Combining Machine Learning and Econometrics Application to Commercial Real Estate Prices

Marc Francke* and Alex van de Minne**

Tokyo, January 2024

Seminar at Hitotsubashi University

* University of Amsterdam & Ortec Finance

** University of Connecticut

Motivation & Contribution

Motivation (1)

- Modeling CRE transaction prices. Different dimensions:
  1. Cross-section
  2. Spatial patterns
  3. Temporal dynamics
  4. Repeat transactions of the same property
- Item (1)
  - Functional form of property characteristics (size, type, construction year, . . .) unknown, probably nonlinear and interaction effects

$\rightarrow$ use ML algorithms

- Items (2)–(4)
  - Violate i.i.d. assumption, implicitly assumed in standard ML approaches
  - The field of time-series, spatial, and panel data econometrics

$\rightarrow$ Use econometric modeling

Motivation (2)

- Econometric models
  - Requires a priori specification of Data Generating Process (DGP)
    - transformations of variables and potential interaction effects, and
    - distributional assumptions on the error term.
  - Focus is on parameter estimates (causality)
    - credible intervals
  - Enables formal testing of hypothesis
- ML algorithms
  - No DGP specification is required
    - model's structure and parameter values simultaneously (Athey, 2018)
  - Focus on out-of-sample prediction
    - ML algorithms can easily outperform *simple* econometric models
    - No testing possible
  - Hungry for data, both in observations and features; the more, the better
- CRE: Small number of sales and features, and very heterogeneous

  A lot of unobserved heterogeneity $\rightarrow$ Property random effect

Contribution (1)

- Is it possible to combine the two cultures of statistical modeling? Yes
- Different components:
  - Econometric part
    1. Common trend: random walk
    2. Property type specific trends: random walk
    3. Spatial effects: Besag model (Besag and Kooperberg, 1995)
    4. Property effects: random effects
  - Machine Learning part
    5. Property characteristics
- Some ML algorithms, like NN, are parametric models
  - Can be estimated by likelihood based or Bayesian methods
  - Computationally not the most efficient method

Contribution (2)

- However, most ML are non-parametric
- Iterative procedure works quite well
  1. Estimate the joint model with a linear function for the property characteristics component (5).
  2. Take the ML ‘residual’ as the observed value minus the estimates of components 1 to 4, and calibrate a ML algorithm
  3. Take the DGP ‘residual’ as the observed value minus the prediction of component 5, and estimate the econometric model, consisting of components 1 to 4.
  - Repeat these steps until convergence occurs
- No guarantee of convergence to ‘true’ value, however
  - Check for neural network: full Bayesian and iterative approach
  - Simulate data with some nonlinear function and check whether iterative procedure finds the same components (not in paper, has been done for residential properties)

Contribution (3)

- Out-of-sample model performance
  - Fully econometric model: worst performance
  - ML algorithm only: second best
  - Mixed model: best
- Mixed model has components that are easy interpretable
  - Location values: heat maps
  - Constant quality price indexes
- Iterative procedure works quite well, also in small samples
  - The differences in results between the full Bayesian estimation, and the iterative approach is negligible

Related literature

Related literature

- Econometric part
  - Structural time series part: replace time fixed effects by stochastic processes (among others, Francke and De Vos, 2000; Bollerslev, Patton, and Wang, 2016)
  - Spatial structure and random effects (Francke and Van de Minne, 2021).
- A few examples of ML for real estate
  - Kok, Koponen, and Martínez-Barbosa (2017): AVM for CRE
  - Pace and Hayunga (2020): Analyze residuals from HPM & spatial models by ML
  - Deppner and Cajias (2022): HPM for RRE and spatial dependence
  - Deppner et al. (2023): AVM for CRE and feature importance
  - Lorenz et al. (2023): Residential rents and interpretation of ML algorithms
- Interpretable ML
  - Mullainathan and Spiess (2017): ML algorithms lack explainability of predictions due to their complex structure that varies for each repeated calibration
  - Despite a substantial body of iML literature (Molnar, 2019), there is still a large gap with econometrics in terms of complexity and standardization of methods.

Model specifications

Neural Network

- Single-layer feed-forward network, NN

$$y_i = f^{NN}(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{1a}$$

$$f^{NN}(x_i) = \lambda_0 + f(a_{1,i})\lambda_1 + \cdots + f(a_{M,i})\lambda_M, \tag{1b}$$

$$a_{m,i} = \omega_{1,m} x_{i,1} + \cdots + \omega_{K,m} x_{i,K} + \omega_{0,m}, \tag{1c}$$

$$f(a_{m,i}) = \left(1 + \exp(-a_{m,i})\right)^{-1}, \tag{1d}$$

- Hidden node: $f(a_{m,i})$ with corresponding coefficients λ_m, $m = 1, \ldots, M$
- Activation: $a_{m,i}$ linear function of indep. vars. with weights $\omega_{k,m}$, bias $\omega_{0,m}$
- Activation function, in our case a sigmoid
- Comments
    - Transform variables: Take log and demean
    - How many hidden nodes? We took $M = K \times 2 + 1$
    - Multi-modality: use different starting values

Random Effects model

- REM

$$
\begin{align}
y_{itp} &= x'_{itp}\beta + \mu_t + \delta_{tp} + \theta_i + \phi_i + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon), \tag{2a}\\
\mu_t &\sim \mathcal{N}(\mu_{t-1}, \sigma^2_\mu), \quad u_{t=1} = 0, \tag{2b}\\
\delta_{t,p} &\sim \mathcal{N}(\delta_{t-1,p}, \sigma^2_\delta), \quad \delta_{t=1,p} = 0, \tag{2c}\\
\theta_i|\theta_{-i}, \sigma^2_\theta &\sim \mathcal{N}\left(\frac{1}{m_i}\sum_{q\in\Omega_i} w_{iq}\theta_q, \frac{\sigma^2_\theta}{m_i}\right), \sum_{i=1}^{I}\theta_i = 0 \tag{2d}\\
\phi_i &\sim \mathcal{N}(0, \sigma^2_{\epsilon_\phi}). \tag{2e}
\end{align}
$$

- $x'_{itp}\beta$: linear specification for property characteristics
- μ: common trend; δ: property type trend, RW
- θ: spatial random effect: $w_{iq} = 1$ when distance $\leq$ 800m, 0 otherwise
- ϕ: property random effect

Combining NN and REM

- Full Bayesian approach: fNNREM
    - Replace $x'_{itp}\beta$ in Eq. (2a) by neural network $f^{NN}(x_{itp})$ in Eq. (1)
- Iterative approach: iNNREM
    0. **Initialize:** Estimate REM (Eq. 2) including linear term
    1. – **Train** ML algorithm on $\{y_{itp} - \hat{\mu}_t - \hat{\delta}_{tp} - \hat{\theta}_i - \hat{\phi}_i, x_{itp}\}$, gives $\widehat{f(x_{itp})}$
       – **Compute** $y^*_{itp} = y_{itp} - \widehat{f(x_{itp})}$
    2. **Estimate:** REM (Eq. 2) without linear term $x'_{itp}\beta$ with dependent variable y^*_{itp}
    - Repeat 1 and 2 until convergence in e.g. MAPE
- In the iterative approach NN can be replaced by any ML algorithm for supervised learning

Estimation

- Packages available in R / Python / Stan
  - NN: iterative backpropagation algorithm (Bishop and Nasrabadi, 2006)
  - REM: Integrated Nested Laplace Approximation (INLA) due to its computation speed (Rue, Martino, and Chopin, 2009).
    Only applicable for Gaussian Markov Random Fields.
  - fNNREM: No-U-Turn-Sampler (Hoffman and Gelman, 2014)
  - iNNREM: Combination of INLA and backpropagation algorithm
- Out-of-sample performance
  - 10 K-folding: train on 90% of data, and predict 10%
  - 10 times sampling without replacement
- Quality of the index
  - Take the average of index values
  - Show the range: min to max

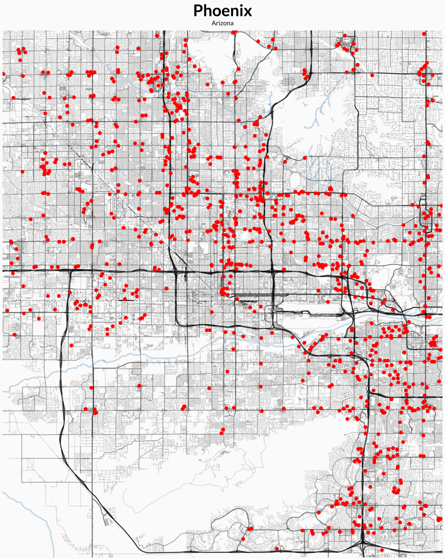
Data

Data

- Source: MSCI/RCA
- CRE in Phoenix: 2,652 pref-filtered transactions, period: 2001 – 2021
- NOI, property type (apartment / industrial / office / retail), age, latitude & longitude, unique property identifier, and walk score

Statistic	**Mean**	**St. Dev.**	**Min**	**Max**
Sales price	$ 17,702,190	$ 21,448,988	$ 787,000	$ 280,000,000
NOI (p. Sqft)	$ 11.309	$ 8.837	$ 1.239	$ 115.914
Building size (Sqft)	129,127	132,461	2,000	1,366,600
Age of building	19.627	13.572	1	78
Walk score	47.006	16.470	0	96
Property type count	Apartment	Industrial	Office	Retail
	1,174	263	484	731

Data



- Figure covers about 70% of transactions
- No clear center. Black square (representing highways) in the middle typically seen as CBD
- Little transactions took place there
- Phoenix's CRE is “spread out” (growth and little supply constraints)

Results

Results

1. Out-of-sample model fit
2. Heat map
3. Indexes

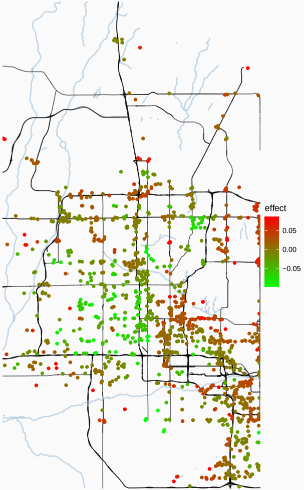
Model fit (1)

	(1) **NN**	(2) **REM**	(3) **fNNREM**	(4) **iNNREM**
MAPE, out of sample, all	0.139	0.147	0.111	0.109
MAPE, out of sample, k = 0	0.142	0.157	0.111	0.114
MAPE, out of sample, k = 1	0.136	0.138	0.113	0.106
MAPE, out of sample, k >1	0.133	0.117	0.102	0.090
MAPE, in sample, all	0.127	0.095	0.104	0.079
Obs, all		2,652		
Obs, k = 0		1,694		
Obs, k = 1		647		
Obs, k >1		311		

Model fit (2)

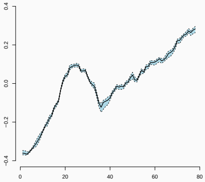
- MAPE for NN is lower than for REM
  - except for properties that have at least 1 sale in training set
  - REM controls for unobserved heterogeneity
- In-sample MAPE is lower compared to the out-of-sample one for REM
  - Overfitting due to property random effects
- fNNREM and iNNREM perform similarly, and better than both NN and REM
  - High correlation in predicted values & residuals
  - Iterative: much faster and easier to generalize
- Less overfitting in fNNREM

Heat map

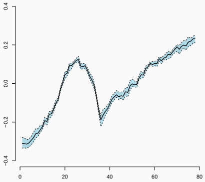


- Heat map of θ
- Red is expensive
- Range is appprox. -10% to + 10%
- Note that this conditional on NOI

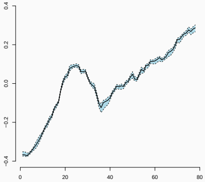
Indexes



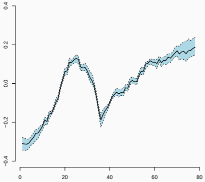
Industrial, REM



Industrial, iNNREM



Office, REM



Office, iNNREM

- NN does not provide indexes
- Left: REM; Right: iNNREM
- Indexes also available for Apartments and Retail
- Note that this conditional on NOI
- Model includes rents, so index related to cap rate
- REM: more co-movement between property types
- iNNREM: more uncertainty in index values

Robustness checks

Robustness check: Limited features (1)

- Limited features (without NOI and walk score)
  - REM are good at explaining away UH and are less affected by dropping variables having a spatial and/or time component.
  - These variables are difficult to get by in many other countries, asset types, or lower level geographies
- MAPE increase
  - NN: from 0.14 to 0.31
  - REM: from 0.15 to 0.23
  - iNNREM: from 0.11 to 0.15
- Heat-map: range of spatial effect
  - Full set: -10% to +10%,
  - Limited set: -40% to +20%,

Robustness check: Limited features (2)

Contribution of the Individual Elements to the Total Fit of the iNNREM, for Both Full and Reduced Datasets.

		(1)	(2)
Description	Parameter	**Full**	**Reduced**
Common trend	$\hat{\mu}$	15.36%	14.94%
Property type subtrends	$\hat{\delta}$	2.92%	8.46%
Spatial random effect	$\hat{\theta}$	3.96%	11.55%
Property random effect	$\hat{\phi}$	0.07%	2.80%
Property characteristics	$\widehat{f(x_{itp})}$	77.69%	62.26%

Robustness checks: Including time and space in NN

- Adding dummy variables for locations and quarters: too many parameters
  - Add linear trend and use coordinates

$$s^x = \cos(\text{Lat}) \times \cos(\text{Lon}), s^y = \cos(\text{Lat}) \times \sin(\text{Lon}), s^z = \sin(\text{Lat}).$$

	(1) **Full**	(2) **Reduced**
MAPE, out of sample, all	0.121	0.250
MAPE, out of sample, k = 0	0.125	0.265
MAPE, out of sample, k = 1	0.116	0.239
MAPE, out of sample, k >1	0.114	0.190
MAPE, in sample, all	0.095	0.214

- Fit improves, but not as good as iNNREM, specifically when a property got sold more than once
- Not possible to provide indexes and heat maps

Robustness checks: Other ML algorithms

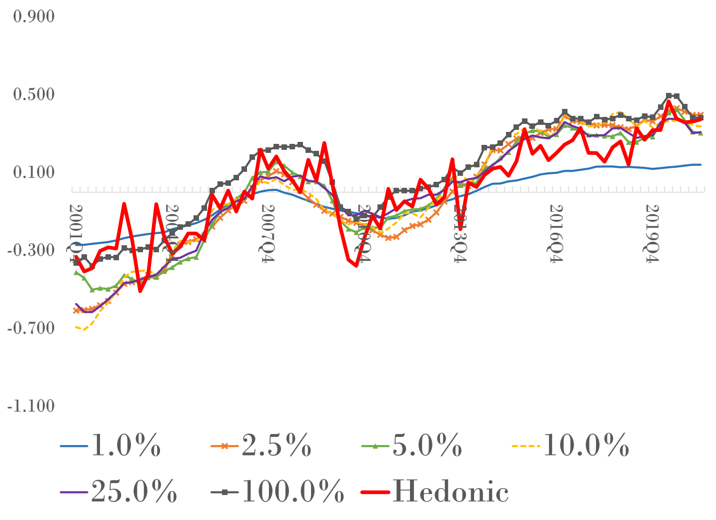
	(1) iSVM	(2) iRT	(3) iGBM
MAPE, out of sample, all	0.110	0.116	0.116
MAPE, out of sample, k = 0	0.114	0.122	0.120
MAPE, out of sample, k = 1	0.107	0.114	0.112
MAPE, out of sample, k >1	0.092	0.092	0.099
MAPE, in sample, all	0.085	0.085	0.082

- Similar performance compared to iNNREM

Robustness checks: sampling subsets (1)

- Select New York metro area: around 24,000 observations
- Sample without replacement 25% (6,250 obs), 10% (2,400 obs), 5% (1,200 obs), 2.5% (600 obs) and 1% (240 obs) from this data and re-estimate the subtrends.
- Proposed methodology produces reliable and stable indexes even in small sample environments (from 2.5%)
- Example for offices

Robustness checks: sampling subsets (2)



Conclusions

Conclusions

- Main goal: illustrate that it is possible to combine econometrics and ML
- Iterative approach works quite well
  - Check with full Bayesian approach (400 times slower)
  - Simulated data using nonlinear function for property characteristics
- Hybrid approach has the advantage (over ML only) of providing indexes and heat maps
- Adding random effects improves the out-of-sample prediction for properties having at least 1 sale in the training set
- Augmenting the baseline NN with spatio-temporal variables, does improve its fit, but it is not as good as the hybrid approach
- Room for improvement
  - All components can be refined
  - Ensemble of models

References

Athey, S. (2018). “The impact of machine learning on economics”. In: *Economics of Artificial Intelligence*. University of Chicago Press.

Besag, J. and C. Kooperberg (1995). “On conditional and intrinsic autoregressions”. In: *Biometrika* 82.4, pp. 733–746.

Bishop, C. M. and N. M. Nasrabadi (2006). *Pattern recognition and machine learning*. Springer.

Bollerslev, T., A. J. Patton, and W. Wang (2016). “Daily house price indices: Construction, modeling, and longer-run predictions”. In: *Journal of Applied Econometrics* 31.6, pp. 1005–1025.

Deppner, J. and M. Cajias (2022). “Accounting for spatial autocorrelation in algorithm-driven hedonic models: A spatial cross-validation approach”. In: *The Journal of Real Estate Finance and Economics*, pp. 1–39.

Deppner, J., B. von Ahlefeldt-Dehn, E. Beracha, and W. Schaefers (2023). “Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach”. In: *The Journal of Real Estate Finance and Economics*, pp. 1–38.

References ii

Francke, M. K. and A. F. De Vos (2000). “Efficient Computation of Hierarchical Trends”. In: *Journal of Business and Economic Statistics* 18, pp. 51–57.

Francke, M. K. and A. M. Van de Minne (2021). “Modeling unobserved heterogeneity in hedonic price models”. In: *Real Estate Economics* 49.4, pp. 1315–1339.

Hoffman, M. D. and A. Gelman (2014). “The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.”. In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Kok, N., E. L. Koponen, and C. A. Martínez-Barbosa (2017). “Big Data in real estate? From manual appraisal to automated valuation”. In: *The Journal of Portfolio Management* 43.6, pp. 202–211.

Lorenz, F., J. Willwersch, M Cajias, and F. Fuerst (2023). “Interpretable machine learning for real estate market analysis”. In: *Real Estate Economics* 51.5, pp. 1178–1208.

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub, https://christophm.github.io/interpretable-ml-book/.

Mullainathan, S. and J. Spiess (2017). “Machine learning: An applied econometric approach”. In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Pace, R. K. and D. Hayunga (2020). “Examining the information content of residuals from hedonic and spatial models using trees and forests”. In: *The Journal of Real Estate Finance and Economics* 60, pp. 170–180.

References iii

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.