Big data, machine learning, and econometrics: applications to real estate

Marc Francke

m.k.francke@uva.nl

University of Amsterdam and Ortec Finance

June 12, 2019

BSCPM Doctoral Conference 2019

Outline

1. Big data
2. Statistical modeling
3. Algorithmic modeling
4. Examples
   - AVM
   - Street View
   - Text data
   - Clicks on internet pages
5. Concluding remarks
6. References

Big data

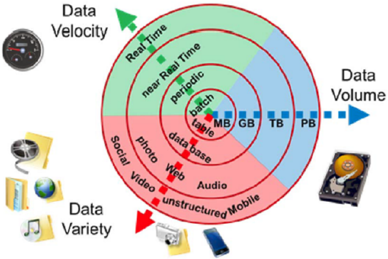
Big data and Machine learning (ML) are buzzwords

‘Big data’ is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing software applications. (Source: wikipedia)

What is so different about big data?

The four V's (source: wikipedia)

1. **Volume** ($n = \infty$);
2. **Variety**: text, images, audio, video; it completes missing pieces through data fusion;
3. **Velocity**: frequency of generation and frequency of handling, recording, and publishing;
4. **Veracity**: data quality of captured data can vary greatly, affecting the accurate analysis.



"Walmart generates more than 2.5 petabytes (250 bytes) of data relating to more than 1 million customer transactions every hour."

What is the value of big data?

- Higher frequency
- More detailed information
- Narrower segmentation
  - compare to census data, household surveys, ...
- Big data complement, do not replace, small data
- Big data (text, images, audio, video, sensor data ...) require new techniques for analysis

Two cultures: statistical and algorithmic modeling

Statistical modeling

- Input $\rightarrow$ Statistical model $\rightarrow$ Output
- Input: transaction price, house size, number of rooms, ...
- Model specification / data generating process assumed to be known, depending on unknown coefficients
  - Functional form
  - Stochastic assumptions

$$y|X,\beta,\sigma \sim N(X\beta,\sigma I)$$

- Model estimation (of β,σ): OLS, WLS, maximum likelihood, minimize loss function, ...
- Output: Predicted values, coefficients, confidence bounds, (in sample) model fit, ...

Statistical modeling in economics

Econometrics: theory driven

- Econometrics is concerned with the empirical determination of economic laws (Theil, 1971, p. 1)
- It is the unification of statistics, economic theory, and mathematics (Greene, 2008, p. 1)
- Econometrics is the interaction of economic theory, observed data and statistical methods (Verbeek, 2008, p. 1)
- At a broad level, econometrics is the science and art of using economic theory and statistical techniques to analyze economic data (Stock and Watson, 2012, p. 43)

Statistical modeling

- Focus on parsimonious model specifications ($k << n$)
  - to identify parameters;
  - to avoid multicollinearity
- Model selection is typically behind the scenes:
  only a few models are being presented
- Commitment to stochastic data models (Breiman, 2001)
  - has led to irrelevant theory, questionable conclusions,
  - has kept statisticians from working on a large range of interesting current problems,
- ‘The belief in the infallibility of data models was almost religious. It is a strange phenomenon once a model is made, then it becomes truth and the conclusions from it are infallible.’

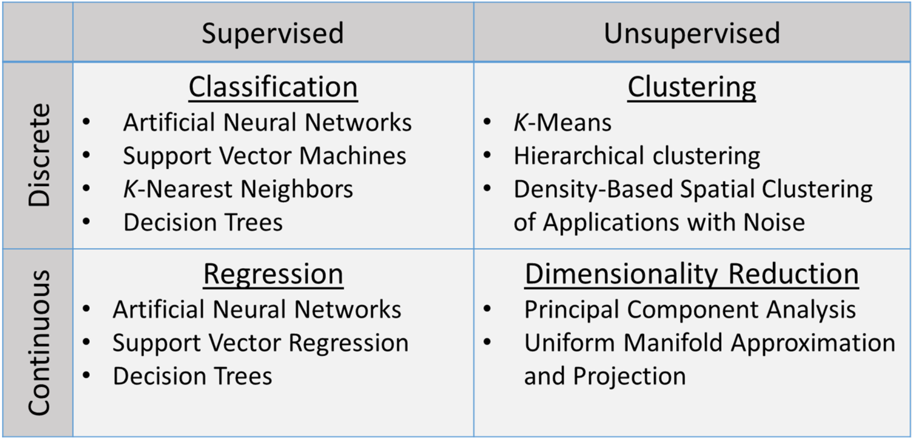
Breiman (2001), Statistical modeling: The two cultures

- Algorithmic modeling
  - can be used both on large complex data sets
  - gives a more accurate and informative alternative to data modeling on smaller data sets
  - If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools
- 'The criterion for any model is what is the predictive accuracy.'
- 'Accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors.'
- 'Instead of reducing dimensionality, increase it by adding many functions of the predictor variables.' (no curse of dimensionality)
- 'Higher predictive accuracy is associated with more reliable information about the underlying data mechanism. Weak predictive accuracy can lead to questionable conclusions.'

What is machine learning? (Athey, 2018)

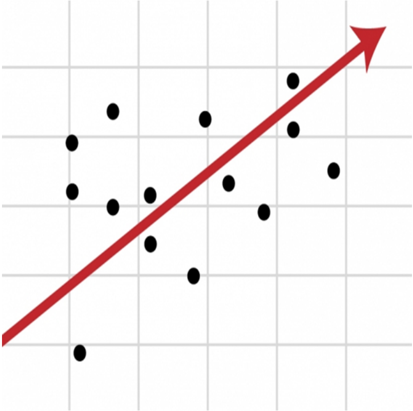
- Machine Learning (ML) is a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction, classification, and clustering or grouping tasks
- Unsupervised ML involves finding clusters of observations that are similar in terms of their covariates, and thus can be interpreted as dimensionality reduction
  - very useful as an intermediate step in empirical work in economics
  - create variables that can be used in economic analyses
  - outcome data is not used at all; thus, concerns about spurious correlation between constructed covariates and the observed outcome are less problematic.
- Supervised ML typically entails using a set of features or covariates (X) to predict an outcome (Y).
  - The observations are assumed to be independent, and the joint distribution of X and Y in the training set is the same as that in the test set.

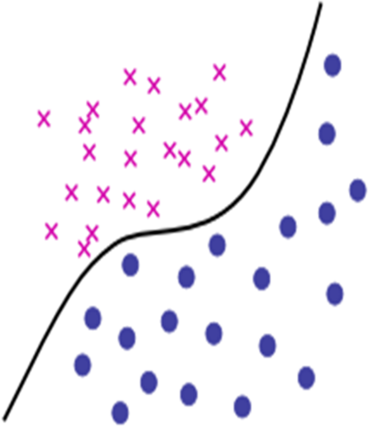
Machine learning



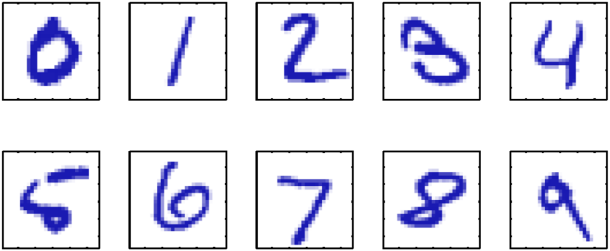
	Supervised	Unsupervised
Discrete	Classification • Artificial Neural Networks • Support Vector Machines • *K*-Nearest Neighbors • Decision Trees	Clustering • *K*-Means • Hierarchical clustering • Density-Based Spatial Clustering of Applications with Noise
Continuous	Regression • Artificial Neural Networks • Support Vector Regression • Decision Trees	Dimensionality Reduction • Principal Component Analysis • Uniform Manifold Approximation and Projection

Regression and classification

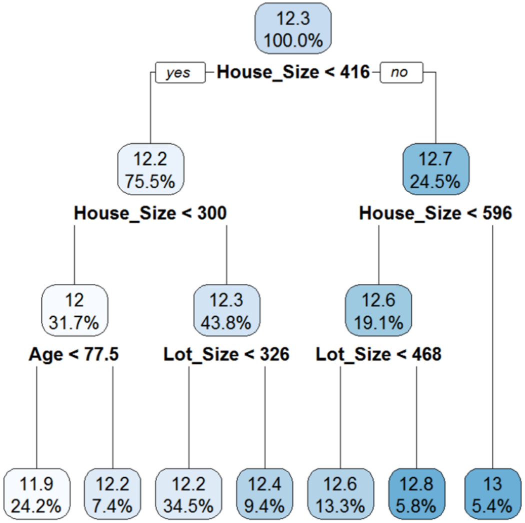




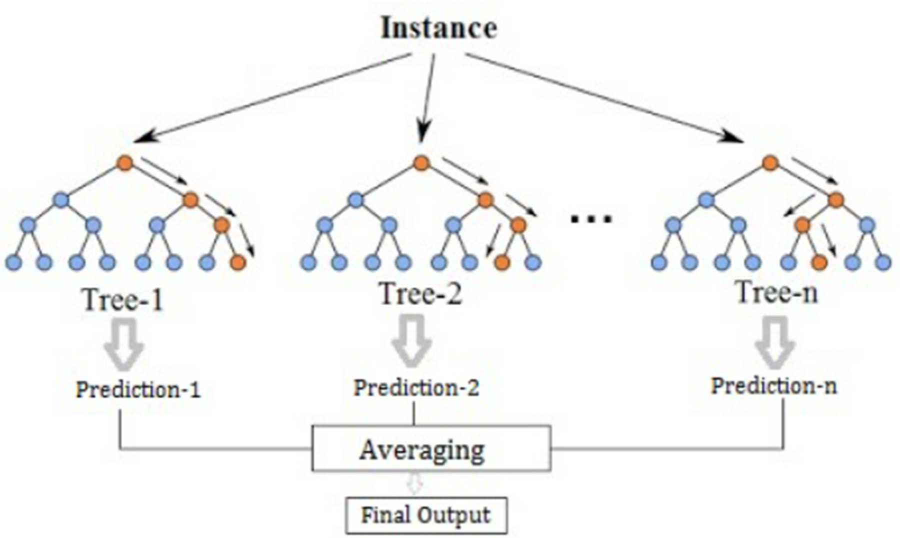
Classification



Decision trees (Random Forest en Gradient Boosting)

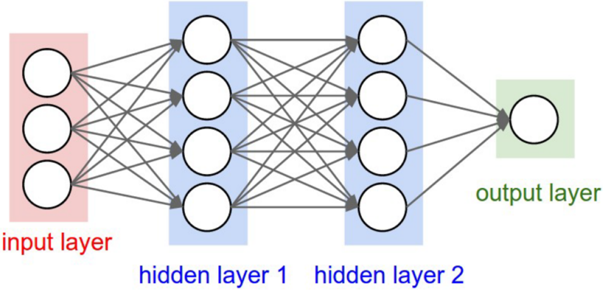


Random Forest

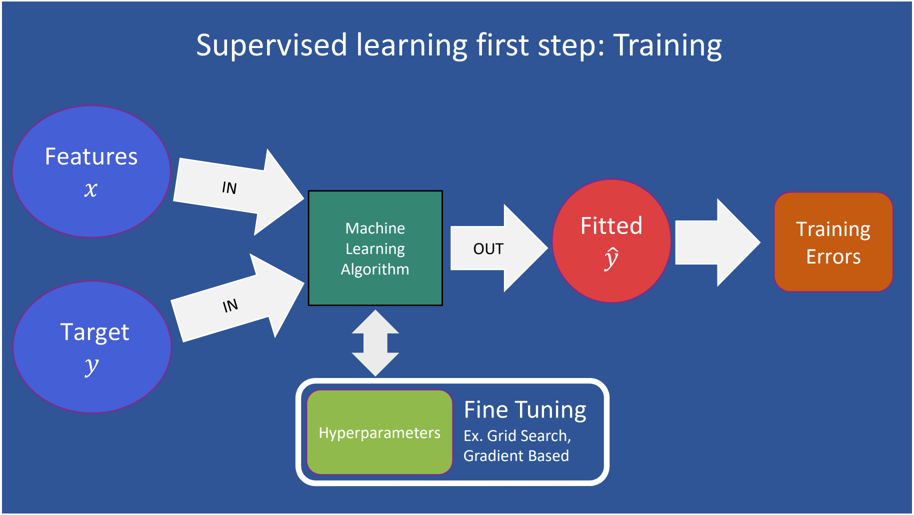


Neural networks

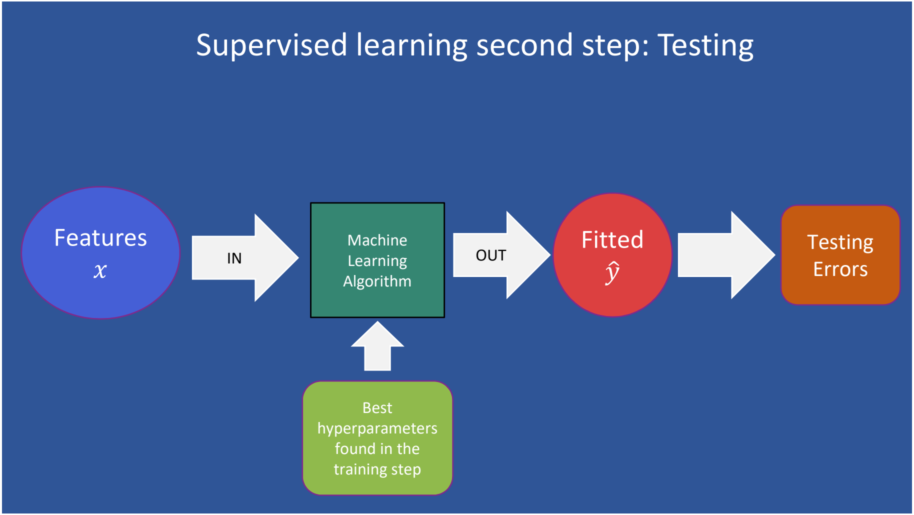
- Collection of computational units: artificial neurons
- Neurons ‘receive’ a weighted sum of outputs from preceding ‘layer’ (non-linear)
- ‘Learning’ is optimizing connection weights



Supervised Machine Learning: Training



Supervised Machine Learning: Testing



Supervised learning second step: Testing

Features
x
IN
Machine Learning Algorithm
OUT
Fitted
$\hat{y}$
Testing Errors
Best hyperparameters found in the training step

Varian, 2014, Big data: New tricks for econometrics

- Data analysis in statistics and econometrics can be broken down into four categories
  - 1) prediction, 2) summarization, 3) estimation, 4) hypothesis testing.
- ‘The focus of [supervised: MF] machine learning is to find some function that provides a good prediction of y as a function of x.’
- Focus on cross-section and data independently distributed
- ‘Usually “good” means it minimizes some loss function’
- ‘Our goal with prediction is typically to get good out-of-sample predictions.’
  - penalize models for complexity: regularization
  - split sample in training (estimate model), validation (choose model), and testing (evaluate model) set
  - if we have an explicit numeric measure of model complexity, we can view it as a parameter that can be “tuned” to produce the best out of sample predictions. (k-fold cross-validation).
- The test-train cycle and cross-validation should be used much more in economics, particularly when working with large datasets

Added value ML

- Machine learning algorithms are now technically easy to use: ‘off the shelf’ packages available in R and Python
- Systematic data cleaning procedures
- Scaling / Transformation / Normalization of variables requires expert opinion
- Feature engineering
  - Additional explanatory variables from text, pictures, audio, ...
  - These variables can be used within (econometric) model
- Flexible functional forms / many variables / interactions
  - “Simply including all pairwise interactions would be infeasible as it produces more regressors than data points. ... Machine learning searches for these interactions automatically ” (JEP Mullainathan and Spiess, 2017)
- Systematic model comparison
  - Econometrics’ practice: only the best performing model is showed

Weak points ML

- Observations are assumed to be independent, and the joint distribution of X and Y in the training set is the same as that in the test set. (Athey, 2018)
    - Time series / spatial / panel
    - Econometrics: Huge literature covering
        - non-stationary data, spurious regressions, unit root testing, Granger causality (predictive causality: predict the future values of a time series using prior values of another time series), co-integration, mixed frequencies, temporal aggregation, dynamic factor models, impulse-response functions, forecasting, ...
    - Mainstream ML literature: relatively silent
        - References to Bayesian structural time series / State Space / (Hidden) Markov models (Bishop, 2006; Murphy, 2012; Varian, 2014; Barber, 2017)
- Explainability: Many ML algorithms are hard to explain

Weak points ML: causality

- Goal ML: prediction, and that is different from explanation (IV-regression does not have highest R^2)

Athey (2018)

Suppose that an analyst has historical data from hotel prices and occupancy rates. Typically, prices and occupancy are positively correlated because the existing pricing policy for hotels ... specifies that hotels raise their prices as they become more fully booked. Off-the-shelf applications of SML techniques are designed to answer the following type of question: If an analyst is told that on a particular day, prices were unusually high, what is the best prediction of occupancy on that day? The correct answer is that occupancy is likely to be high. By contrast, the question of the effect of changing the pricing policy is a causal question, and common experience indicates that if the firm implemented a new policy to systematically raise prices by 5% everywhere, it would be unlikely to sell more hotel rooms. A different set of statistical techniques is required to answer this question, perhaps exploiting “natural experiments” in the data or an approach known as “instrumental variables”

Causality

- What is the effect of a treatment?
  - observed difference in outcome = average treatment effect on the treated + selection bias
  - challenge: to get rid of selection bias
- Econometrics/statistics developed methods to deal with causality
  - instrumental variables, regression discontinuity, diff-in-diff, and various forms of natural and designed experiments
  - Causal effect of treatment is not identified without making (non-testable) assumptions: Justification is important in research
- Instrumental variable approach
  - Linear equation $y_i = x_i'\beta + \varepsilon_i$
  - However, $E[x_i\varepsilon_i] \neq 0$ and $E[z_i\varepsilon_i] = 0$
  - Solution 2SLS:
    1. regress x on z resulting in prediction $\hat{x}$
    2. regress $\hat{x}$ on y to get estimate of β
  - First step is essentially a prediction problem: supervised ML
    - ★ weak instruments
    - ★ overfitting

Comparison

(Supervised) ML	Econometrics
Algorithms	Assumption of DGP (structure)
ML algorithms technically easy to use	Inference in advanced models difficult
Prediction $\hat{Y}$	Causal relations / associations $\hat{\beta}$
Functional form flexibility	Simple (rigid) models
Out-of-sample performance	In-sample fit statistics
Regularization (penalty complexity)	Information criteria (effective $\#$ of pars)
Many models (averaging)	One model
Lack of standard errors of coef.	Inference on coeffients possible
Similar predictions using different vars	Multicollinearity
Large databases & missing data	Smaller databases
Data IID, cross-sectional	Complex structures (time, space, panel)

Machine learning is alchemy

AI researchers allege that machine learning is alchemy

- ‘Rahimi [working for Google] charged that machine learning algorithms, in which computers learn through trial and error, have become a form of alchemy.’
- ‘Researchers, he said, do not know why some algorithms work and others don’t, nor do they have rigorous criteria for choosing one AI architecture over another.’
- I’m trying to draw a distinction between a machine learning system that’s a black box and an entire field that’s become a black box.’

Econometrics is alchemy

Hendry (1980, Economica, Econometrics - Alchemy or science)
Spurious regressions

Keynes: ‘No one could be more frank, more painstaking, more free from subjective bias or parti pris than Professor Tinbergen. There is no one, therefore, so far as human qualities go, whom it would be safer to trust with black magic. That there is anyone I would trust with it at the present stage, or that this brand of statistical alchemy is ripe to become a branch of science, I am not yet persuaded. But Newton, Boyle and Locke all played with Alchemy. So let him continue.’

Automated valuation model

- Precision of valuations:
  example for CRE (Kok, Koponen, and Martínez-Barbosa, 2017)
- Hybrid valuation model for RRE (Ortec Finance, 2017)
  - Development of hybrid automated valuation model: (time series and spatial) econometrics and ML
  - Focus on out-of-sample model performance (precision) and interpretability
  - Transaction prices of 740,000 houses in the Netherlands in the period 2009–2016
  - Model trained on transactions up to 2015
  - Out-of-sample prediction of transactions in 2016

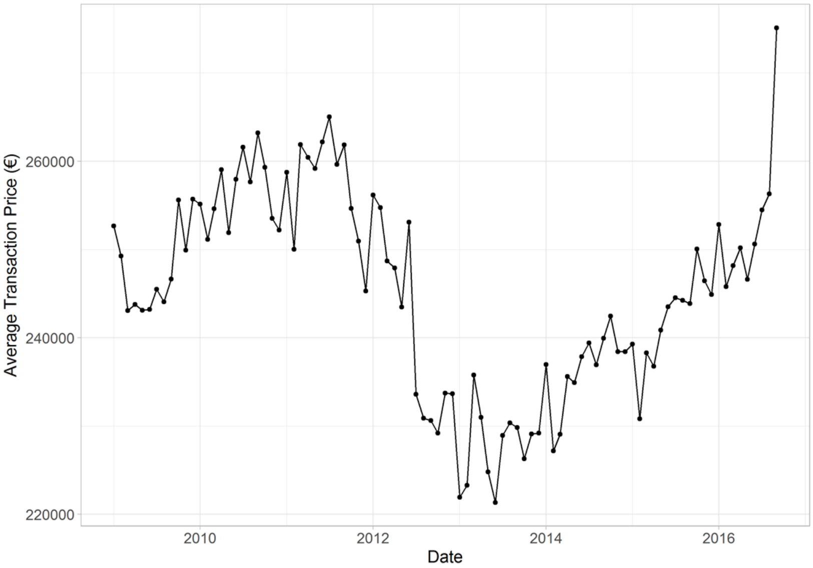
Hybrid valuation model

- (Econometric) time series model deals with temporal (and spatial) dependence
- ML deals with cross-section

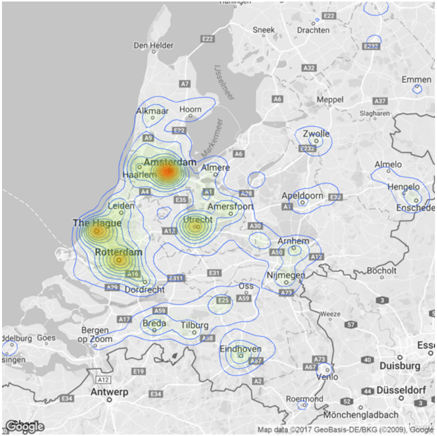
$$\begin{aligned}
y_t &= \mathbf{i}\mu_t + D_t\lambda_t + \dot{D}_t\theta_t + \ddot{D}_t\phi + f(X_t, \beta) + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2 I), \\
\mu_{t+1} &= \mu_t + \kappa + \eta_t, \qquad \eta_t \sim N(0, q_1\sigma^2), \\
\lambda_{t+1} &= \lambda_t + \varsigma_t, \qquad \varsigma_t \sim N(0, q_2\sigma^2 I), \\
\theta_{t+1} &= \theta_t + \omega_t, \qquad \omega_t \sim N(0, q_3\sigma^2 I),
\end{aligned}$$

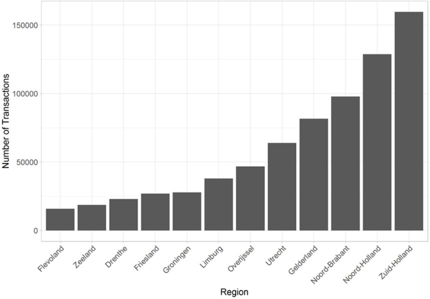
- y_t vector of log prices, μ_t common trend, λ_t house type trend, θ_t district trend (Time series model)
- X_t covariates (ML)
- iterative estimation procedure

Average house price

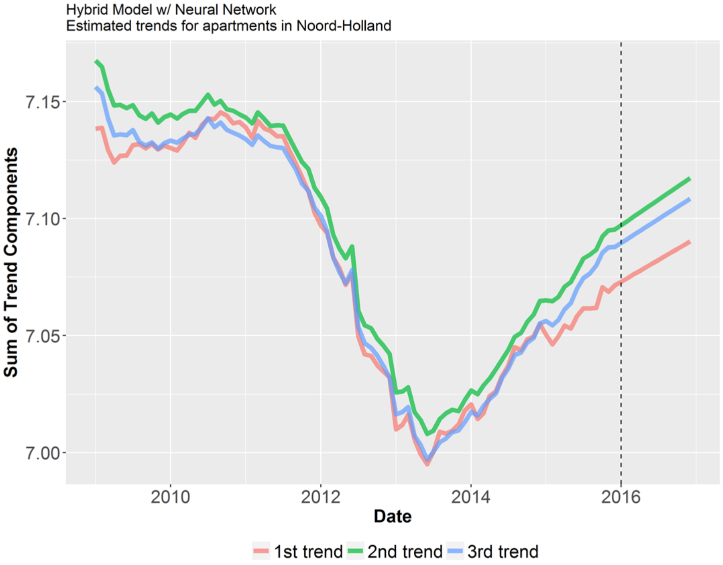


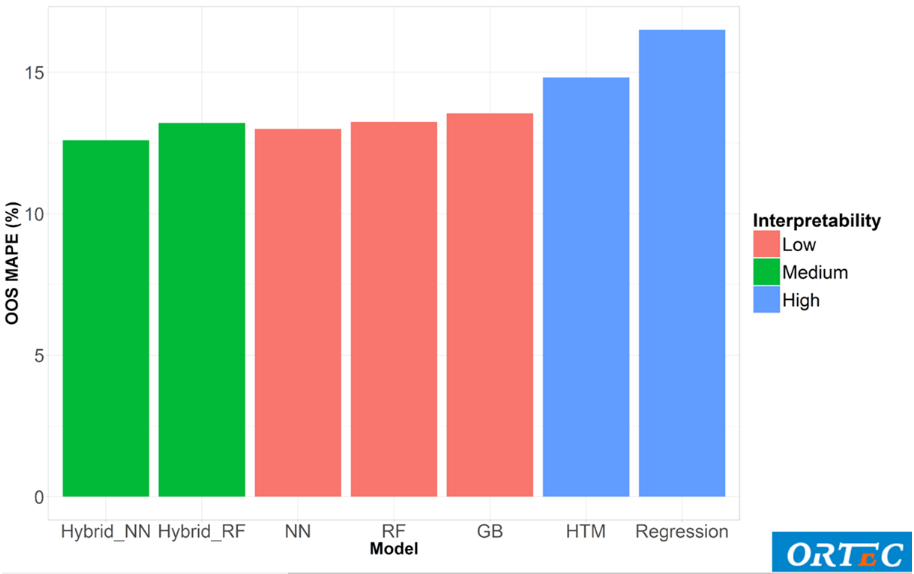
Data


Sneek
Drachten
Den Helder
IJsselmeer
Emmen
Meppel
Slagharen
Alkmaar
Hoorn
Markermeer
Zwolle
Amsterdam
Haarlem
Almere
Almelo
Hengelo
Enschede
Apeldoorn
Leiden
Amersfoort
The Hague
Utrecht
Arnhem
Rotterdam
Nijmegen
Bocholt
Dordrecht
Oss
Weeze
Breda
Tilburg
Bergen op Zoom
Goes
Eindhoven
Venlo
Duisburg
Düsseldorf
Antwerp
Roermond
Mönchengladbach
Google
Map data ©2017 GeoBasis-DE/BKG (©2009), Google


Number of Transactions
150000
100000
50000
0
Flevoland
Zeeland
Drenthe
Friesland
Groningen
Limburg
Overijssel
Utrecht
Gelderland
Noord-Brabant
Noord-Holland
Zuid-Holland
Region

Convergence of trend components



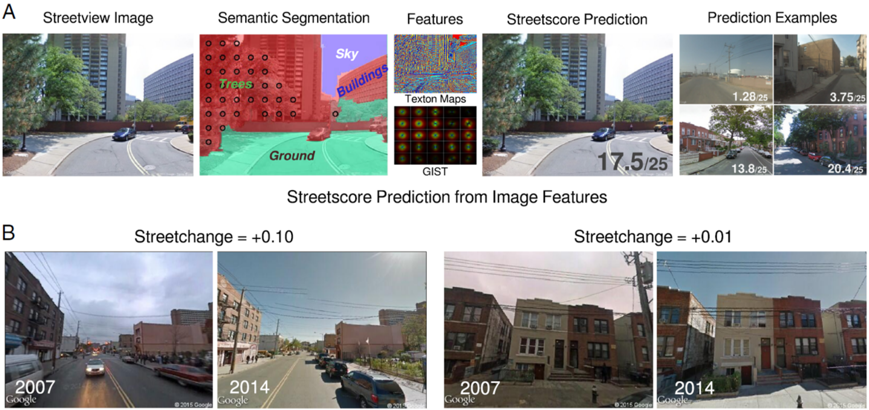

OOS MAPE (%)
15
10
5
0
Hybrid_NN
Hybrid_RF
NN
RF
GB
HTM
Regression
Model
Interpretability
Low
Medium
High

Street View

Naik et al. (2017): Computer vision uncovers predictors of physical urban change, Proceedings of the National Academy of Sciences

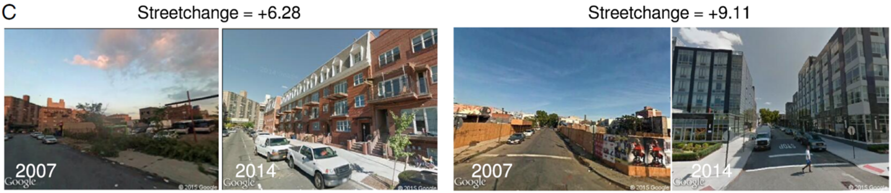
- Variable of interest: Streetchange
- Feature extraction:
  Measurement of physical urban change using Google Street View
- Baltimore, Boston, Detroit, New York, Washington DC
- Images classes: ground, buildings, trees, sky
- Streetscore: 0–25
- Validation by using surveys and external data

Street View



Streetscore Prediction from Image Features

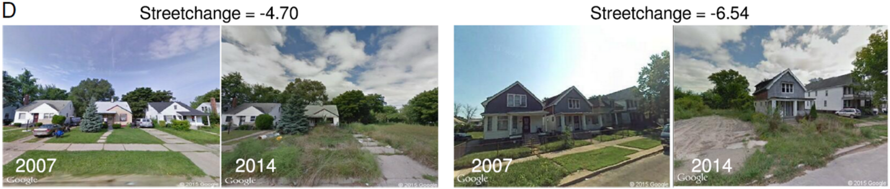
No Significant Change in Streetscore - Examples



Streetchange = +6.28

Streetchange = +9.11

Significant Improvement in Streetscore - Examples



Streetchange = -4.70

Streetchange = -6.54

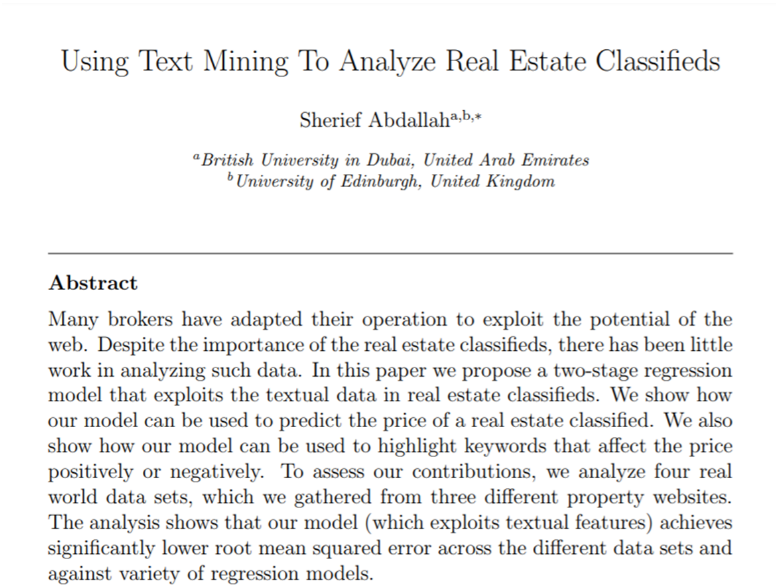
Significant Decline in Streetscore - Examples

Testing theories

- neighborhoods that are densely populated by college-educated adults are more likely to experience physical improvements' theory on human capital & local success
- neighborhoods with better initial appearances experience, on average, larger positive improvements – "tipping" theories of urban change
- neighborhood improvement correlates positively with physical proximity to the central business district and to other physically attractive neighborhoods – "invasion" theories of urban sociology

Only possible by using computer vision methods and street-level imagery (high spatial resolution compared to surveys, census data)

Text data



Using Text Mining To Analyze Real Estate Classifieds

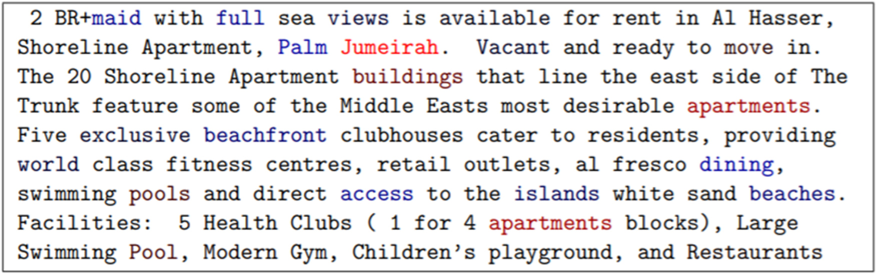
Sherief Abdallah[a,b,*]

[a] *British University in Dubai, United Arab Emirates*
[b] *University of Edinburgh, United Kingdom*

Abstract

Many brokers have adapted their operation to exploit the potential of the web. Despite the importance of the real estate classifieds, there has been little work in analyzing such data. In this paper we propose a two-stage regression model that exploits the textual data in real estate classifieds. We show how our model can be used to predict the price of a real estate classified. We also show how our model can be used to highlight keywords that affect the price positively or negatively. To assess our contributions, we analyze four real world data sets, which we gathered from three different property websites. The analysis shows that our model (which exploits textual features) achieves significantly lower root mean squared error across the different data sets and against variety of regression models.

Text: positive and negative impact on price



2 BR+maid with full sea views is available for rent in Al Hasser, Shoreline Apartment, Palm Jumeirah. Vacant and ready to move in. The 20 Shoreline Apartment buildings that line the east side of The Trunk feature some of the Middle Easts most desirable apartments. Five exclusive beachfront clubhouses cater to residents, providing world class fitness centres, retail outlets, al fresco dining, swimming pools and direct access to the islands white sand beaches. Facilities: 5 Health Clubs (1 for 4 apartments blocks), Large Swimming Pool, Modern Gym, Children's playground, and Restaurants

Table 1: Sample classified with important words highlighted. Words in red affect the price of the classified negatively (e.g. Jumeirah), while words in blue affect the price positively (e.g. Palm).

Text: regression results

Feature extraction

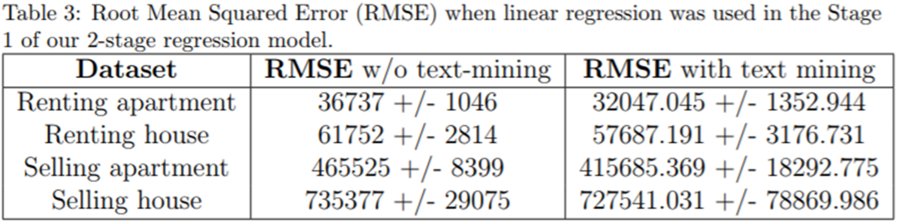
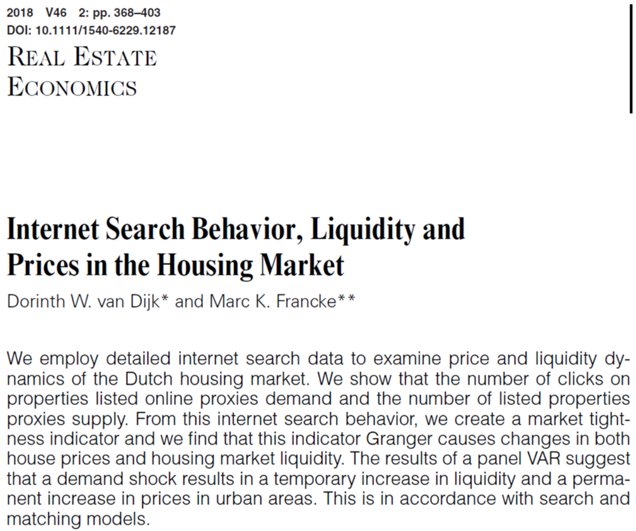


Table 3: Root Mean Squared Error (RMSE) when linear regression was used in the Stage 1 of our 2-stage regression model.

Dataset	**RMSE** w/o text-mining	**RMSE** with text mining
Renting apartment	36737 +/- 1046	32047.045 +/- 1352.944
Renting house	61752 +/- 2814	57687.191 +/- 3176.731
Selling apartment	465525 +/- 8399	415685.369 +/- 18292.775
Selling house	735377 +/- 29075	727541.031 +/- 78869.986

Funda



2018 V46 2: pp. 368–403
DOI: 10.1111/1540-6229.12187
REAL ESTATE
ECONOMICS

Internet Search Behavior, Liquidity and Prices in the Housing Market

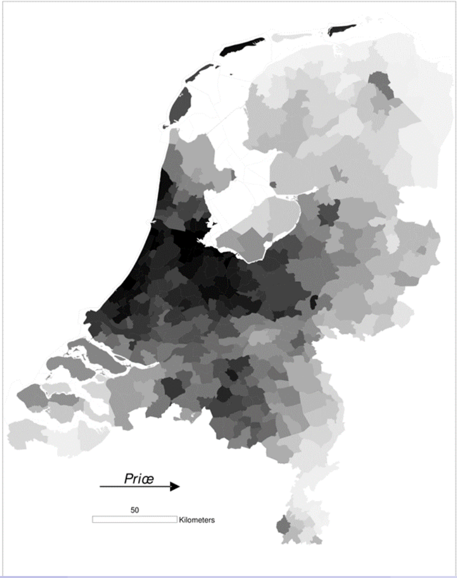
Dorinth W. van Dijk* and Marc K. Francke**

We employ detailed internet search data to examine price and liquidity dynamics of the Dutch housing market. We show that the number of clicks on properties listed online proxies demand and the number of listed properties proxies supply. From this internet search behavior, we create a market tightness indicator and we find that this indicator Granger causes changes in both house prices and housing market liquidity. The results of a panel VAR suggest that a demand shock results in a temporary increase in liquidity and a permanent increase in prices in urban areas. This is in accordance with search and matching models.

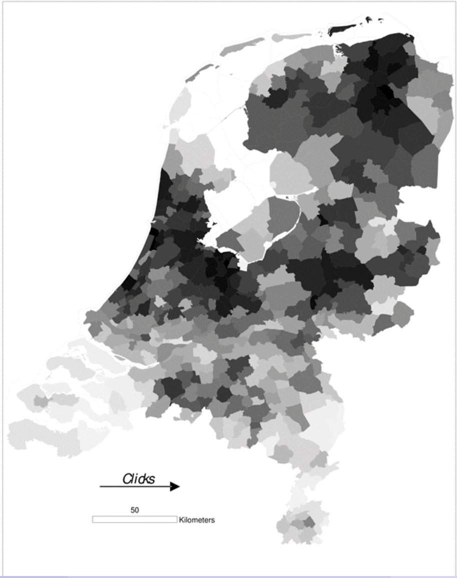
Clicks on search pages

- Forecast of price and market liquidity changes
- Liquidity: Number of transactions as a fraction of the number houses for sale
- The use of clicks enables to do a local analysis
- Market tightness: Number of clicks per period as a fraction of the number houses for sale at the start of the period
- Findings
  - An increase in the number of clicks leads to a temporary increase of liquidity and a permanent impact on price
  - Results in line with search and matching models
- Data
  - Clicks: Funda data 2011 – 2013 (60% share)
  - Transaction data: NVM (70% of transactions)

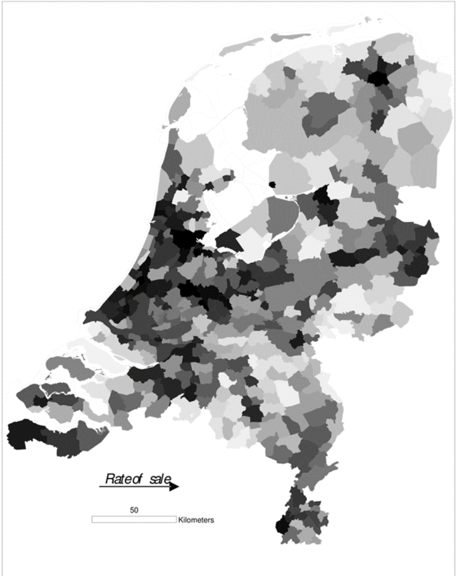
Prices



Clicks



Liquidity



Concluding remarks

- Relabeling
- OLS versus ML: applied research oversimplifies
- ML provides additional toolkit for econometricians
  - Functional form flexibility (at the cost of interpretation)
  - Creation of additional covariates (unsupervised ML)
    Satellite & language → new variables (sentiment in media)
  - Focus on out-of-sample performance and model averaging (although partly unsolved issues)
- A priori structure often needed when having limited # observations
- Do invest in knowledge on Bayesian statistics & econometrics
  - Variable selection: spike-and-slab regression
  - Large literature on model averaging (Steel, 2011)
  - Model comparison: training & test samples (O'Hagan, 2004, Ch. 7)
  - Complex model structures for large data sets: Variational methods (Spantini, Bigoni, and Marzouk, 2017) are feasible

Books

- Books
  - Bishop (2006, book):
    Pattern recognition and machine learning
  - Müller and Guido (2017, book):
    Introduction to machine learning with Python: A guide for data scientists
- Books online available
  - Hastie, Tibshirani, and Friedman (2009):
    The elements of statistical learning: Data mining, inference, and prediction
  - Barber (2017):
    Bayesian reasoning and machine learning
  - Murphy (2012)
    Machine Learning. A probabilistic approach

References I

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

Barber, David (2017). *Bayesian reasoning and machine learning*. Cambridge University Press.

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York.

Breiman, L. (2001). “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. In: *Statistical Science* 16.3, pp. 199–231.

Greene, W. H. (2008). *Econometric Analysis, 6/E*. 6th. Prentice Hall.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. Springer series in statistics. Springer-Verlag New York.

Hendry, D. F. (1980). “Econometrics – Alchemy or science?” In: *Economica* 47.188, pp. 387–406.

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.

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

Müller, A. C. and S. Guido (2017). *Introduction to machine learning with Python: A guide for data scientists*. O’Reilly Media.

Murphy, K. P. (2012). *Machine Learning. A probabilistic approach*. MIT Press.

References II

Naik, N., S. D. Kominers, R. Raskar, E. L Glaeser, and C. A. Hidalgo (2017). "Computer vision uncovers predictors of physical urban change". In: *Proceedings of the National Academy of Sciences* 114.29, pp. 7571–7576.

O'Hagan, A. (2004). *Kendall's Advanced Theory of Statistics*. Second Edition. Vol. 2B, Bayesian Inferencev. London: Arnold.

Spantini, A., D. Bigoni, and Y. Marzouk (2017). "Inference via low-dimensional couplings". In: *arXiv preprint arXiv:1703.06131*.

Steel, M. J. (2011). "Bayesian model averaging and forecasting". In: *Bulletin of EU and US Inflation and Macroeconomic Analysis* 200, pp. 30–41.

Stock, J .H. and M. H. Watson (2012). *Introduction to Econometrics*. 3rd. Pearson.

Theil, H. (1971). *Principles of Econometrics*. John Wiley & Sons.

Varian, H. R (2014). "Big data: New tricks for econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

Verbeek, M. (2008). *A Guide to Modern Econometrics*. 3th Edition. John Wiley & Sons, Ltd.