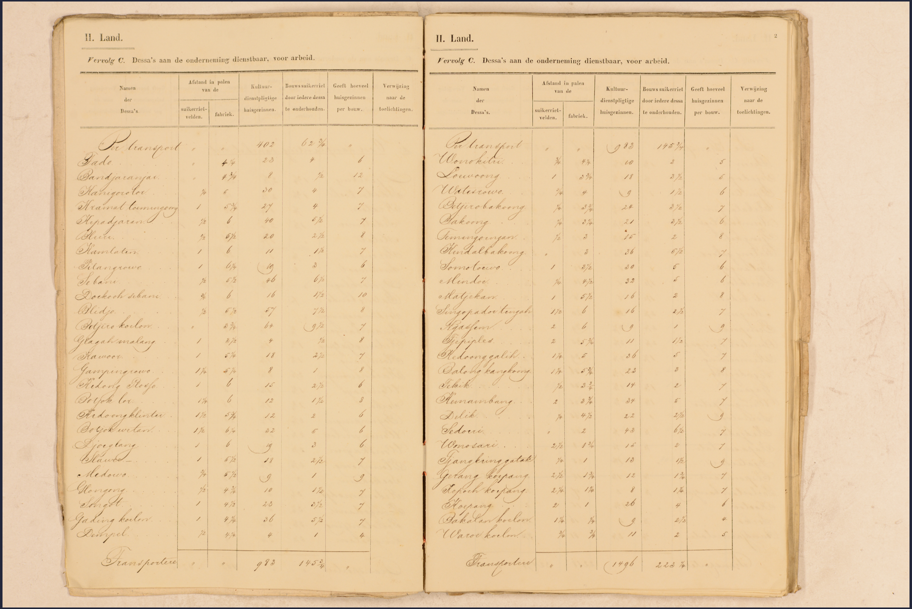

1.2
0.8
0.4
0
−0.4
−0.8
−1.2
−1.6
1860
1865
1870
1875
1880
1885
1890
1895
1900
Steam
Sail

Findings

- Rich countries did not benefit on average
- Similar impact of trade on agricultural and non-agricultural countries
- Institutions might reflect economic development differences

red

- Investigates Dutch Cultivation System impacts
- Farmers forced to cultivate export crops: Sugar
- Areas near factories more industrialized today
- Residents near factories have higher education



II. Land.

Vervolg C. Dessa's aan de onderneming dienstbaar, voor arbeid.

Namen der Dessa's.	Afstand in palen van de		Kultuur-dienstpligtige huisgezinnen.	Bouws suikerriet door iedere dessa te onderhouden.	Geeft hoeveel huisgezinnen per bouw.	Verwijzing naar de toelichtingen.
	suikerriet-velden.	fabriek.				
Per transport	"	"	402	62 3/4	"	
Dado	"	4 1/4	23	4	6	
Bandjaranjar	"	4 3/4	8	1/2	12	
Kanigorolor	1/4	5	30	4	7	
Kramat toemengoeng	1	5 3/4	27	4	7	
Kepadjaran	1/2	6	40	5 1/2	7	
Bleer	1/2	5 1/2	20	2 1/2	8	
Kamlaten	1	6	11	1 1/2	7	
Pilangrowo	1	6 1/4	19	3	6	
Sebani	1/2	5 1/2	46	6 1/2	7	
Doekoeh sebani	3/4	6	16	1 1/2	10	
Blidjo	1/2	5 1/2	57	7 1/2	8	
Petjiro koelon	"	3 3/4	64	9 1/2	7	
Glagah malang	1	3 1/2	4	1/2	8	
Kawoor	1	5 1/4	18	2 1/2	7	
Gampingrowo	1 1/4	5 1/4	8	1	8	
Kedong Rosso	1	6	15	2 1/2	6	
Potjok lor	1 1/4	6	12	1 1/2	8	
Kedoongklinter	1 1/2	5 3/4	12	2	6	
Potjok wetan	1 1/2	6 1/4	32	5	6	
Djoeglang	1	6	19	3	6	
Slawe	1	5 1/2	18	2 1/2	7	
Medowo	3/4	5 1/2	9	1	9	
Glongong	1/2	4 3/4	10	1 1/2	7	
Songot	1	4 1/2	23	3 1/2	7	
Gading koelon	1	4 1/4	36	5 1/2	7	
Dempel	1/2	4 1/4	4	1	4	
Transportere	"	"	983	145 3/4	"	

II. Land.

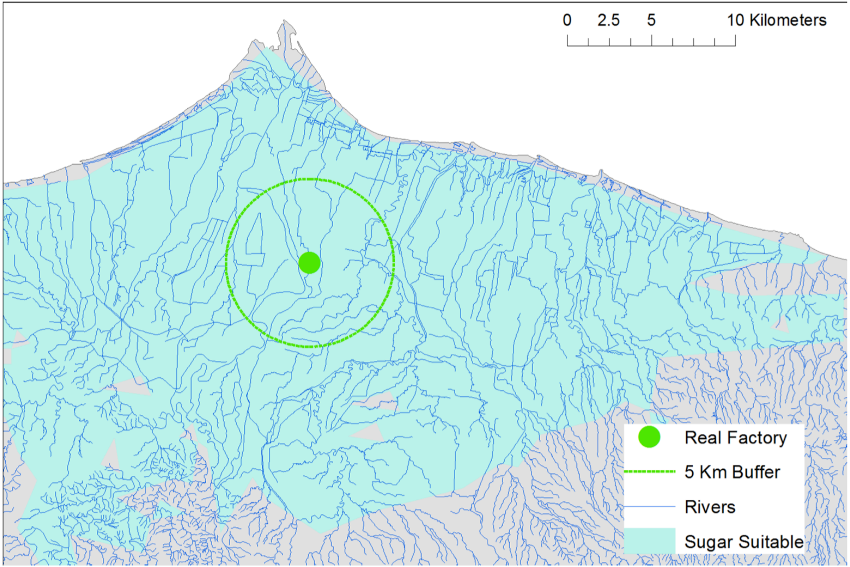
2

Vervolg C. Dessa's aan de onderneming dienstbaar, voor arbeid.

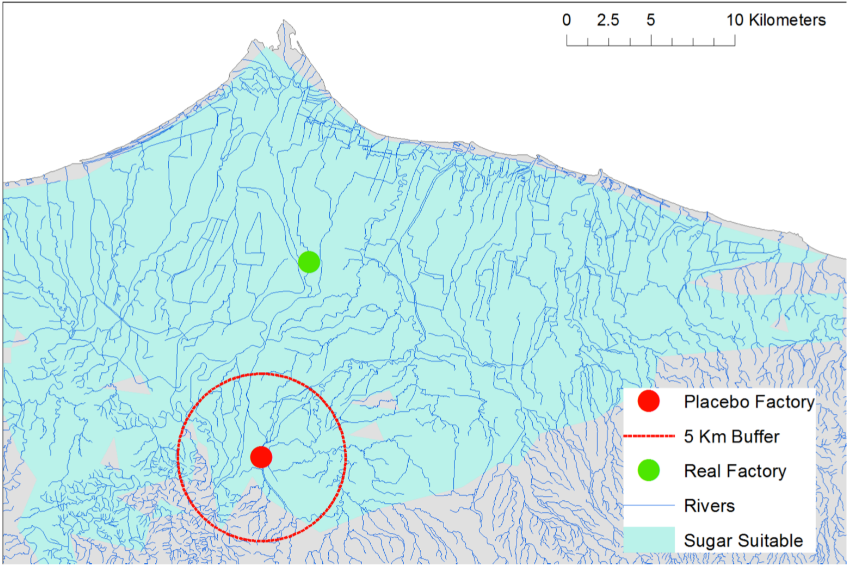
Namen der Dessa's.	Afstand in palen van de		Kultuur-dienstpligtige huisgezinnen.	Bouws suikerriet door iedere dessa te onderhouden.	Geeft hoeveel huisgezinnen per bouw.	Verwijzing naar de toelichtingen.
	suikerriet-velden.	fabriek.				
Per transport	"	"	983	145 3/4	"	
Wonokitri	3/4	4 1/4	10	2	5	
Loewoong	1	3 3/4	18	3 1/2	5	
Watesrowo	3/4	4	9	1 1/2	6	
Petjirobakoong	1/4	3 3/4	24	3 1/2	7	
Bakoong	1/4	3 1/4	21	3 1/2	6	
Pemengoengan	1/2	3	15	2	8	
Kendalbakoong	"	3	36	5 1/2	7	
Somoloewo	1	3 1/2	30	5	6	
Mindoe	1/4	4 1/2	32	5	6	
Matjekan	1	5 1/2	16	2	8	
Singopadoe tengah	1 1/2	6	16	2 1/2	7	
Ngasem	2	6	9	1	9	
Tjepiples	2	5 3/4	11	1 1/2	7	
Kedoonggalih	1 1/4	5	36	5	7	
Balong kangkoong	1 1/4	5 3/4	23	3	8	
Tekik	1/2	3 3/4	14	2	7	
Kemambang	2	3 3/4	34	5	7	
Delik	1/4	4 1/2	22	2 1/2	9	
Sedoeri	"	2	43	6 1/2	7	
Wonosari	2 1/2	1 3/4	15	2	7	
Tjangkringgatak	1/4	1	13	1 1/2	9	
Gerang koepang	2 1/2	1 3/4	12	1 3/4	7	
Kepoeh koepang	2 1/4	1 1/4	8	1 1/4	7	
Koepang	2	1	26	4	6	
Bakalan koelon	1 1/4	1/4	9	2 1/2	4	
Waroe koelon	3/4	3/4	11	2	5	
Transportere	"	"	1496	223 1/4	"	

- Data combines historical and contemporary sources
- Traces long-term impacts of Cultivation System
- Geographic distance to factories measures exposure
- Uses contemporary data for long-term impacts

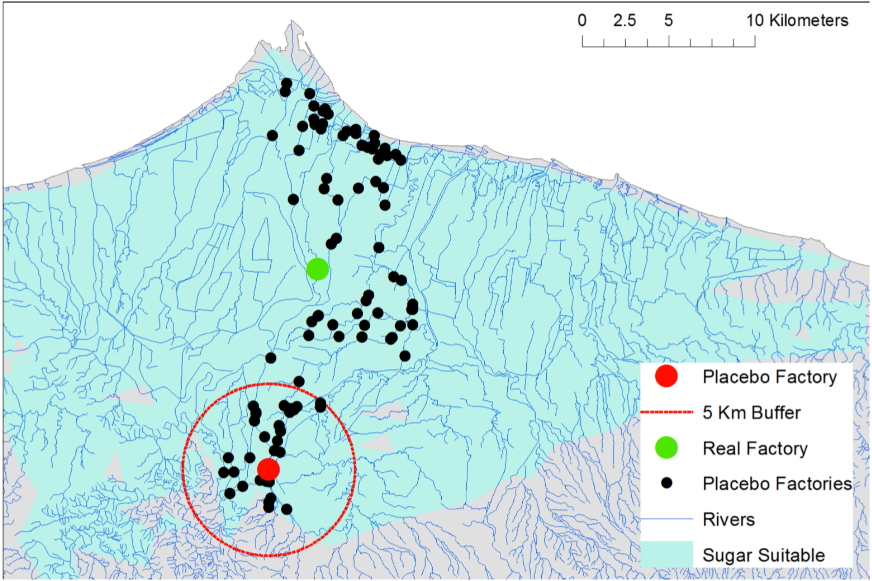
(a) Real Factory



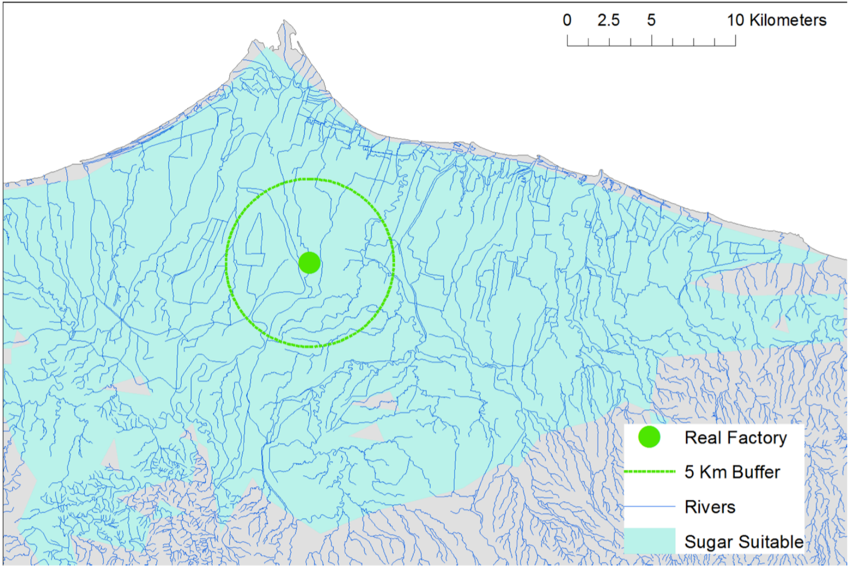
(b) Placebo Factory Suitability



(c) Placebo Factories



(a) Real Factory



Discussion

- Study focuses on specific colonial institution
- Findings may not generalize to other institutions
- Pre-existing differences between areas not ruled out
- Unobserved factors could influence results

red

- Novel approach to estimate the locations of lost cities from the Bronze Age
- Structural gravity model to estimate the locations of lost cities based on trade data from ancient texts
- Ancient city sizes are persistent, meaning that large ancient cities tend to be located at or near large modern cities

Data and its Novelty

- Sample of 9,728 digitized texts and approximately 2,000 additional non-digitized texts
- ancient texts to extract information about trade routes and city locations
- data mentions 79 unique settlements, with the analysis restricted to 25 Anatolian cities in Turkey

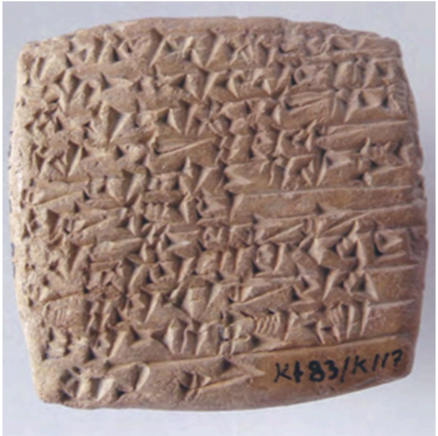
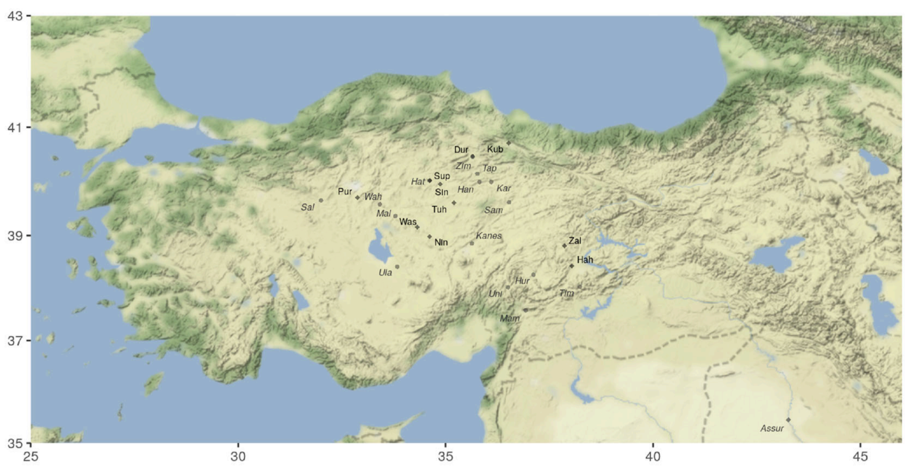
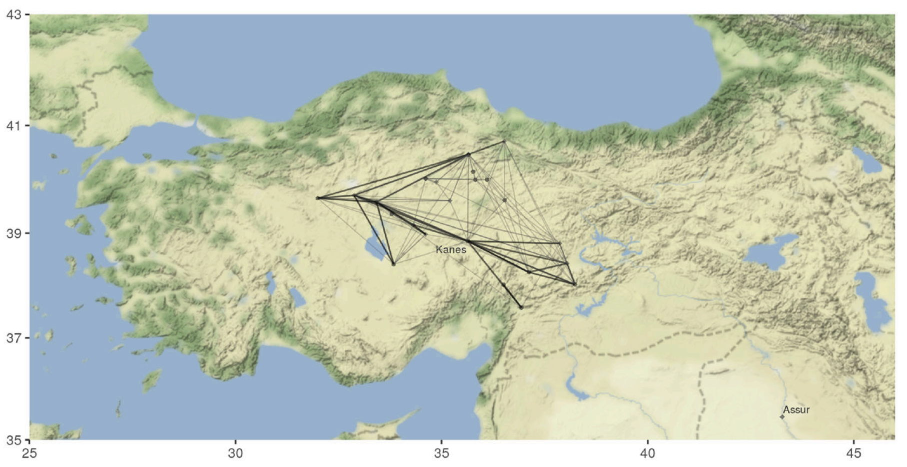


FIGURE I

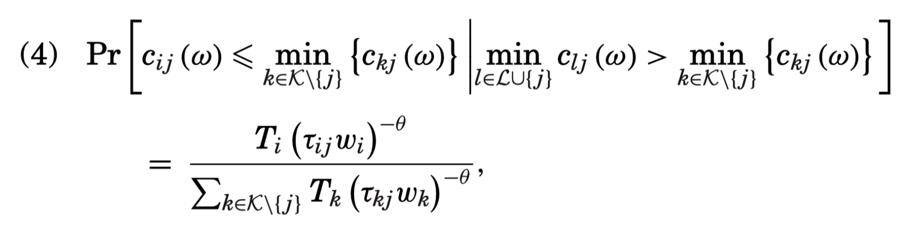
Tablet Kt 83-k 117


43
41
39
37
35
25
30
35
40
45
Dur
Kub
Zlm
Tap
Sup
Hat
Han
Kar
Sln
Pur
Wah
Sal
Tuh
Sam
Mal
Was
Nln
Kanes
Zal
Hah
Ula
Hur
Unι
Tlm
Mam
Assur


43
41
39
37
35
25
30
35
40
45
Kanes
Assur

Empirical Strategies

- structural gravity model to estimate the locations of lost cities
- detailed data on the topography of the entire region surrounding Anatolia to compute travel times



$$\Pr\left[c_{ij}\left(\omega\right) \leqslant \min_{k\in\mathcal{K}\setminus\{j\}} \left\{c_{kj}\left(\omega\right)\right\} \,\middle|\, \min_{l\in\mathcal{L}\cup\{j\}} c_{lj}\left(\omega\right) > \min_{k\in\mathcal{K}\setminus\{j\}} \left\{c_{kj}\left(\omega\right)\right\}\right] = \frac{T_i\left(\tau_{ij}w_i\right)^{-\theta}}{\sum_{k\in\mathcal{K}\setminus\{j\}} T_k\left(\tau_{kj}w_k\right)^{-\theta}}, \tag{4}$$

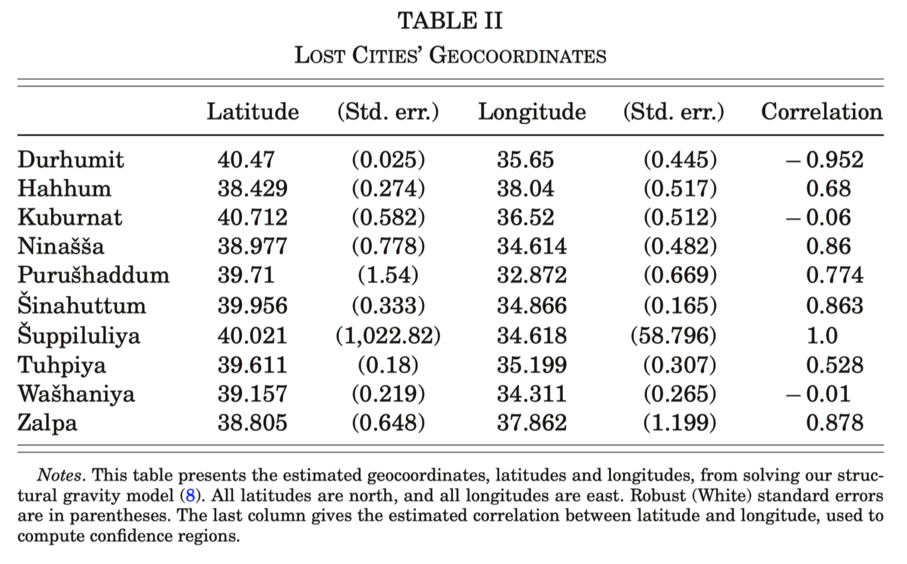
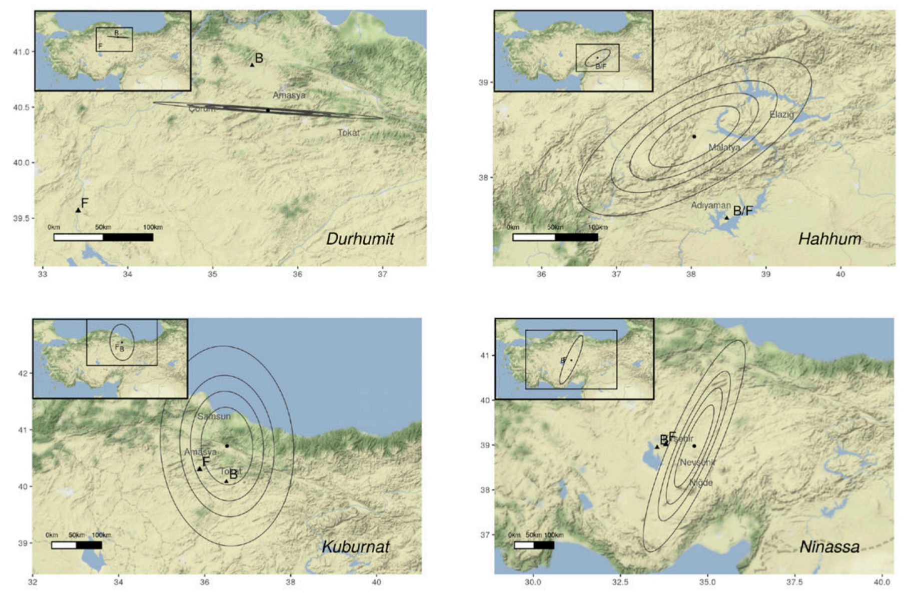


TABLE II

LOST CITIES' GEOCOORDINATES

	Latitude	(Std. err.)	Longitude	(Std. err.)	Correlation
Durhumit	40.47	(0.025)	35.65	(0.445)	– 0.952
Hahhum	38.429	(0.274)	38.04	(0.517)	0.68
Kuburnat	40.712	(0.582)	36.52	(0.512)	– 0.06
Ninašša	38.977	(0.778)	34.614	(0.482)	0.86
Purušhaddum	39.71	(1.54)	32.872	(0.669)	0.774
Šinahuttum	39.956	(0.333)	34.866	(0.165)	0.863
Šuppiluliya	40.021	(1,022.82)	34.618	(58.796)	1.0
Tuhpiya	39.611	(0.18)	35.199	(0.307)	0.528
Wašhaniya	39.157	(0.219)	34.311	(0.265)	– 0.01
Zalpa	38.805	(0.648)	37.862	(1.199)	0.878

Notes. This table presents the estimated geocoordinates, latitudes and longitudes, from solving our structural gravity model (8). All latitudes are north, and all longitudes are east. Robust (White) standard errors are in parentheses. The last column gives the estimated correlation between latitude and longitude, used to compute confidence regions.



Durhumit
B
F
Amasya
Çorum
Tokat
41.0
40.5
40.0
39.5
33
34
35
36
37
0km 50km 100km
Hahhum
Elazığ
Malatya
Adıyaman
B/F
39
38
36
37
38
39
40
0km 50km 100km
Kuburnat
Samsun
Amasya
Tokat
F
B
42
41
40
39
32
34
36
38
40
0km 50km 100km
Ninassa
B
F
Kırşehir
Nevşehir
Niğde
41
40
39
38
37
30.0
32.5
35.0
37.5
40.0
0km 50km 100km

Discussion

- may be a systematic bias for larger cities to be more or less likely to have been unambiguously located by historians
- large ancient cities may never be discovered, as they lay buried under modern cities
- data does not observe internal transactions, a purchase in a city of a good sourced locally in the same city