Measuring Housing Demand Using Multimodal Geospatial Networks

Aaron Bramson

Workshop on Statistical Network Analysis and Beyond, June 3nd, 2025

GA TECHNOLOGIES

Aaron Bramson, PhD

Principal Investigator
Advanced Innovation Strategy Center
GA technologies, Inc. Tokyo, Japan

Research Interests:
Complex Systems, Agent-Based Modeling, Network Theory

Actual Work:
Geospatial Data Analysis, AI Model Design, Algorithm Design

Background

- **Residential property demand (and price) estimation is extremely important, but it is complicated by heterogeneous preferences, unique goods, low turn-over, intangible features, geospatially distributed alternatives, and many other factors.**
- **Demand for a location is usually estimated using physics-based gravity modeling approaches, neo-classical economic market models, or hedonic pricing models.**
- **Hedonic pricing models are based on the theory that the value of a property is the sum of the values of the property's features; neighborhood features and transportation convenience should be core valuable features.**

Claims

- **Demand (and therefore prices) should be based on four levels of features:**
    1. **the unit**
    2. **the building**
    3. **the surrounding area**
    4. **accessibility to other areas of interest**

- **Accessibility-based demand should utilize multiple modes of transportation and multiple possible objectives/interests.**

Objective

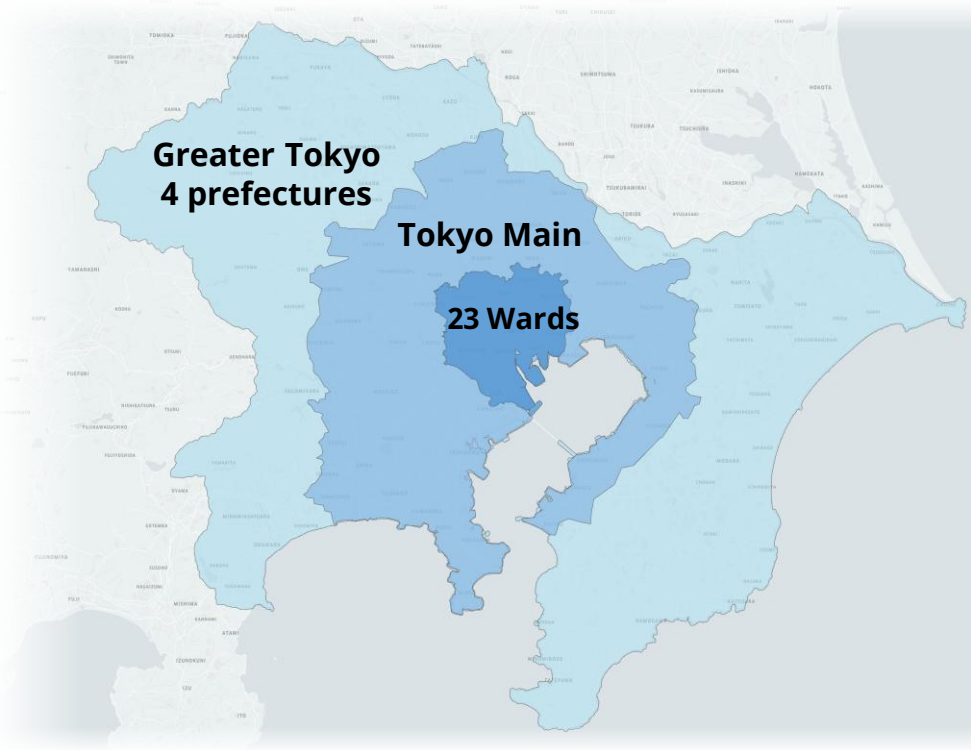
- **In this first pass, we aim to estimate demand for locations…**
    1. **for the main Tokyo Metro area using**
    2. **travel by walking and train**
    3. **based on employment locations.**

- **We estimate location demand using network diffusion to propagate potential access to jobs across an integrated network to each location in the region.**
- **Leverage the predictive power of machine learning methods, while still addressing the demands of reason and responsibility by using “explicable AI” methods.**

Creating the Demand Score

Tokyo Main Area

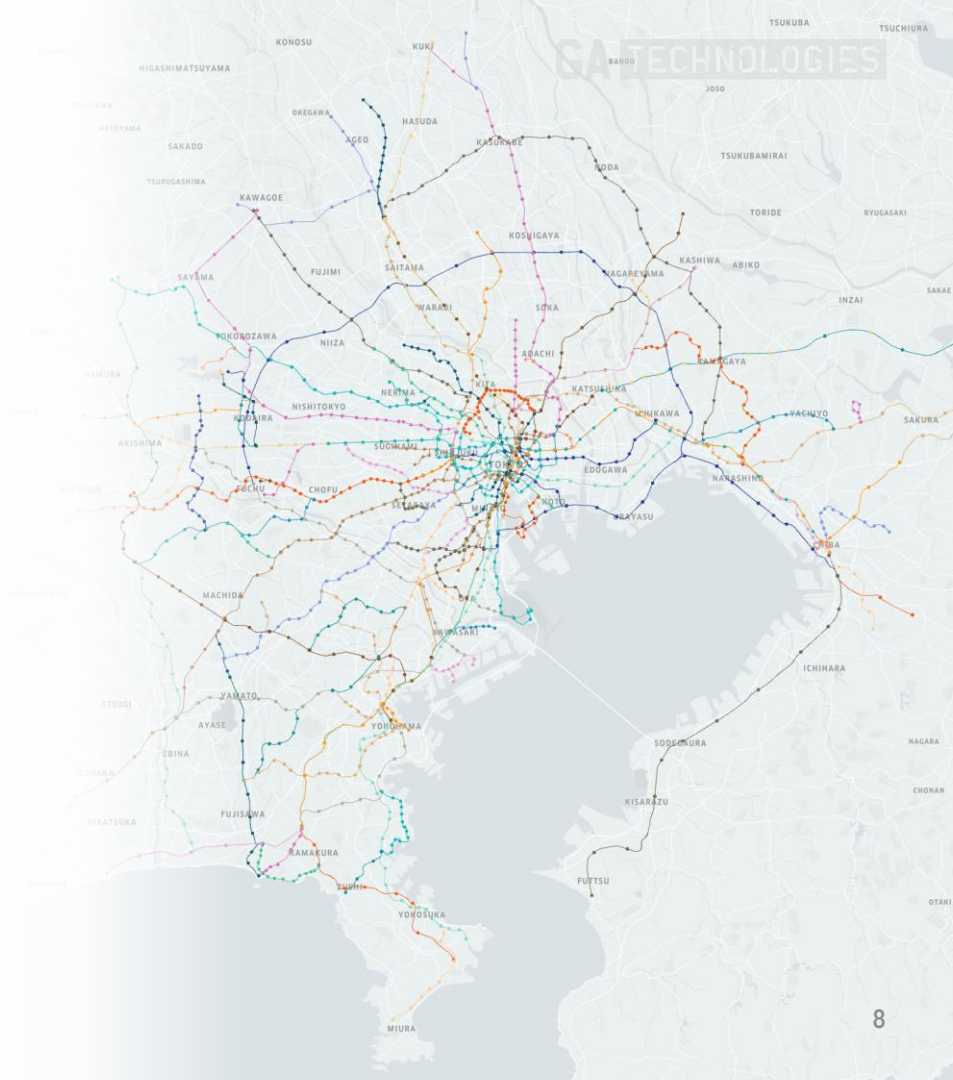
- **Our region of analysis is “Tokyo Main”; a custom region roughly following the E16 circuit route.**
- **This region covers 36% of the land area of the four prefectures but includes 92% of the population.**
- **Surface area is 4893 km².**
- **The population is 32,197,448 people! (25% of Japan’s population).**
- **There are 15,080,305 salaried employees in the region.**



By comparison, both the population and surface area are roughly four times the combined five boroughs of New York City.

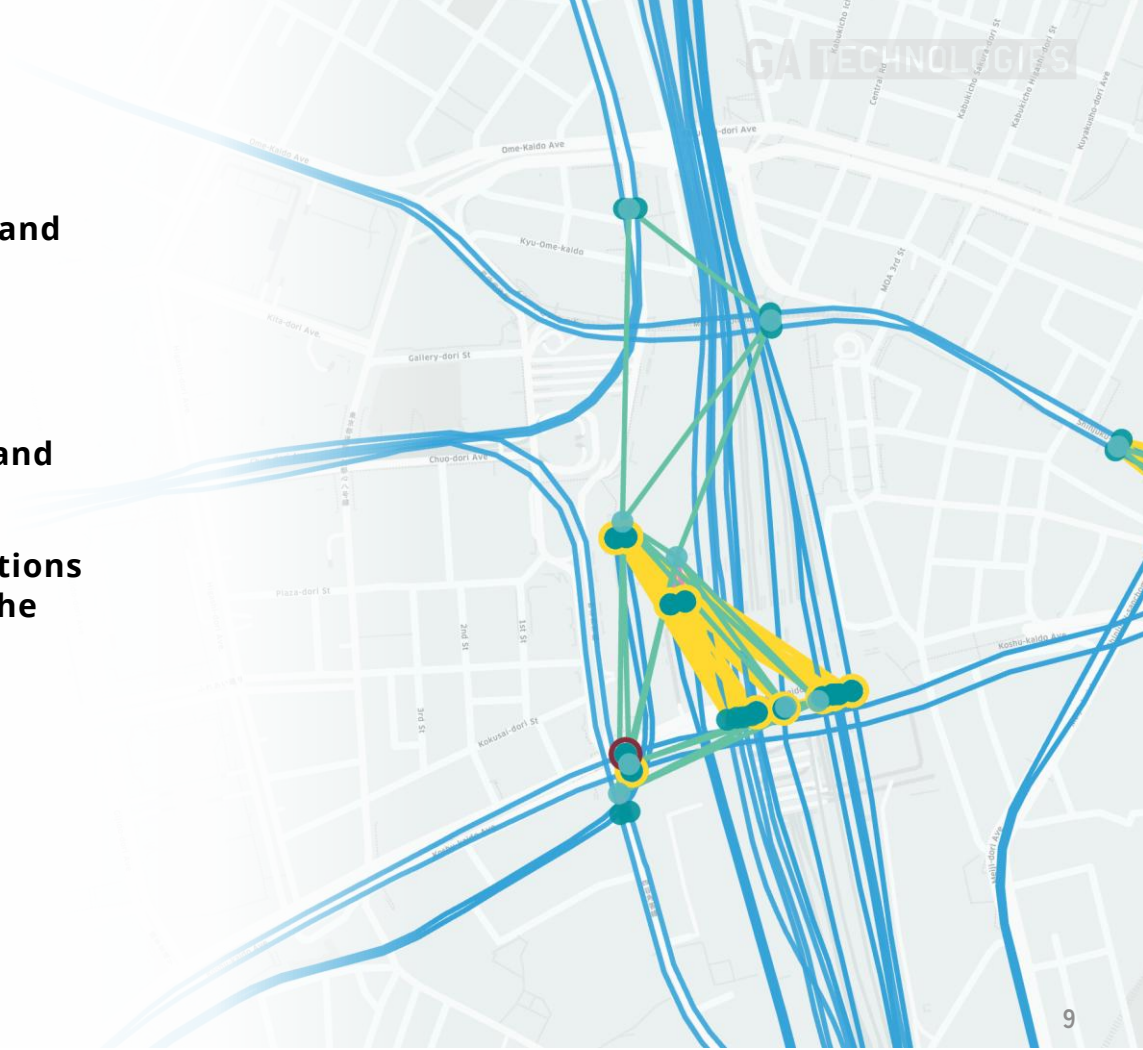
Train Network

- **Build from open data (OSM and Wikipedia) with traversal times bought from Ekitan.**
- **The region contains 1440 train, subway, and tram stations (1247 by name).**
- **Network includes through service, time needed to transfer, express trains, etc.**



Train Network

- **Network is directed; traversal times and cost differ by direction.**
- **Train edges have platform nodes for stopping position.**
- **Platforms are connected to stations and other platforms.**
- **Stations are connected to nearby stations and station entrance/exit nodes on the road network.**

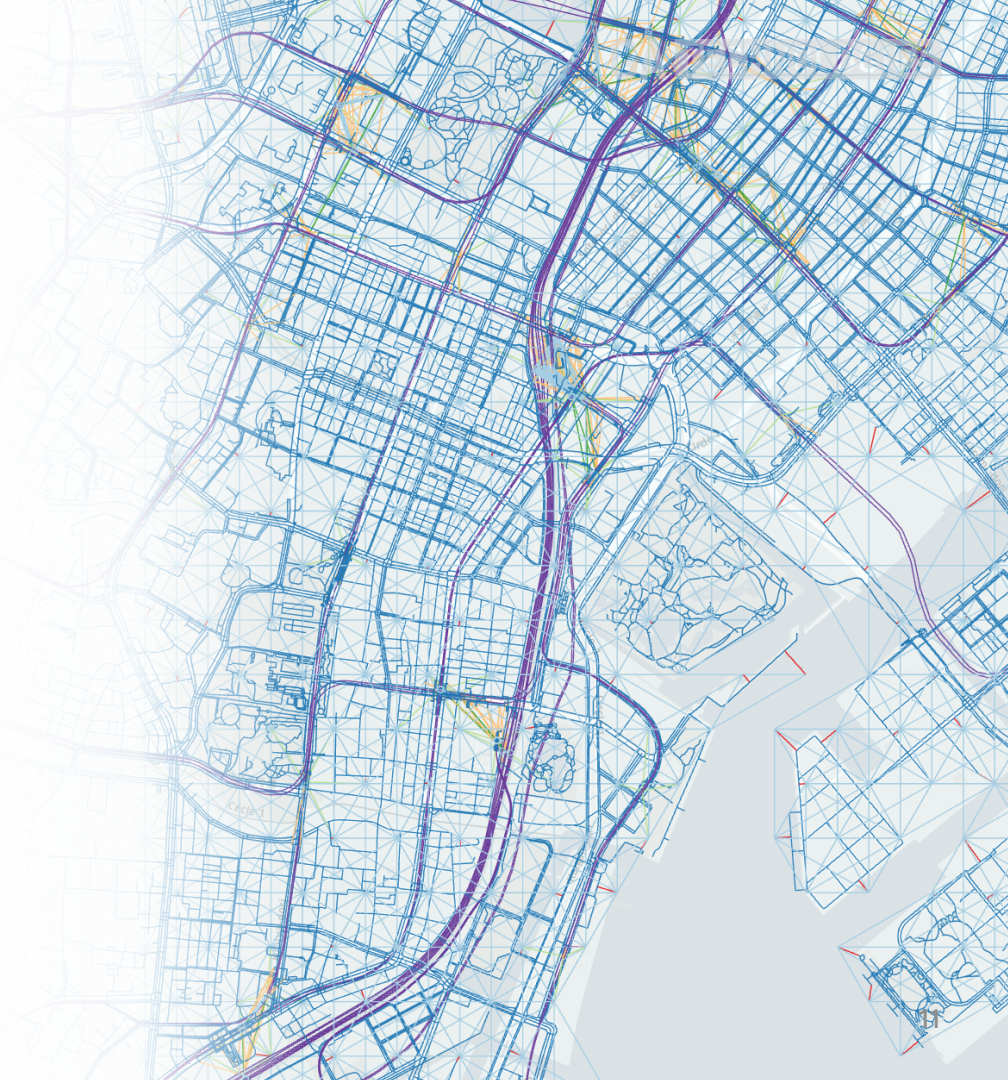


Walking (Road) Network

- **All walkable edges from OSM.**
- **Manually improved connectivity of stations, ensuring each one has proper exits and those exits are integrated into the larger network.**
- **It’s awesome, but the road network too big to use directly.**

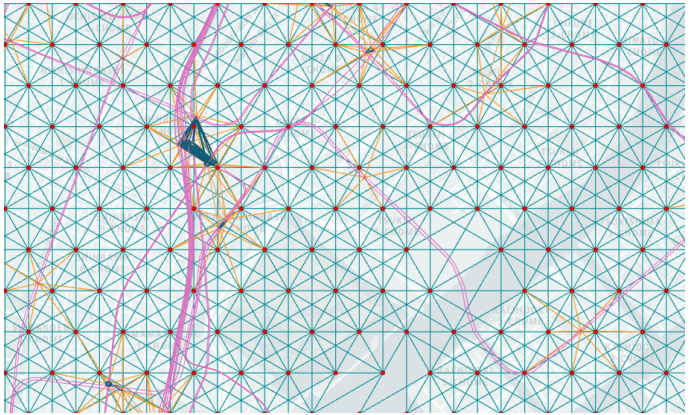
Hex Network Construction

- Create a grid of 250m inner diameter hexagons, resulting in 144,849 hexes.
- Connect each hex centroid to the closest point on the road network.
- Connect hexes to neighboring hexes within 550m (radius-2 neighbors)
- Connect hexes to station entrances/exits within 550m.
- Inter-hex traversal times are generated from road traversals.
- Keep links with traversal times less than 15 minutes by foot.
- Remove the road network.



Hex+Train Network

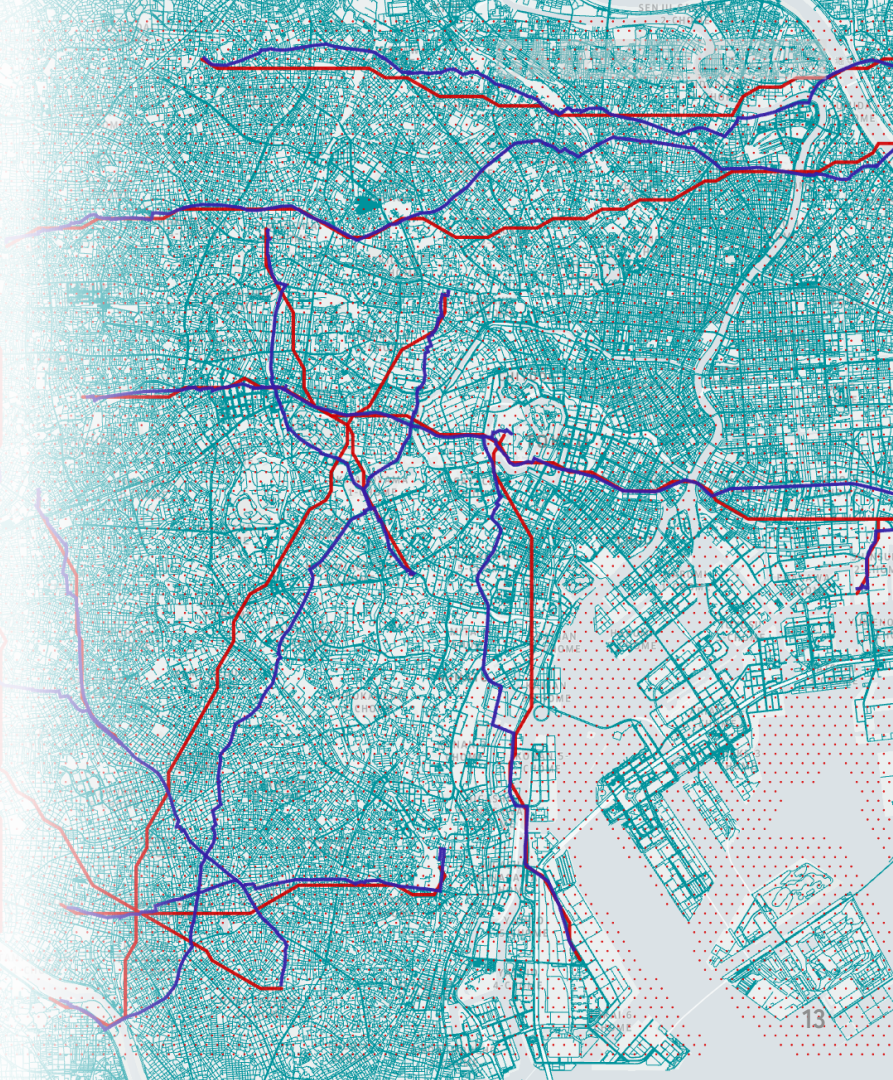
- **Both the hex and the train network include station entrance/exit nodes, so they are easy to integrate.**



Hex Network Accuracy

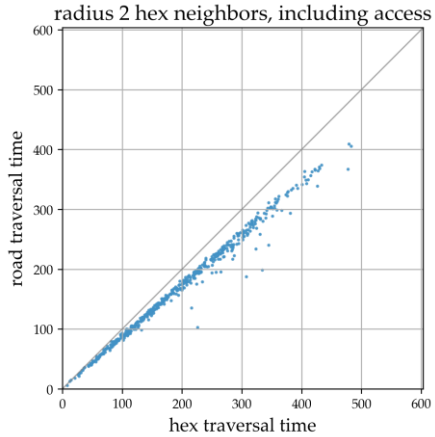
- We calculated the accuracy of 10,000 trips using
  - radius-1 vs radius-2 connectivity
  - including vs excluding the hex link time
  - walking only vs walking+train

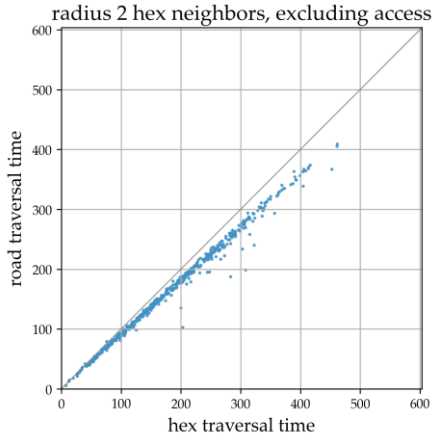
- When including the train network, the average trip time was 59.8 minutes, closely matching the average commute time.
- Only actually accessible routes exist.



Hex Network Accuracy

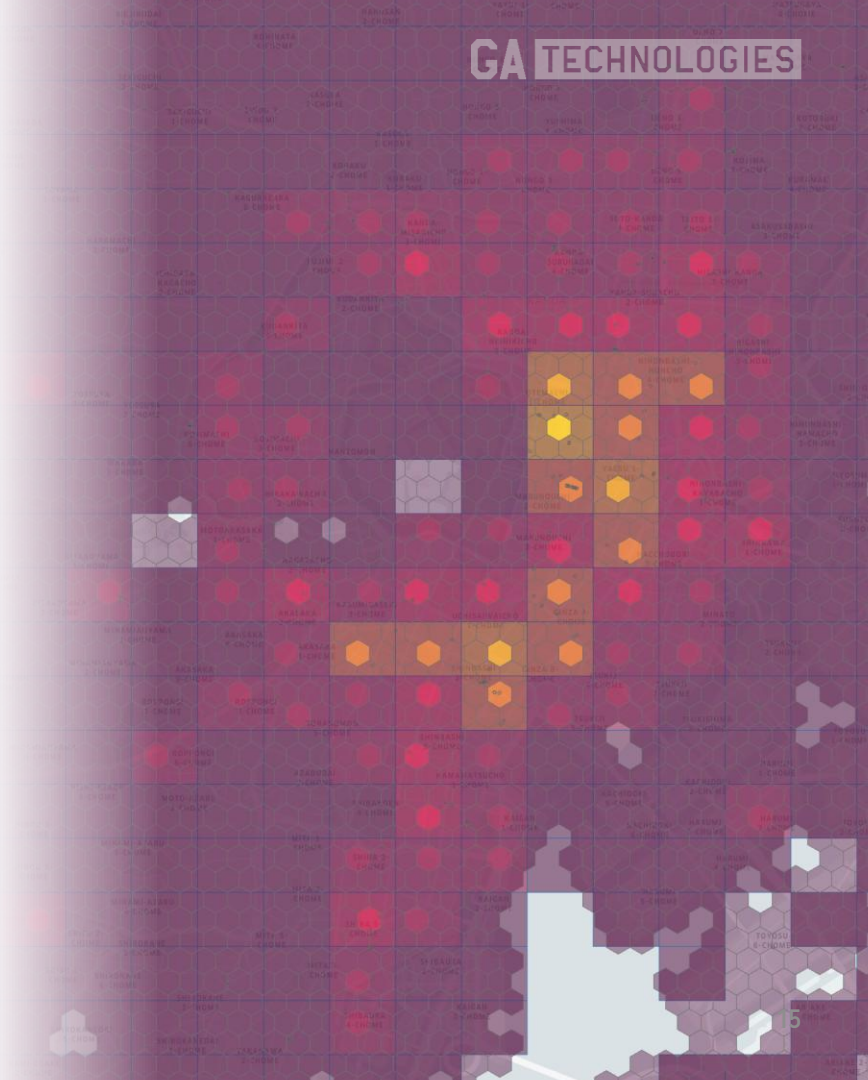
- **When including the train network and the hex link time, the mean absolute error between the road+train times and radius-2 hex+train times was 2.5 minutes of a 60-minute trip (4% error).**
- **There is a consistent bias in the estimates, so further refinement should be able to achieve even closer agreement.**





Employment Data

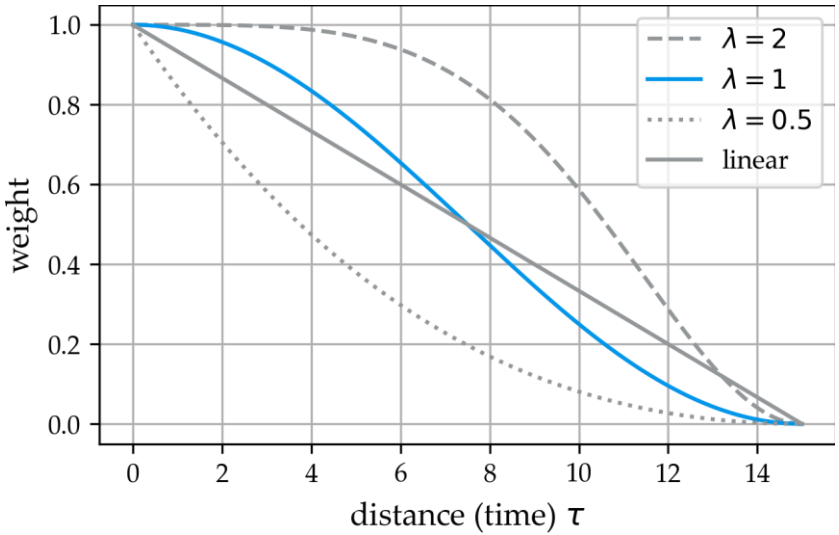
- The employment data comes as a 500m square grid.
- I resample the data to the hex grid by assigning a grid's value to the hex that overlaps the grid's centroid, or closest hex if none does.
- The result is 17,433 hex nodes with source demand.



Propagation Function

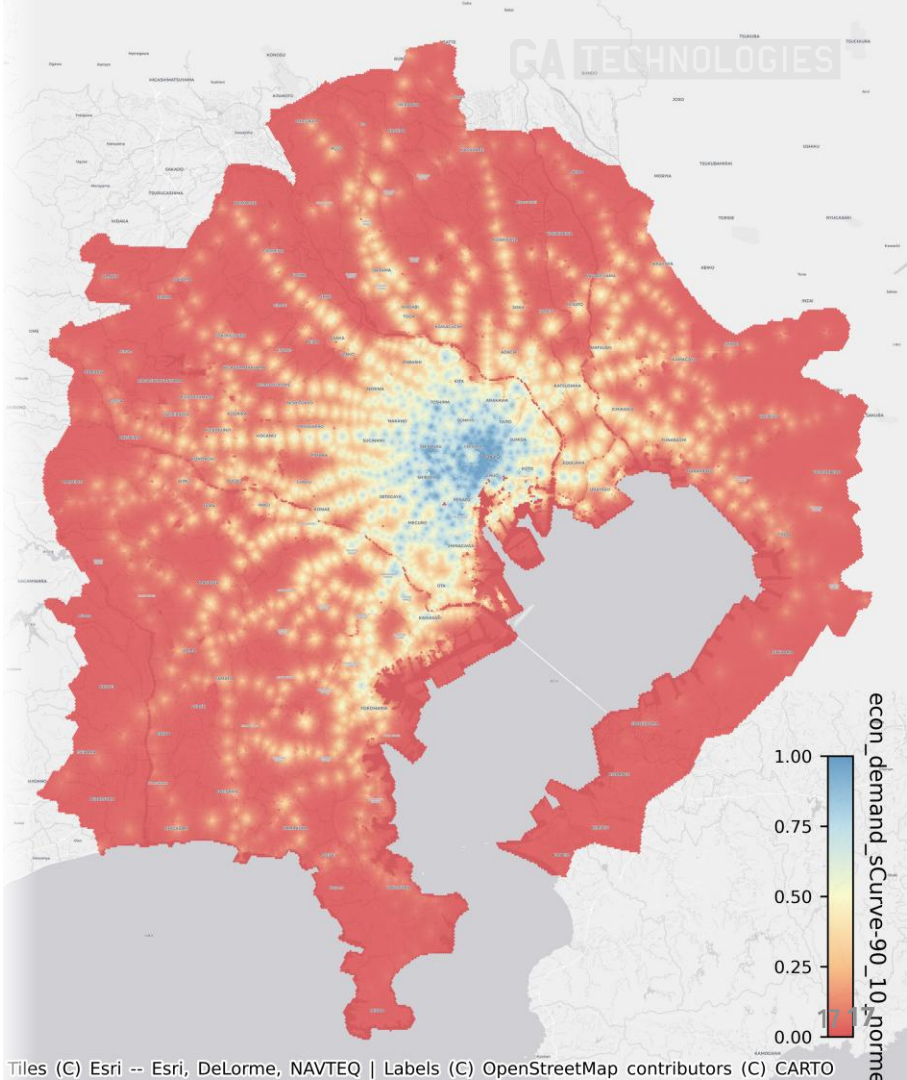
- **We use four different discount functions: a linear function and a cosine based curve parameterized with three values (λ=0.5, 1.0, 2.0).**
- **All four are run with two different horizons: 60 minutes and 90 minutes.**

$$W_{ij} = \frac{1}{2}\left(1 + \cos\frac{\pi\tau_{ij}^{\lambda}}{T^{\lambda}}\right) \text{ if } \tau_{ij} < T \text{ else } 0$$



Employment Diffusion Results

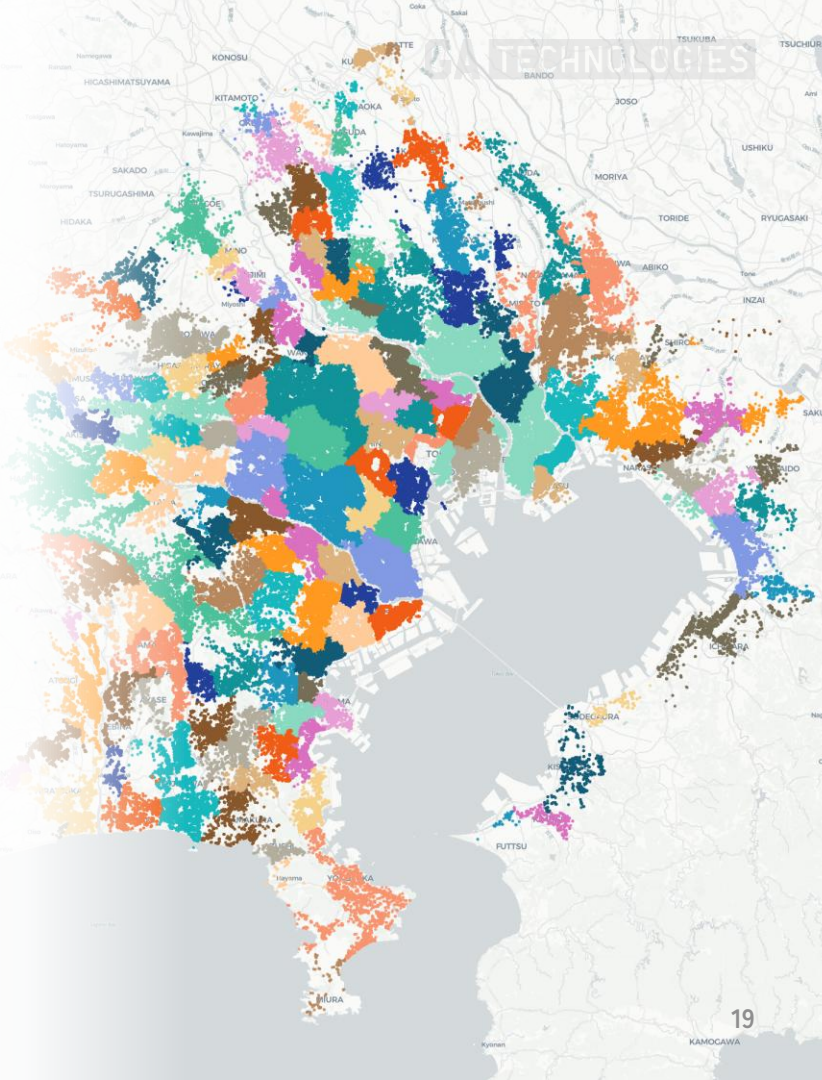
- We calculate **estimated employee demand** as the time-weighted potential flow of salaried employees from places of work, across the transportation network, to each location.
- Results are qualitatively similar for all four curves, and increasing the horizon from 60 to 90 minutes has the obvious effect.
- Although there are jobs all over, the high concentration in the city center and secondary centers dominates the flow.



Simple Validation of the Demand Score

Property Data

- Rental properties in concrete structures from 2022-01 to 2024-01.
- There are 806,863 total records, 793,375 entries with complete data.
- We filter available rental prices to reduce heterogeneity.
    - Prices published since January 2022.
    - Built within the years 2010 through 2014.
    - Footprints of 25-28 m^2 (one room apartments)
- We find 356,744 units of similar size and age.
- There are 8,412 hexes with prices (~6.65%)



Validation with Rent Prices

Pearson Correlation with Prices

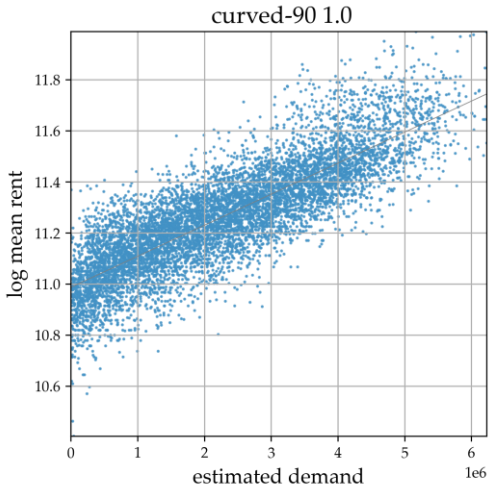
Linear 60	0.825
sCurve 60, 0.5	0.788
sCurve 60, 1.0	0.804
sCurve 60, 2.0	0.82
Linear 90	0.81
sCurve 90, 0.5	0.832
sCurve 90, 1.0	0.826
sCurve 90, 2.0	0.8

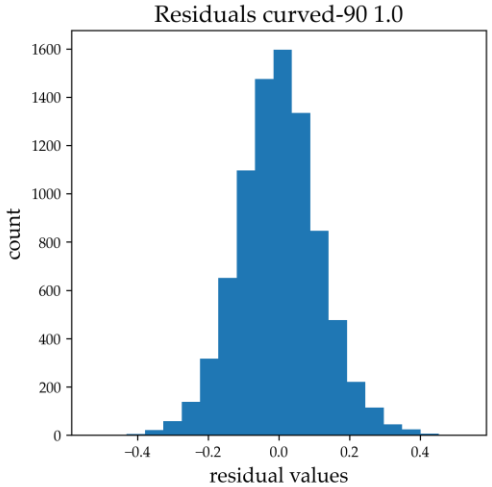
Pearson Correlation with log Prices

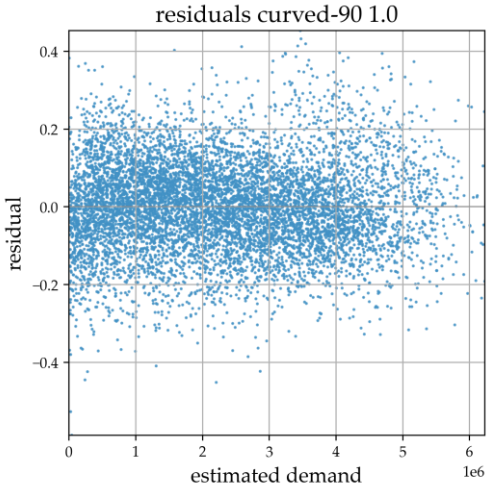
Linear 60	0.81
sCurve 60, 0.5	0.758
sCurve 60, 1.0	0.779
sCurve 60, 2.0	0.802
Linear 90	0.825
sCurve 90, 0.5	0.828
sCurve 90, 1.0	0.832
sCurve 90, 2.0	0.819

Validation with Rent Prices

- **Not only is the fit high at 83.2%, but the residuals are nicely distributed.**







Full Pricing Model Using the Demand Score

with Masayoshi Mita

Pricing Model Using Four Levels of Data

- **Demand (and therefore prices) should be based on four levels of features:**
    1. **the unit**
    2. **the building**
    3. **the surrounding area**
    4. **accessibility to other areas of interest**

- **We use machine learning models to determine the usefulness of the accessibility scores.**
- **We compare opaque models using lon, lat, and station IDs to “explicable models” using the geospatial data.**

Rich Geospatial Data

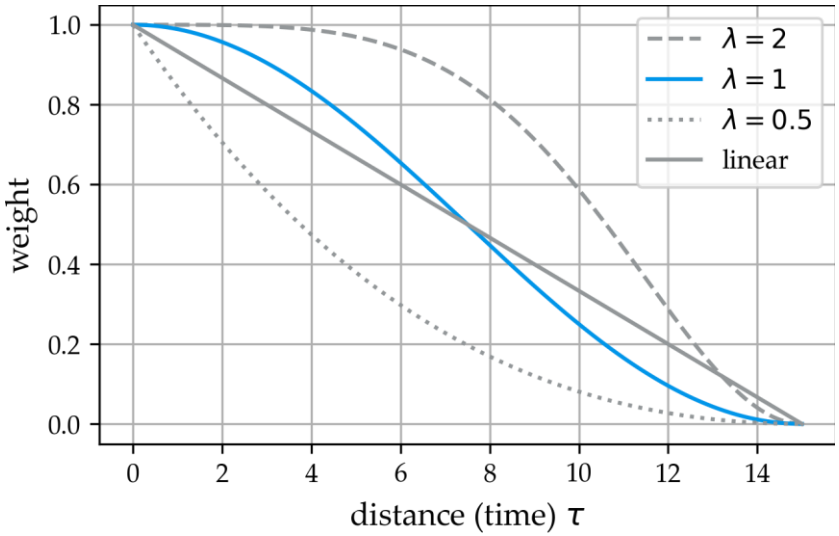
- **In addition to the property and network data, we have amassed a large amount and variety of data for the greater Tokyo area including:**
    - **Zoning**
    - **Land use**
    - **Vegetation**
    - **Natural Hazards**
    - **Population**
    - **Jobs and Companies**
    - **Stores and Amenities**
    - **Embassies**
    - **Undesirable Establishments (pachinko, yakuza, etc.)**
    - **Building Structure Sizes and Shapes**

Aggregate Neighborhood Data

- **For some features, we aggregate within 15 minutes (1200m at 80 meters per minute) using the hex network (about 4 hexes).**
- **For others (not based on access) we aggregate over hexes within a 1200m radius.**
- **We use other values in some special cases; e.g. embassy score uses 30 min, train accessibility scores use 60 min, estimated demand uses 90 min.**
- **In all cases, we use the same equation to weight the contribution of reachable nodes to the focal node.**

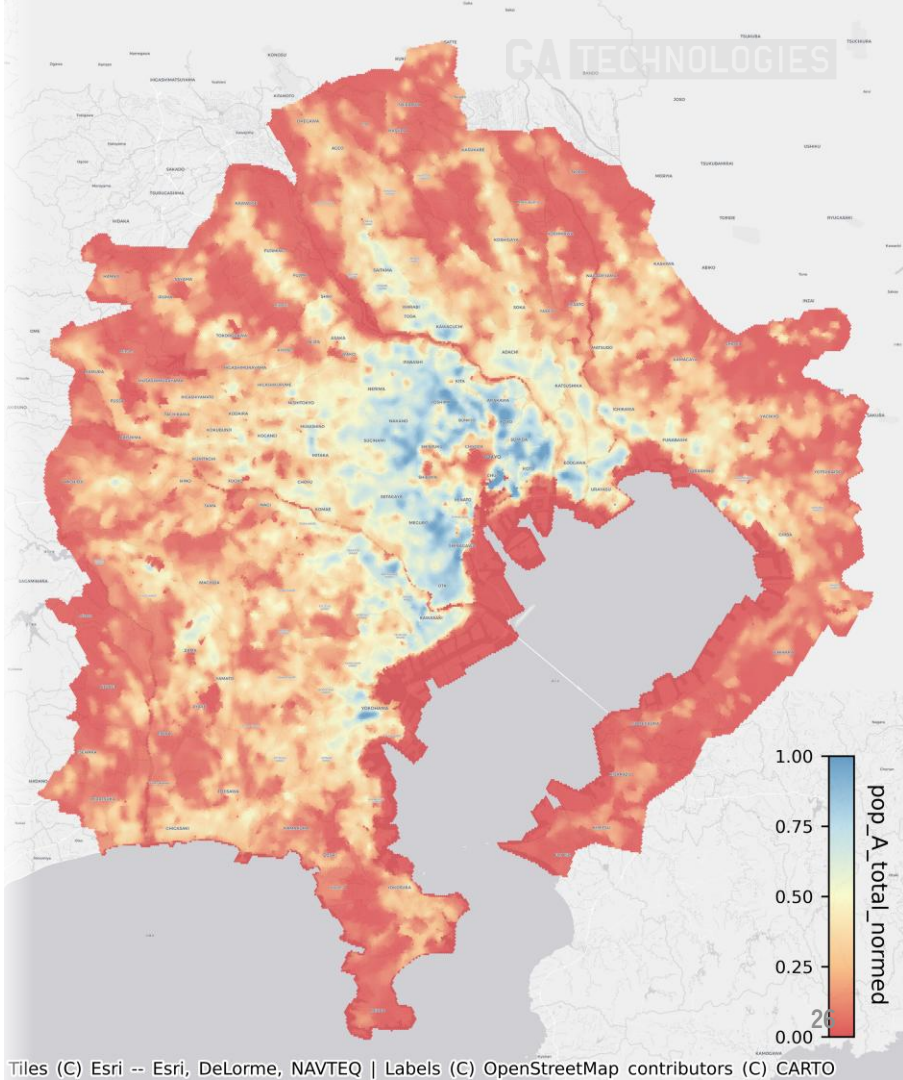
  X_j is the value at hex *j*
  T is the threshold (e.g. 15 min)
  τ is the distance from *i* to *j*
  λ is a tuning parameter =1.0 here

$$W_{ij} = \frac{1}{2}\left(1 + \cos\frac{\pi\tau_{ij}^{\lambda}}{T^{\lambda}}\right) \text{ if } \tau_{ij} < T \text{ else } 0$$



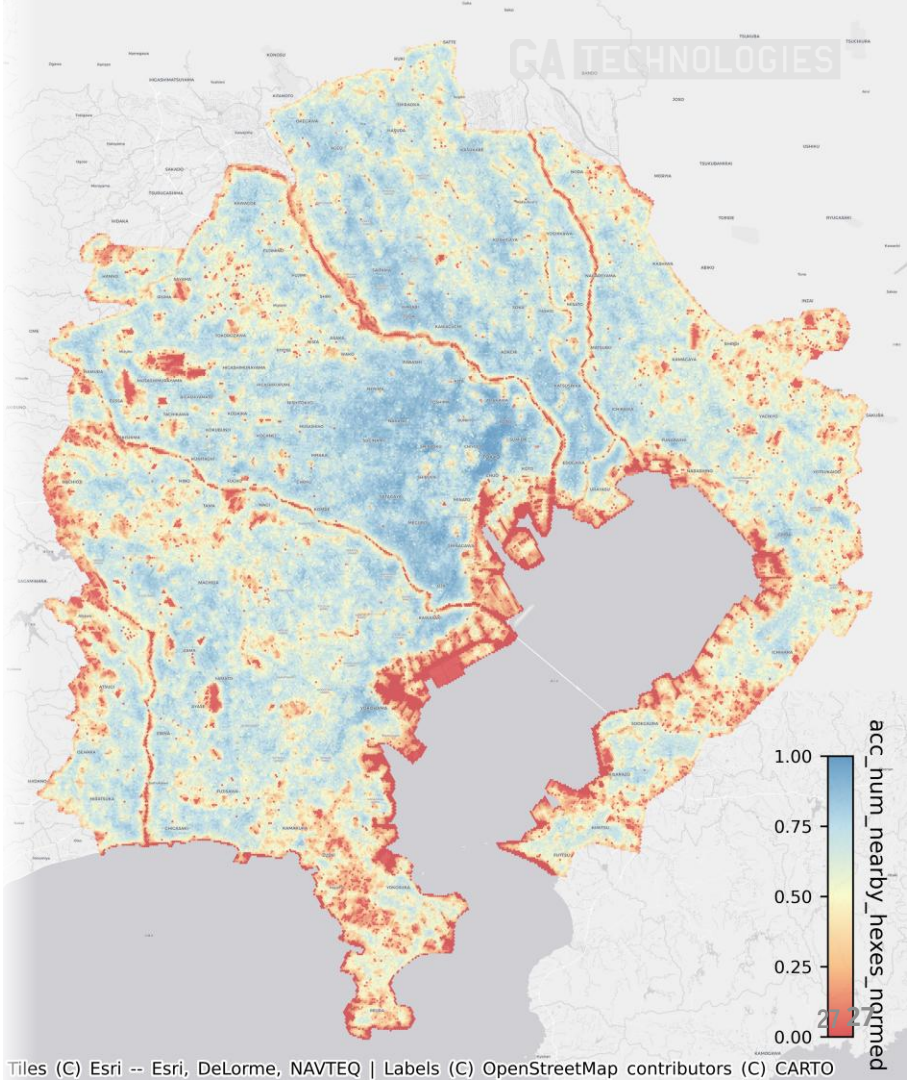
Normalized Scores

- The result is a smoothing (aliasing) of the features across hexes in a way that
    1. Respects movement constraints and
    2. Captures the strength of influence.
- We normalize the data by first identifying the lower and upper 0.1 percentile mark
- We use those limits to map the values to the [0, 1] range
- We clip lower values to 0 and larger values to 1.



Network Based Features

- In addition to using the network to generate the time/distance weighting of geospatial data, we also create additional features.
  - Number of hexes reachable within 15 min.
  - Number of stations reachable within 15 min.
  - Furthest point reachable within 60 min.
  - Total track length coverable within 60 min.



Data Summary

- **Our full collection of 79 explanatory features spans all four levels:**
    1. **unit variables**: ***number of rooms, unit floor, surface area***
    2. **building variables**: ***building floors, build year, closest station time, age in months***
    3. **neighborhood variables**: ***store score, station area score, number of buildings, percent area covered by buildings, mean building surface area, embassy scores, unpleasant score, percent of each zoning type, percent of each land use type, percent of each vegetation type, reachable hexes 15min, number of nearby stations***
    4. **wider accessibility variables**: ***estimated employee demand, station accessibility score***
- **Our prediction target is *log adjusted rent per square meter* (including fees and key money)**

The Analysis

Methods

- **Our analysis uses gradient boosted decision trees from LightGBM (LGBM) for the prediction tasks in three cases:**
    1. **The model using *longitude*, *latitude*, and *station IDs* (implicit).**
    2. **The model using my geospatial data.**
    3. **The model using both.**

- **Recognizing the temporal aspect of rent prices…**
    - **We use the last 30 days of data as the test set.**
    - **This gives us 750,411|47,826 split.**

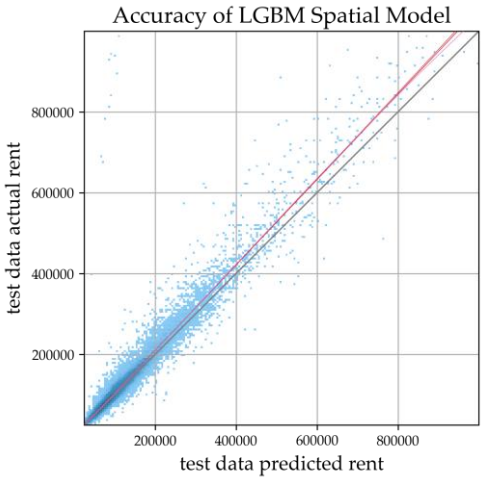
(We also do spatial holdout experiments.)

Main Results

- **The LGBM yields a MAPE of 7% and an R^2 of 0.934.**
- **The LGBM performance is practically identical for all spatial data types.**
- **Using only room and building data, the LGBM yields a MAPE of 15.696% and an R^2 of 0.802, so clearly the geospatial relationships are playing an important role.**
- **These results show that the rich geospatial data accounts for all the spatial features implicitly captured using station IDs and coordinates.**

TABLE I
ACCURACY COMPARISON OF OUR OLS, NN, AND LGBM MODELS

model	MAPE %	MAE ¥	R^2
OLS	11.698	15,085	0.804
NN	8.619	10,512	0.916
LGBM-implicit	7.059	8,641	0.934
LGBM-spatial	7.030	8,600	0.934
LGBM-both	7.006	8,568	0.935



Feature Importance

- **We measure feature importance using a variety of methods, but we report the results from *permutation feature importance* from the *Scikit Learn* package.**
- **This technique randomly re-sorts the values for each feature, allowing one to measure the resulting change in prediction values (similar to a variable dropout analysis).**
- **Unlike the "split" and "gain" importance measures provided by LGBM, or SHAP values, permutation feature importance is model-agnostic and clearly interpretable.**
- **Primarily we are interested in:**
    1. **The relative importance of geospatial and implicitly spatial features when both are included**
    2. **Whether geospatial estimation analyses really do improve when integrating information from all four levels**
    3. **Which kinds of geospatial features are most relevant to rental prices.**

Feature Importance

- **When including both explicitly and implicitly spatial features, the implicit features DO appear in the top 15.**
- **Features from all four levels DO appear in the top 15, with a wider accessibility feature in the #2 spot.**
- **Room and building features are generally among the strongest features.**
- **Neighborhood features play a minor role; perhaps redundancy is to blame.**
- **Among the neighborhood features, economic and status variables come out on top…correlated with being near city centers.**

TABLE II
LARGEST FEATURE EFFECTS ON LGBM-BOTH PREDICTION

rank	feature	MAPE %	MAE ¥
1	unit surface area	20.459	24, 938
2	estimated employee demand	11.013	11, 966
3	built year	8.111	8, 718
4	age in months	7.830	8, 577
5	**closest station id**	6.004	6, 669
6	building floors	5.535	6, 869
7	number of rooms	5.445	5, 394
8	unit room floor	3.195	3, 715
9	number of companies	2.990	3, 463
10	number of jobs	2.656	3, 421
11	**longitude**	2.214	2, 153
12	embassy score	2.027	3, 229
13	**latitude**	1.876	2, 049
14	embassy score log	1.498	2, 567
15	veg=field	1.325	1, 569

Feature Importance

- **When including only explicitly geospatial features, the top 10 are very similar.**
- **Population variables enter the top 15.**
- **Building size also appears.**
- **Time to closest station (thought to be a key feature in Tokyo) appears as #11 here.**

- **Even though the prediction target is rent per square meter, the unit surface area is the strongest determinant of price because there is a strong non-linear relationship.**

TABLE III
LARGEST FEATURE EFFECTS ON LGBM-SPATIAL PREDICTION

rank	feature	MAPE %	MAE ¥
1	unit surface area	20.512	25,265
2	estimated employee demand	12.568	13,257
3	age in months	7.998	8,732
4	built year	7.966	8,533
5	building floors	5.623	6,959
6	number of rooms	5.129	5,134
7	number of jobs	3.759	4,512
8	unit room floor	3.234	3,726
9	area building surface area	2.395	2,404
10	embassy score log	2.360	3,948
11	closest station time	1.980	2,013
12	veg=field	1.932	2,426
13	population percent female	1.790	1,998
14	number of companies	1.719	1,988
15	population percent house	1.487	1,525

Conclusions

- **Using a hex grid network with traversal times set from the walking traversal times creates a highly accurate simplified geospatial network that enables the detailed analysis of unprecedently large physical areas.**
- **Propagation on that hex network can be used to create novel accessibility features.**
- **Using rich geospatial data can fully replace implicitly spatial variables typically used real estate prediction models without any accuracy penalty.**
- **All four levels of variables are important for price prediction (and presumably other tasks).**
- **One wider accessibility variable (estimated demand) is especially important.**
- **Patterns in the prediction residuals reveal that additional data would likely improve the results, but additional geospatial data is unlikely to improve results; thus, more detailed property data is likely needed. Our in-house analyses confirm this.**

Demand Analysis Expansions

- Include travel by bus, automobile, and bicycle.
- Tweek hex network connectivity to eliminate biased error in route time estimation.
- Determine reasonable multimodal routes beyond walking to the stations and bus stops.
- Include additional sources of demand:
    - other job types
    - include estimated students at colleges and universities
    - access to shopping based on store counts
    - access to amenities such as parks, sports centers, shinkansen, airports, hospitals, etc.
- Expand coverage to the whole country (or at least our service areas).

Discussion

Aaron Bramson a_bramson@ga-tech.co.jp

Supplementary Materials

Network Refinements in Progress

- Enhancing the road network to include slopes for mobility-adjusted accessibility and more accurate traversal times and effort.
- Extracting rich survey data to determine road widths and the presence of sidewalks; these relate to safety and estimating speed limits for cars.

Route Composition Matters for Hedonic Modeling

- **The demand score only considers total time from home to work; however, the two options:**
  - **20 minute walk and 10 minutes by train, and**
  - **5 minutes walk and 25 minutes by train**

    are not actually equivalent.
- **Given the same travel time, people prefer trains to buses.**
- **People will ride longer overall in order to avoid transfers.**
- **Longer walks are tolerated at the begining and end of trips, but not in the middle (e.g. 10-minute walk transfers).**

From Opaque to Explicable

- While machine learning can make fairly accurate predictions using station IDs and coordinates in place of geospatial data, these variables lack meaningful interpretations and render the model location-specific.
- Even though the specific structure of NN and LGBM models are largely inscrutable, using intuitive variables means that the results can still be interpreted by nonspecialists (in terms of accuracy and feature importance).
- In this application, when setting rent prices, this rich foundation allows one explain price differences using features with intuitive relationships to price.
- It still begs certain questions, like why the *percent females* or *proximity to embassies* has the observed effect on rent prices, perhaps they are just correlated with the actual causal factors.
- But unlike coordinates and named areas, these deeper questions can potentially be answered, although that would probably require even richer data.

Spatial Holdout

- **One purported advantage of the explicitly geospatial data is that the model should be more portable to other locations.**
- **So we perform a spatial holdout test by using two administrative areas as the test data instead of using the last 30 days.**

TABLE IV
SPATIAL HOLDOUT ACCURACY COMPARISON

	Shinjuku		Suginami	
model	MAPE %	R^2	MAPE %	R^2
OLS	10.655	0.84	10.518	0.81
NN	9.975	0.903	8.612	0.869
LGBM-implicit	10.168	0.879	8.703	0.875
LGBM-spatial	9.455	0.890	8.717	0.867
LGBM-both	9.549	0.896	8.731	0.875

- **The LGBM models perform worse on the spatial holdout compared to the temporal holdout, but drops least when using only the geospatial data.**
- **This means that, even when relying on informative geospatial characteristics, the predictions may depend on nuanced variations within the training regions.**
- **The OLS model improves on the spatial holdout, and the NN is comparable to LGBM.**
- **The Shinjuku (a city center) holdout is more difficult than the Suginami (suburban) holdout for all models.**

Other Applications

- **Imagine getting turned down for a loan based on an ML algorithm.**
  - **When asked for the reasons, the bank could (maybe) tell you the variables that contributed the most to your risk score (using something like SHAP).**
  - **But what if features like postal code, time at current job, and height are offered as explanation?**
  - **Do these variables have plausible causal impacts on loan risk?**

SHAP Feature Importance

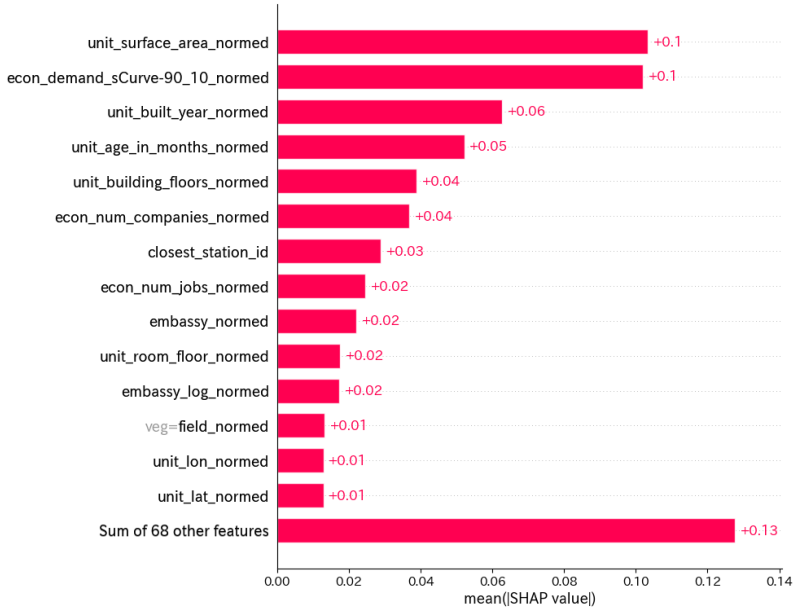
- **SHAP is a popular method of measuring feature importance in machine learning models.**
  - **It measures the effects of values in a particular row in comparison to the whole dataset, thus it can tell you how particular values in a sample change the predicted value compared to the predictions of all other samples.**
  - **Permutation feature importance is based on the decrease in model performance, whereas SHAP is based on the magnitude of feature attributions.**
    - **That is, a high average SHAP score would tell us which feature changes the price the most, but not whether those changes were important for accurately predicting prices.**

SHAP Feature Importance

- **However, the same features are revealed as the most important, but at slightly different levels.**

TABLE II
LARGEST FEATURE EFFECTS ON LGBM-BOTH PREDICTION

rank	feature	MAPE %	MAE ¥
1	unit surface area	20.459	24,938
2	estimated employee demand	11.013	11,966
3	built year	8.111	8,718
4	age in months	7.830	8,577
5	**closest station id**	6.004	6,669
6	building floors	5.535	6,869
7	number of rooms	5.445	5,394
8	unit room floor	3.195	3,715
9	number of companies	2.990	3,463
10	number of jobs	2.656	3,421
11	**longitude**	2.214	2,153
12	embassy score	2.027	3,229
13	**latitude**	1.876	2,049
14	embassy score log	1.498	2,567
15	veg=field	1.325	1,569



SHAP Feature Importance

- **However, the same features are revealed as the most important, but at slightly different levels.**