経済学者によるデータ発掘と
その研究利用、
非統計資料に着目して

神戸大学大学院経済学研究科講師
山﨑潤一

自己紹介

- 2016年から神大経済、ロンドンスクールオブエコノミクス（いわば一橋大学？）PhD
- 経済学の中でもデータを用いた実証研究をメインにしています
- 主な関心は開発経済、都市経済、政治経済です。

今日のお題

- 経済学者が使うデータ: 圧倒的に統計が多い（と思う）
- しかしながら、独自資料/データを用いた研究も重要
    - 政府統計では捕捉できない経済活動/人間行動が捕捉できる
        - 集計/非集計の壁（細かさ）、範囲の問題
- 今日は最近の研究から実際の例を紹介します。
    - 例1: 文書
    - 例2: 画像（地図、挿絵
- 生データとしての書籍、資料の学術的/社会的価値に思いを馳せる…!

1. 文書

文書の研究例

- Hansen et al. (2017) “Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach” Quarterly Journal of Economics
- アメリカの連邦公開市場委員会の議事録を分析
    - ある時を境に、公開が前提になり、それより前の議事録も公開されることとなった
- 公開が前提になると、評判や自分のキャリアなどを気にして人の行動はどう変わるのだろうか？
    - 努力は増えそう
    - 自分の専門に自信があれば、より目立つような発言をしたりするが、そうでなければ目立たないようになるだけ

着目した変数

- テキストから単語数などをカウント、特に具体的な数字を使った発言: 努力の代理変数として重要
- 同時に出てきやすい単語のグループからトピックに分類-> 個人が話しているトピックが全体のトピックとどれくらい違うかを計測: 目立つ行動を捕捉

Topic0[1]	product	increas	wage	price	cost	labor	rise	acceler	inflat	pressur	trend	compens
Topic1[1,2]	growth	slow	economi	continu	expans	strong	trend	inflat	will	recent	slowdown	moder
Topic2[2]	inflat	expect	core	measur	higher	path	slack	gradual	continu	remain	view	suggest
Topic3[1]	percent	year	quarter	growth	month	rate	last	next	state	averag	california	employ
Topic4	number	data	look	chang	measur	use	point	show	revis	estim	gdp	actual
Topic5[1,2]	polici	inflat	monetarpol	need	time	can	monetari	move	tighten	view	action	believ
Topic6[2]	rate	term	expect	real	lower	increas	rise	level	declin	short	nomin	year
Topic7	statement	word	chang	meet	languag	discuss	issu	want	read	sentenc	view	use
Topic8[2]	chairman	support	mr	direct	recommend	agre	asymmetr	prefer	symmetr	move	toward	favor
Topic9[1]	employ	continu	growth	job	nation	region	seem	state	manufactur	greenbook	busi	bit
Topic10	dollar	unitedstates	export	countri	import	foreign	japan	growth	abroad	trade	develop	currenc
Topic11	model	use	simul	shock	effect	scenario	nairu	differ	rule	chang	baselin	altern
Topic12[2]	risk	may	balanc	seem	side	uncertainti	possibl	economi	probabl	reason	upsid	much
Topic13	forecast	greenbook	staff	project	differ	assumpt	littl	assum	somewhat	lower	end	period
Topic14	period	committe	consist	econom	run	maintain	futur	read	slightli	stabil	expect	develop
Topic15	invest	incom	spend	capit	household	consum	busi	hous	consumpt	sector	stock	stockmarket
Topic16[1]	month	report	increas	survey	expect	indic	remain	continu	last	recent	data	activ
Topic17[1]	project	forecast	year	quarter	expect	will	percent	revis	anticip	growth	next	recent
Topic18	question	ask	issu	let	want	answer	rais	discuss	don	start	without	okay
Topic19	peopl	talk	lot	much	comment	around	differ	number	realli	look	thing	hear
Topic20	presid	ye	governor	parri	stern	vice	hoenig	minehan	kelley	jordan	moskow	mcteer
Topic21	move	can	evid	signific	stage	inde	will	issu	economi	may	quit	clearli
Topic22[2]	chairman	thank	mr	time	meet	laughter	comment	let	will	point	call	may
Topic23[1]	year	panel	line	shown	right	chart	expect	project	percent	middl	left	next
Topic24	district	nation	area	continu	sector	construct	manufactur	report	activ	region	economi	remain
Topic25	know	someth	happen	right	thing	want	look	sure	can	realli	anyth	els
Topic26[1,2]	polici	might	committe	market	may	tighten	eas	risk	action	staff	possibl	potenti
Topic27	year	continu	product	price	level	industri	will	sale	increas	auto	last	district
Topic28[1]	inventori	product	sale	level	order	will	sector	come	good	quarter	much	adjust
Topic29	price	oil	increas	energi	effect	import	suppli	product	demand	will	market	oilprices
Topic30	term	might	point	can	sens	run	short	probabl	time	longer	tri	someth
Topic31	seem	may	time	certainli	bit	littl	quit	much	far	perhap	better	might
Topic32	money	aggreg	borrow	seem	rang	reserv	rate	target	time	altern	suggest	million
Topic33[2]	move	market	point	will	fundsrate	rate	basispoints	need	fed	today	basi	time
Topic34[1]	report	busi	compani	year	contact	firm	sale	worker	expect	plan	director	industri
Topic35	will	fiscal	ta	budget	cut	govern	effect	billion	state	spend	deficit	year
Topic36	will	economi	world	rather	problem	believ	can	situat	much	seem	view	good
Topic37	realli	look	side	thing	lot	problem	concern	littl	pretti	situat	kind	much
Topic38	bank	credit	market	loan	financi	debt	lend	fund	concern	financ	problem	spread
Topic39[1,2]	economi	weak	recoveri	recess	confid	eas	neg	econom	will	turn	declin	period

出典: Hansen et al. (2017)

結果

- 景気状況を認識する会合では、評判を気にしないといけない若い人ほど:
  - 具体的な数字を使うようになった
  - 全体の平均からは違うようなトピックを話すようになった
- 政策担当者の行動原理を検証した貴重な論文

大学のシラバスさえも

- Biasi and Ma (2021) “The Education-Innovation Gap”, Mimeo

- 全米の大学のシラバスをOpen Syllabus Projectから収集、その内容と該当分野の年代別トップジャーナル論文を比較

- 新しい研究成果を反映しているシラバスはどこの大学にあるのか？誰のシラバスなのか？

大学のシラバスさえも

- 研究成果を出している/研究費の多い教員のシラバスは新しい論文により”近い”
  - コサイン類似度でシラバスと論文を比較

- 新しい研究成果に基づくシラバスを使っている大学では、学生の卒業率や賃金などが高くなるパターンあり

- シラバス: 教育の中身を計測するための貴重な資料

2. 画像

画像の研究例

- Yamasaki, Nakajima, and Teshima (2021) “From Samurai to Skyscrapers: How Historical Lot Fragmentation Shapes Tokyo,” TDB-CAREE Discussion Paper Series

- 都市における土地区画の変遷と価値

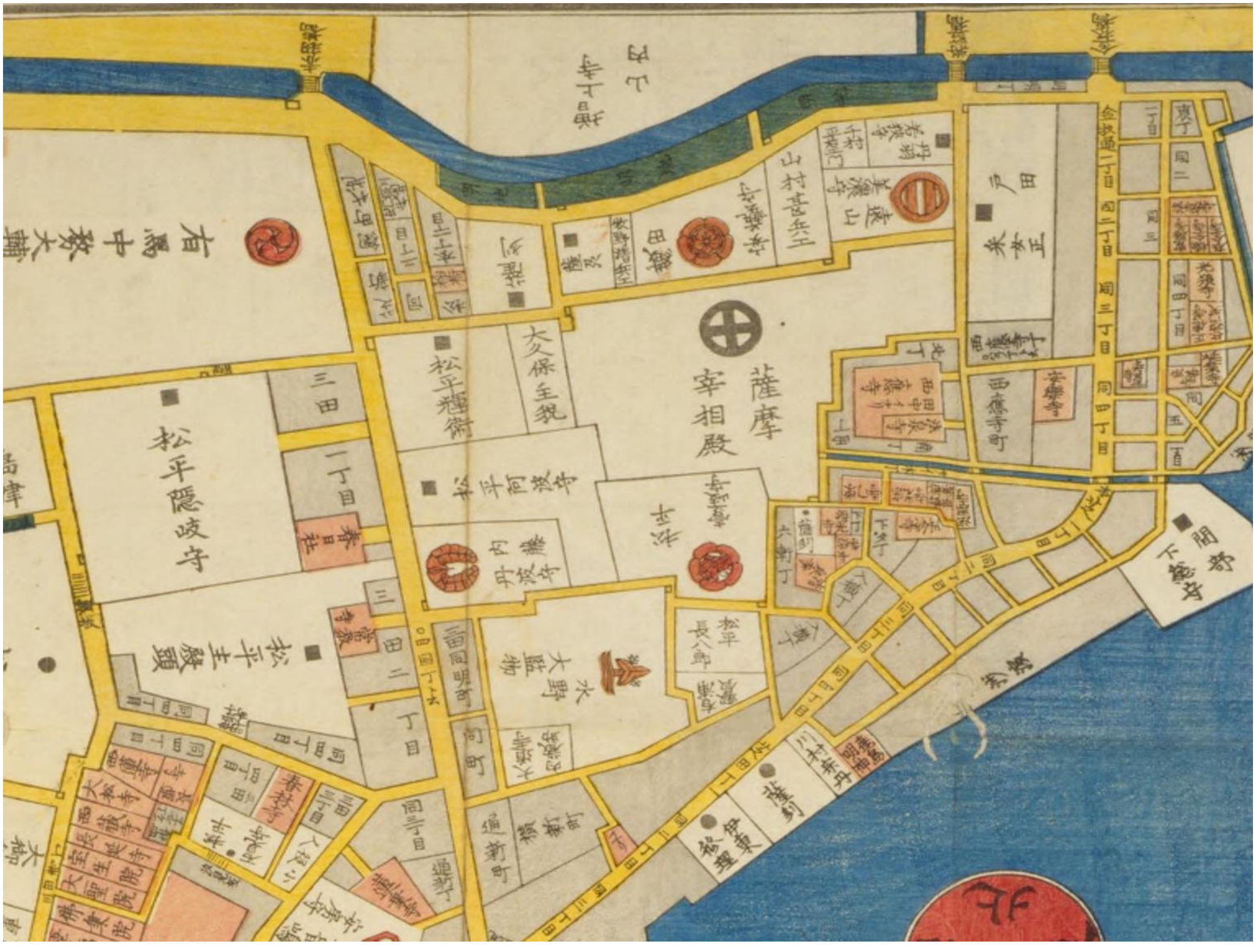
- 利用した主な地図/資料

    - 江戸時代の切絵図->大名屋敷の分布

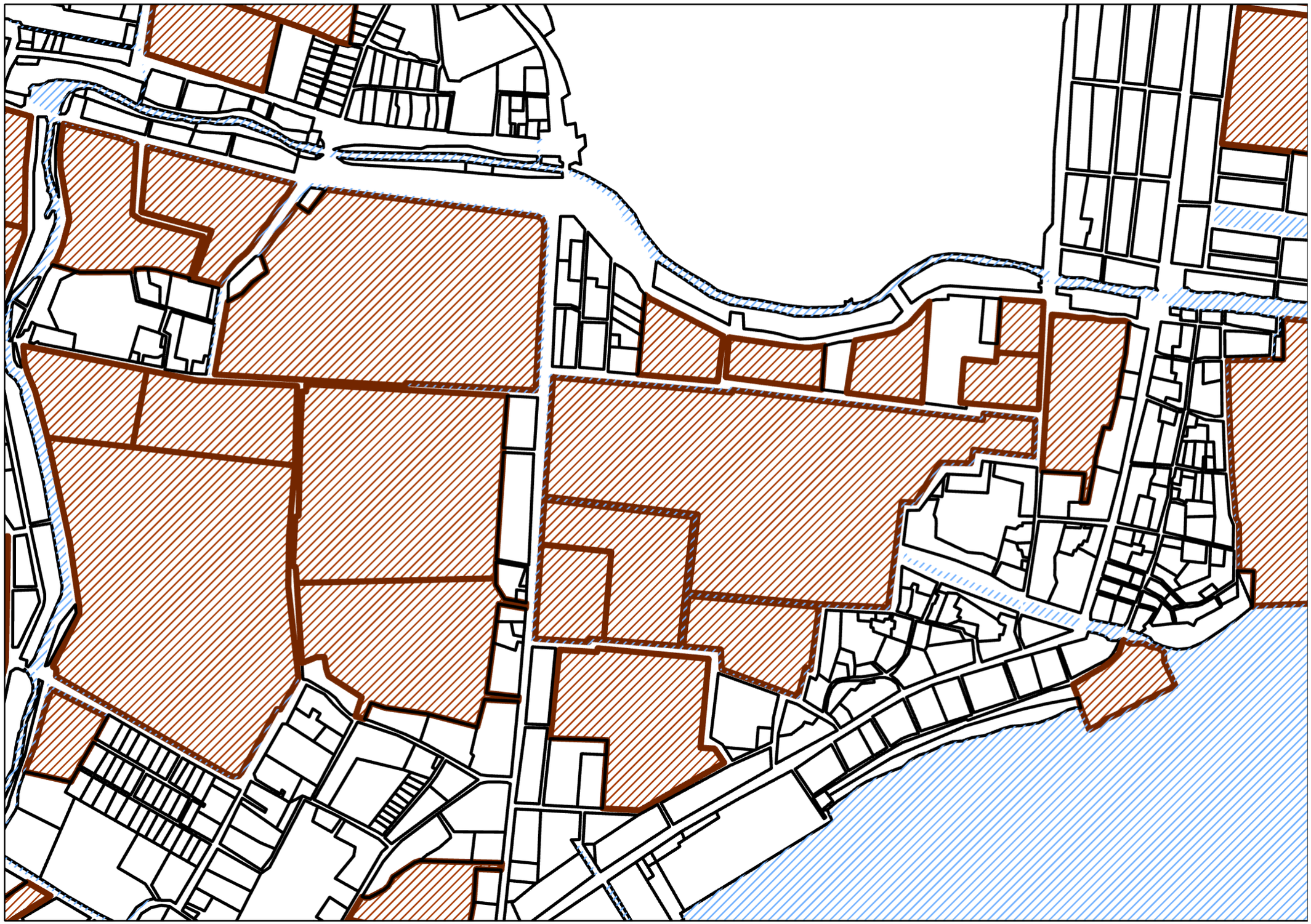
    - 明治以降の地籍図と地価

- これらを位置合わせ（画像の複数ポイントに緯度経度付与）

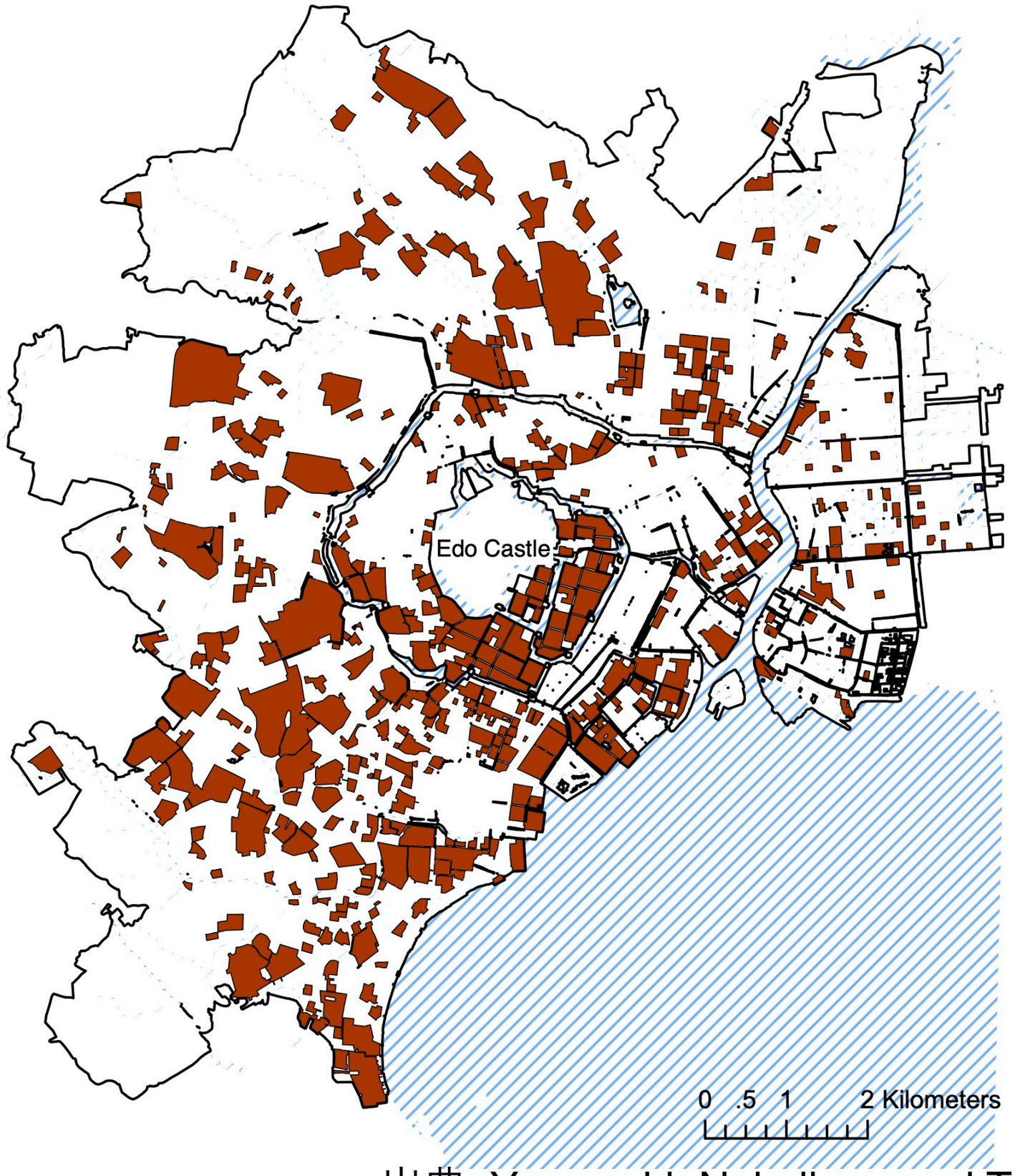
    - →人の手で土地の形をGISソフトウェアで電子化



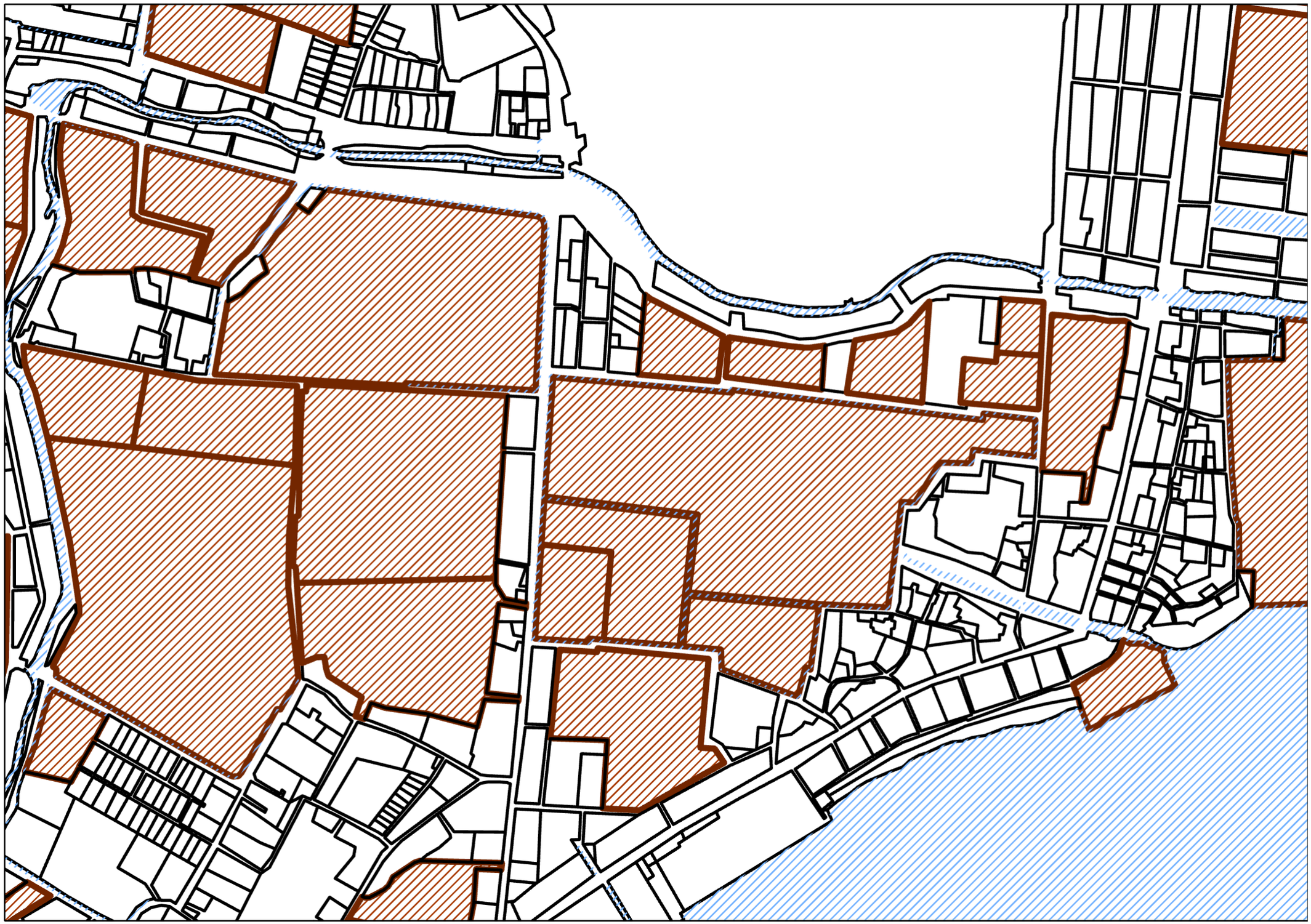
出典: 江戸切絵図　芝高輪絵図



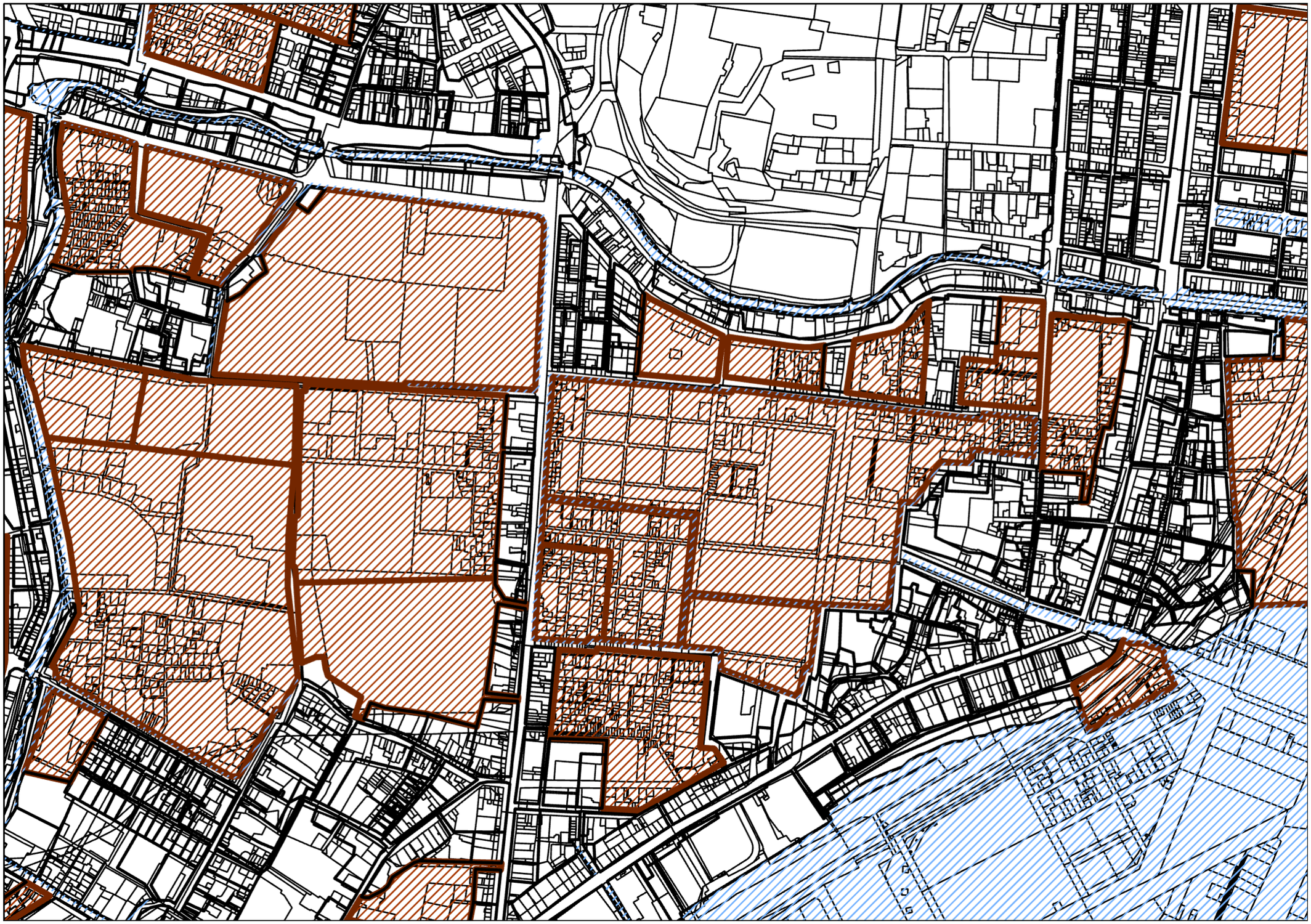
出典: Yamasaki, Nakajima, and Teshima (2021)



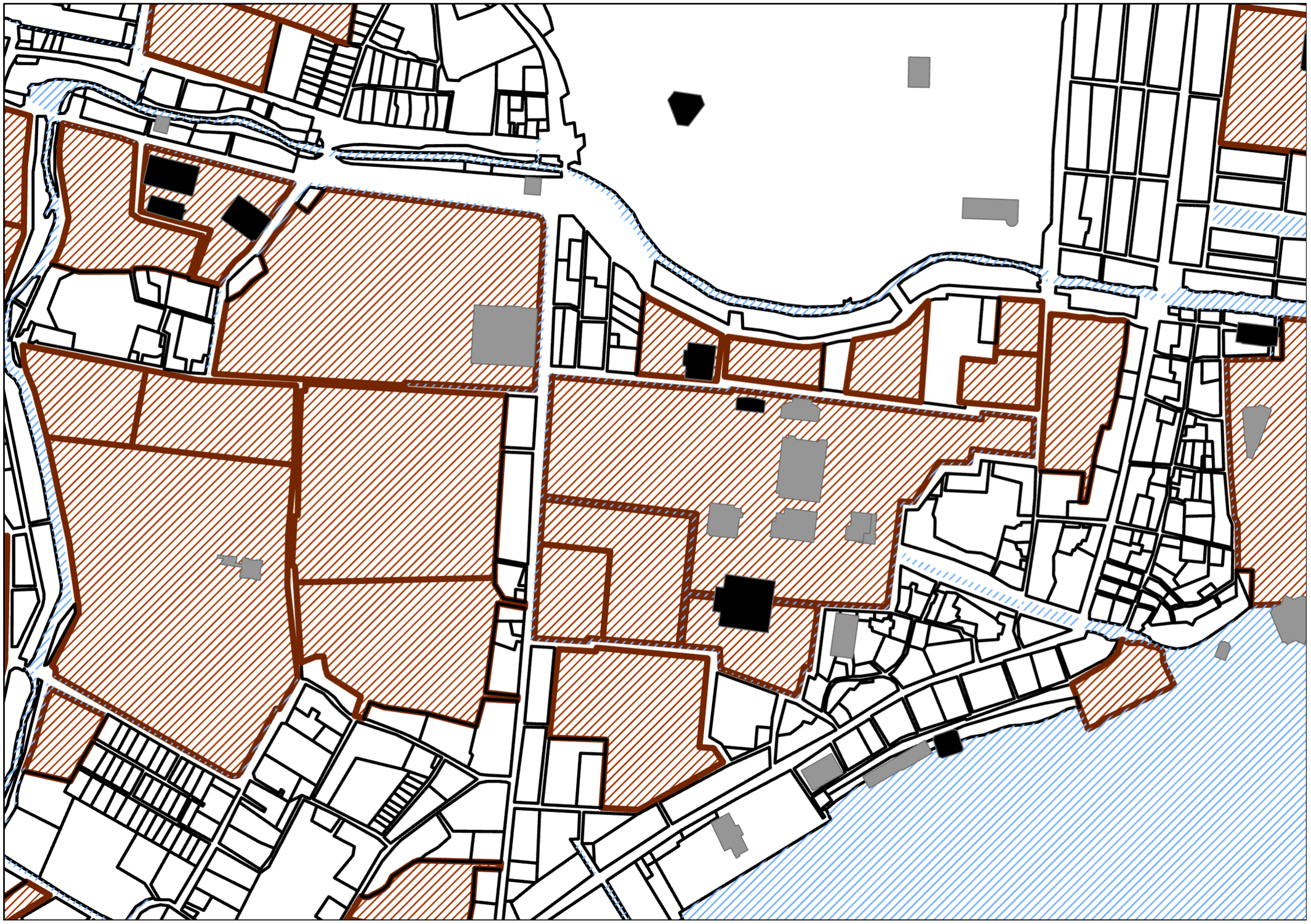
出典: Yamasaki, Nakajima, and Teshima (2021)



出典: Yamasaki, Nakajima, and Teshima (2021)

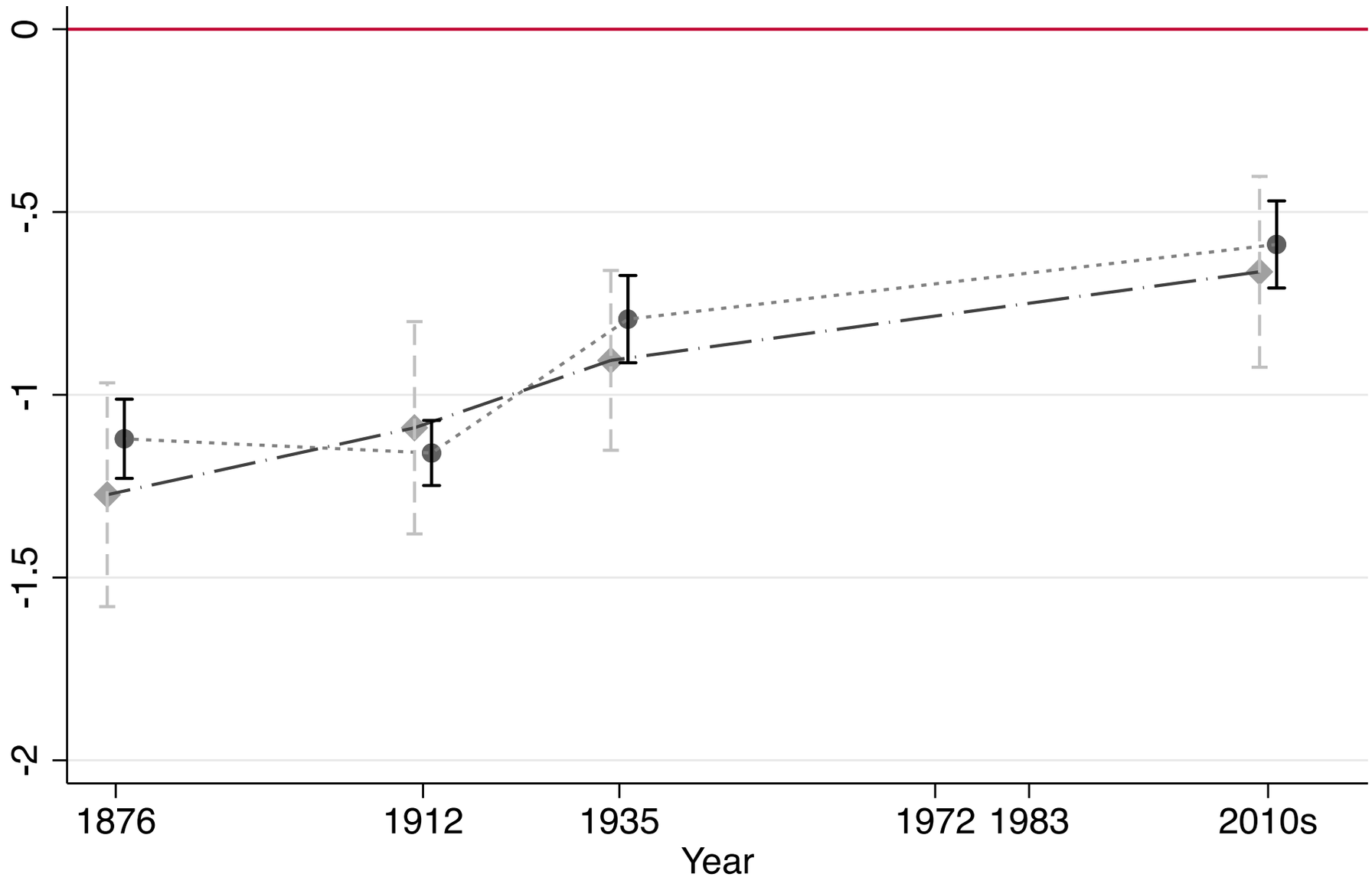


出典: Yamasaki, Nakajima, and Teshima (2021)



出典: Yamasaki, Nakajima, and Teshima (2021)

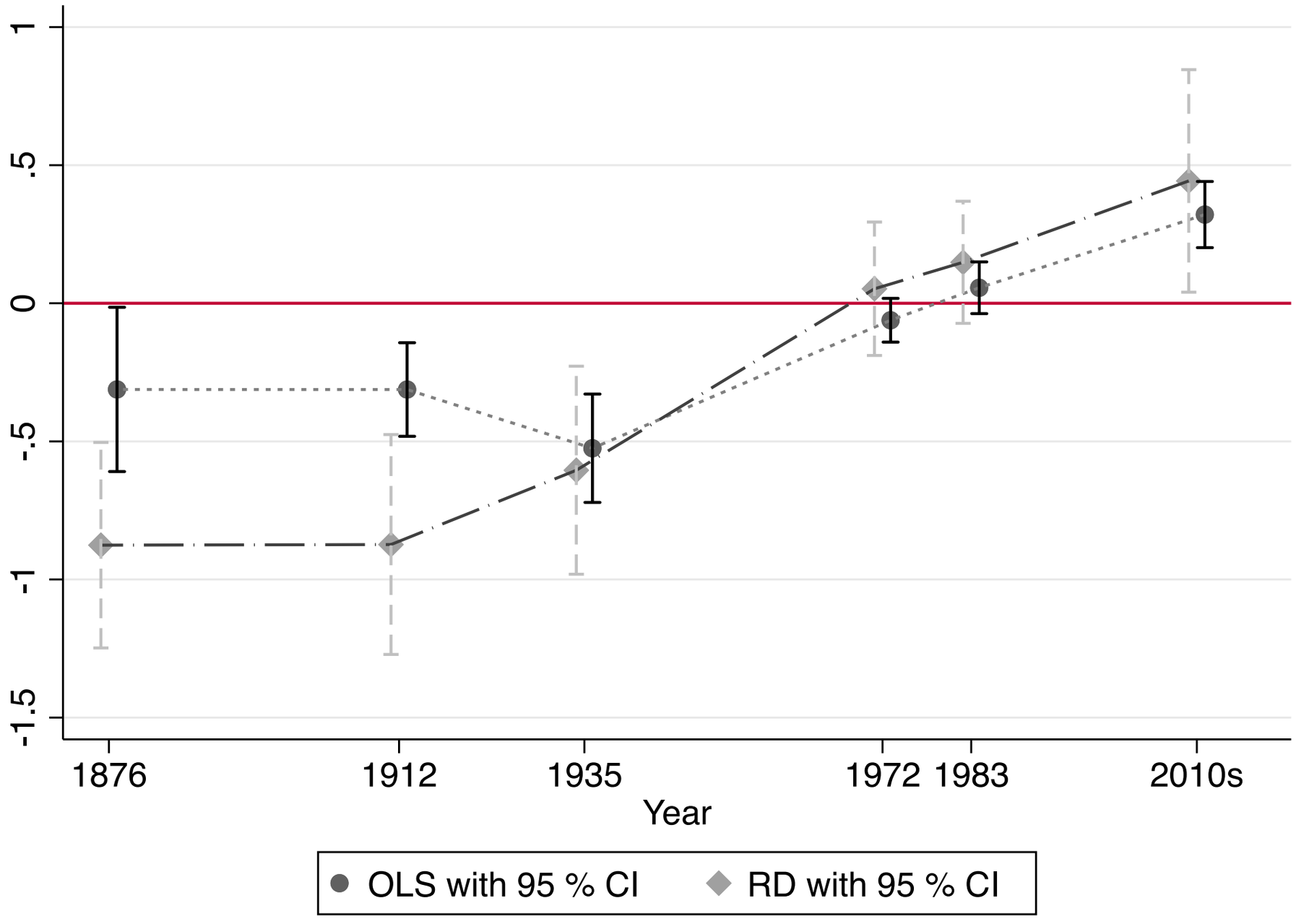
大名屋敷は常に少ない筆数



出典: Yamasaki, Nakajima, and Teshima (2021)

しかし地価が上がりだしたのは

70年代以降



出典: Yamasaki, Nakajima, and Teshima (2021)

地図と経済学

- 地図: 土地利用などを示す貴重な資料
- 都市の発展などは時間がかかる現象なので、過去の地図も有用
- 衛星画像もよく使われる
  - 夜間光
  - 家屋の質

画像の研究例(2)

- Adukia et al. (2021) “What We Teach About Race and Gender: Representation in Images and Text of Children’s Books” NBER WP

- 子供向け図書: 価値観形成に重要な役割

- 子供向けの図書において、どういう人物が描かれてきたのか

  - テキストそのものの解析

  - 画像の解析

- 対象となる書籍: 全米図書館協会の中にある、児童図書館協会が賞に選んだ書籍(1922年から）

- ”主流”と”多様性”の書籍二つに分類

- 多様性: マイノリティーのストーリーに注目

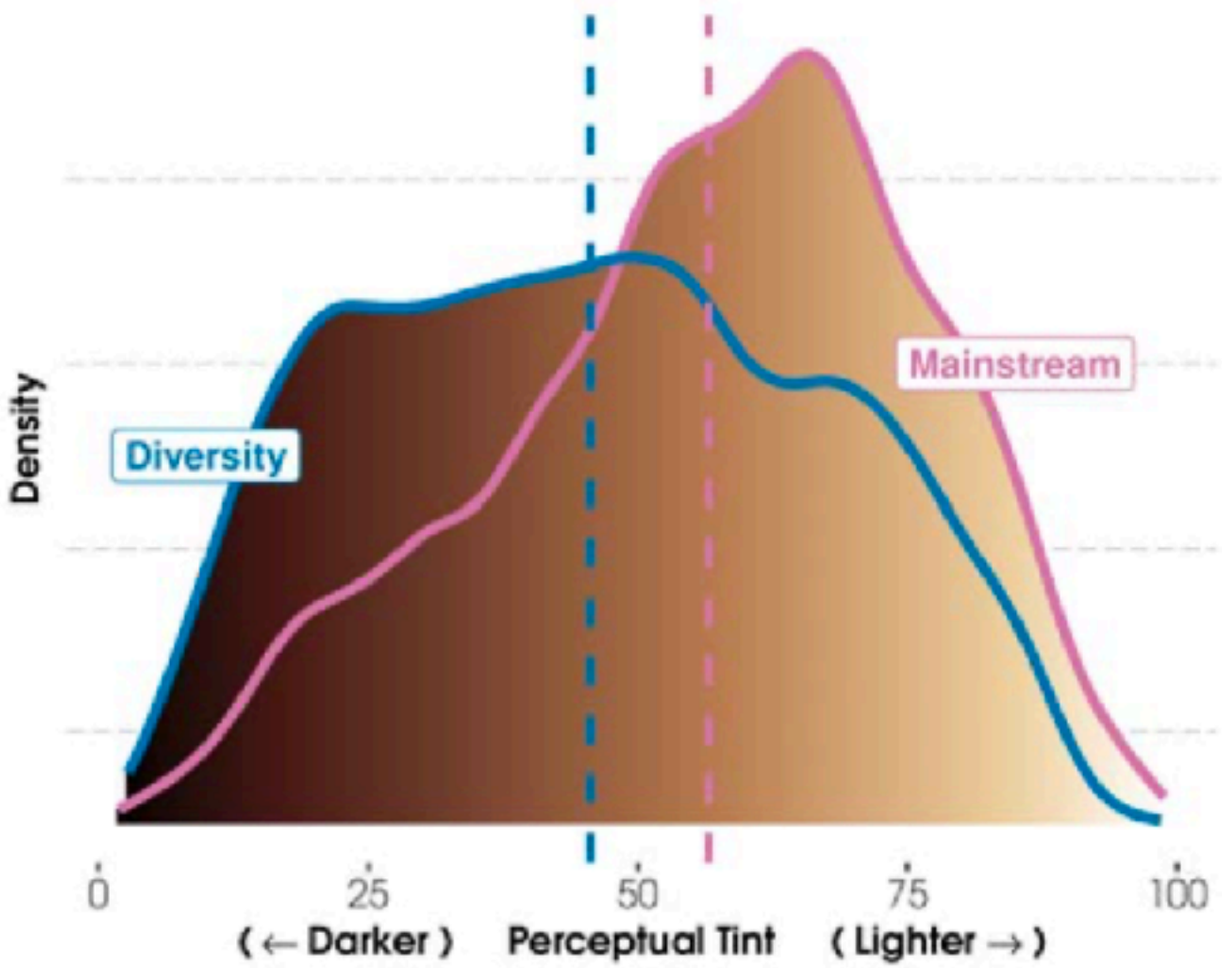
- ちなみに主流の本は常に貸し出し中となるほど人気が高い

- テキスト解析　→登場人物のバックグラウンドを推定
- 挿絵に対して機械学習的処理
- 顔認識->顔部分の色を特定
- 顔認識->性別や人種、年齢を推定

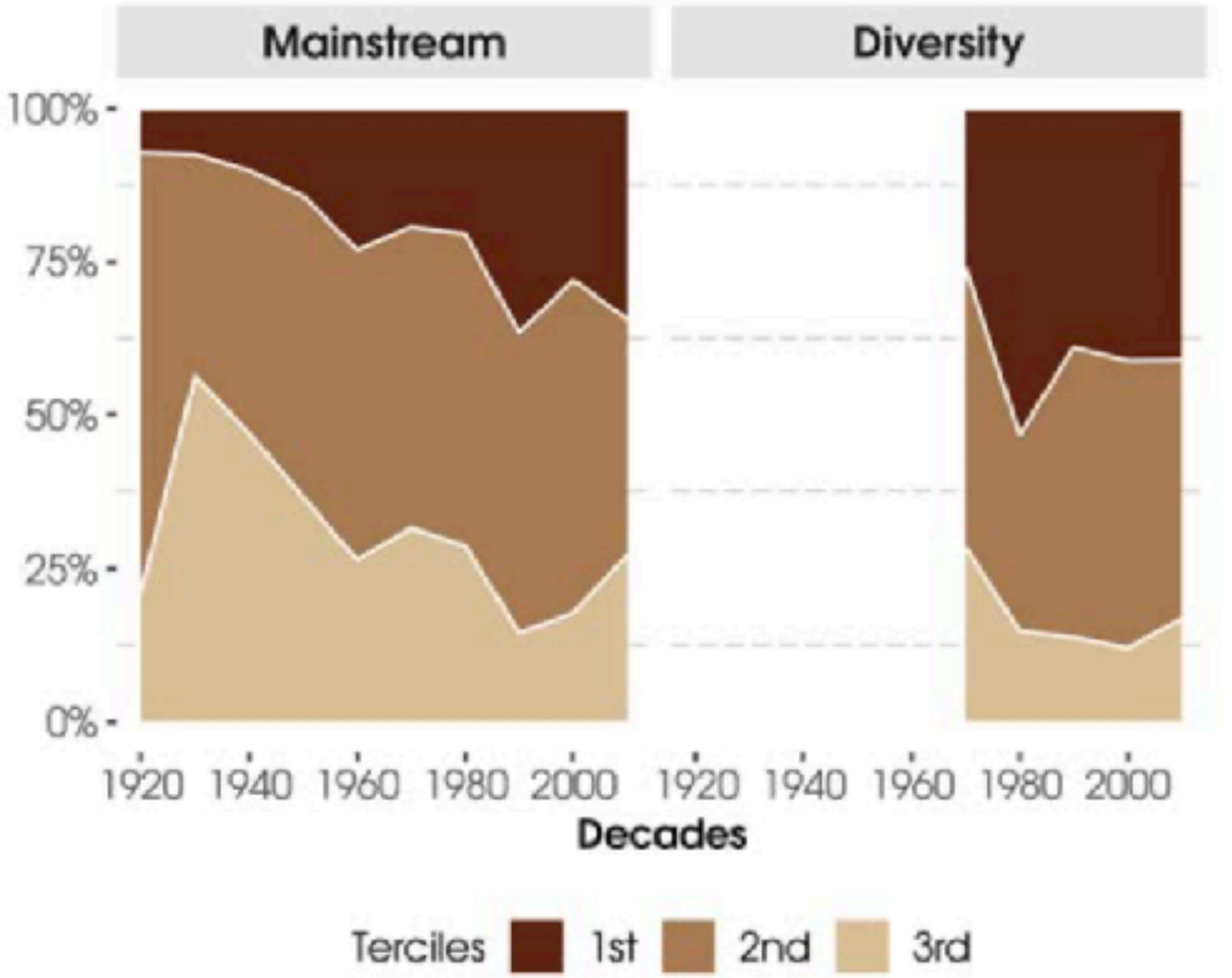
- 主な発見
    - 全体としてはdarker skinの人物が増えているが、主流派の本ではlighter skinの人物がより多く描かれている
    - 女性は画像には増えているが、テキストにはそれほど増えていない
        - 象徴的に使われている可能性？
- センサス人口比率と比べて、黒人、ラテン系は常にunder represented、白人男性は常にover represented

Figure 5. Skin Colors in Faces, by Collection: Human Skin Colors

(a) Distribution of Skin Colors



(b) Mean Proportion in Each Tercile, Over Time



出典: Adukia et al. (2021) “What We Teach About Race and Gender: Representation in Images and Text of Children’s Books” NBER WP Number 29123

まとめ

- 経済学or社会科学にとって有用なデータは統計やOCRによる単純なテキスト情報だけとは限らない
- 電子化による保存/複製のコスト低下 + 機械学習の発展により膨大な資料からのデータセット作成が容易に
- 2021ノーベル経済学賞との関連
  - 自然実験アプローチ: 社会に偶然生まれた変化を実験として見立て、物事の因果関係を探る（実験室実験が難しい社会科学にとって重要なアプローチ）
  - **しかし偶然生まれた変化を描写する資料がないと、分析ができない**
  - うまくいった例: Dell and Querubin (2018) “Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies”
    - アメリカ軍が爆撃対象選定に使ったスコアの計算ミス（丸め誤差）を用いて、たまたま爆撃目標になった村とそうでない村を比較
    - 米軍の資料から正しいスコアを計算する必要あり -> 資料収集

個人的に苦労すること

- 個別の資料/情報に関して:
  - OCRなどを使ったテキスト化はほとんどされていない。構造化はされていない。
  - メタデータ: 例えば地図であれば位置情報
- 資料の全体像に関して:
  - リサーチナビ: ありかがわかるので非常に有用だが、内容まではわからない
  - 内容が分かる良い例: 国会会議録検索システム
  - 例えば地方に散逸している選挙資料などは中央で保存管理すべき
  - 中央集権の欠如？