Automatic Mass Valuation for Non-Homogeneous Housing Markets

A. K. Alexandridis, University of Kent
D. Karlis, Athens University of Economics and Business
D. Papastamos, CERVED Property Services S.A.

karlis@aueb.gr

ISF, Thessaloniki, 2019

Acknowledgment: This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH . CREATE . INNOVATE (project code:T1EDK- 11293501- Real Estate Analytics REA)

Outline

- Motivation
- Introduction
- Methodology
- Data description
- Results
- Conclusions and Future work

Motivation

- In recent years big financial institutions are interested in creating and maintaining property valuation models
- The main objective is
  - to use reliable historical data in order to be able to forecast the price of a new property in a comprehensible manner
  - and to provide some indication for the uncertainty around this forecast
- In this work we develop a mass automatic valuation model for property valuation using a large database of historical prices from Greece

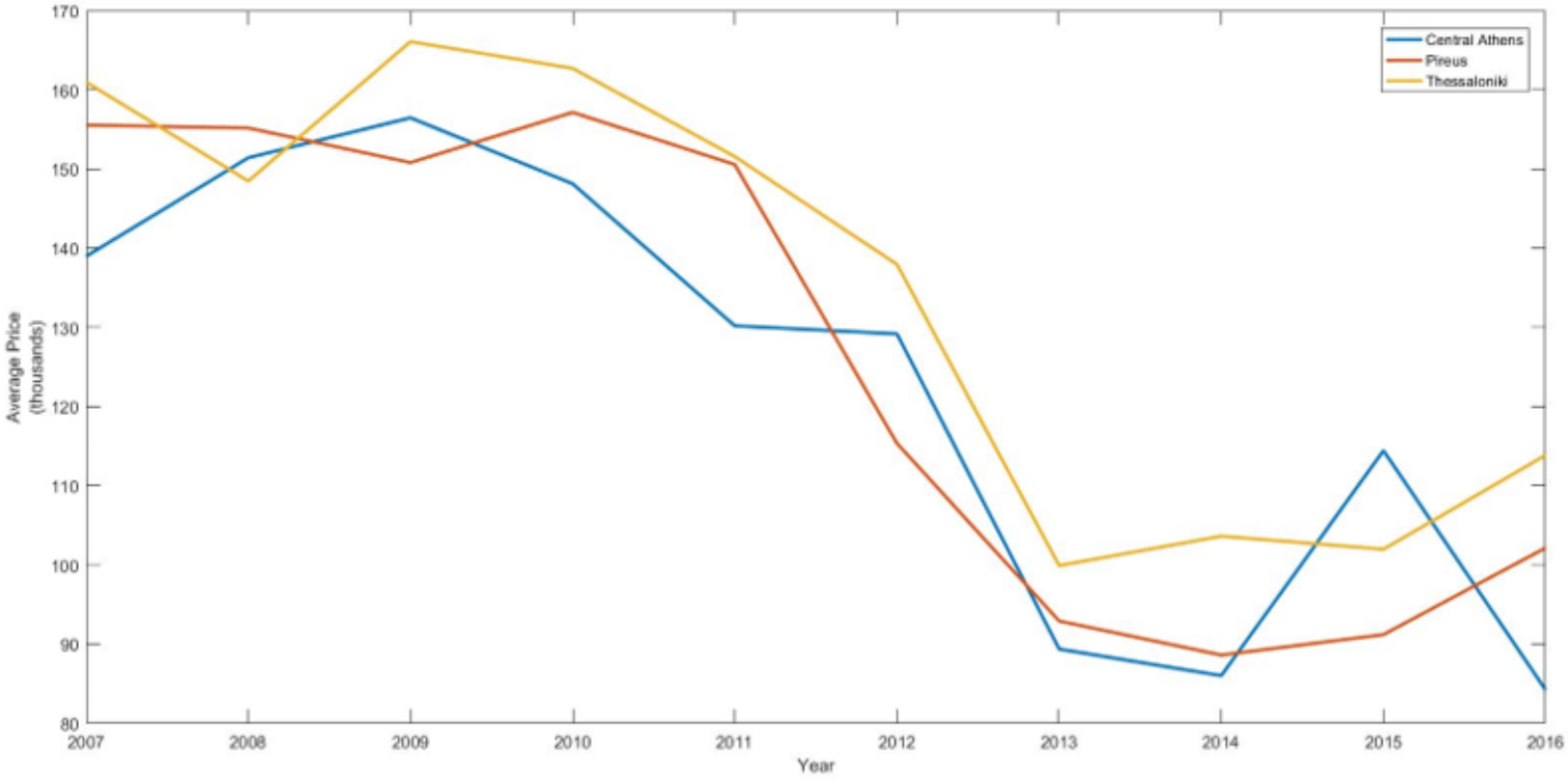
Motivation

- Past studies
  - Focus on large and developed markets
  - Small datasets
- The Greek property market
  - inefficient
  - non-homogeneous market
  - still at its infancy and
  - governed by lack of information
- As a result modelling the Greek real estate market is a very interesting and challenging problem

Introduction

- The global crisis led to a significant decline in house prices
- Financial institutions were the ones most affected with major financial losses
- At the moment the Greek market is experiencing an unprecedented situation regarding the current valuations and the future trends
- The residential market in Greece has experienced significant contraction over the last 8 years

House price decline in metropolitan areas



Decline of the Greek housing market

- Since the start of the financial crisis
  - the private construction activity in Greece is reduced by almost 80%
  - the house prices showed a cumulative decrease of 41%
  - 43.5% and 45.1% in metropolitan areas such as Athens and Thessaloniki respectively
- At the period 2008q1-2015q4
  - the ratio of non-performing loans to total bank loans increased by 30.9% (and by 38.4% if restructured loans are also taken into consideration)
  - 35.6% (and 43.5%, respectively) at the end of that period

Necessity for Automatic Valuation Models

The need for unbiased, objective, systematic assessment of real estate property has always been important

- Banks need assurance that they have appraised a property on a fair value before issuing a loan
- The government needs to know the market value of a property for taxation reasons
- Customers need to be able to have a fair value
- Banks and Real Estate companies want to decrease the cost and improve time efficiency.

Approaches in real estate valuation

- Traditional valuation methods include various expressions of linear regression multiple, stepwise, quantile, robust and additive regression approaches using hedonic models
- Recently more advanced methodologies have been employed : neural networks, fuzzy logic, multi-criteria decision analysis and spatial analysis
- Mixed results: Advanced techniques not always outperform simple linear regression
- GDPR issues apply

Methodology

We have developed and tested several methods. We focus today on some of them to keep the ideas simple. Three main groups of methodologies

- Multiple Regression Analysis
- Similarity Measure Valuation
- Neural Networks

Ensemble methods that combine outputs from the previous method are also available

Multiple Linear Regression

Typical hedonic regression model

- The model takes the form

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_j^i + \epsilon_i$$

where

- X_j^i is the value of the j-th explanatory variable/characteristic for the i-th property,
- Y_i is the logarithm of the value of the property translated to value of the present period and
- β_j, $j = 0, \ldots, k$ are regression coefficients associated with the explanatory variables, β_0 being the intercept.
- The usual assumptions for the errors apply

Forward variable selection based on predictive ability of the model derived by using Predicted Residual Sum of Squares

Predictive Approach

Our approach was the following

1. Start from a model with only the constant.
2. Select as the variable to enter the model the one that minimizes the mean of the relative absolute prediction error for the model k, defined as

$$MAPE_k = \sum_{i=1}^{n_t} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

   where y_i is the observed value of the property from the validation sample and $\hat{y}_i$ is the predicted value of the model. In the above n_t is the cardinality of a validation.
3. With the selected variable in the model we go back to step 2 to find among the other candidates the one that minimizes the MAPE
4. Stop when no further decrease of MAPE is possible.

The final model is used to forecast the values of the properties out-of-sample.

Penalized Methods

Use a penalty in the LS in order to have variable selection (screening)

- Lasso penalty
  The lasso regression minimises

$$\hat{\beta} = \min_{\beta_0, \beta} \left(\frac{1}{2N} \sum_{i}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad (1)$$

- Elastic net penalty

$$\hat{\beta} = \min_{\beta_0, \beta} \left(\frac{1}{2N} \sum_{i}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \left[(1-\alpha) \frac{\beta_j^2}{2} + \alpha |\beta_j| \right] \right) \quad (2)$$

  where the α parameters ranges in $[0, 1]$ and interpolates between the lasso ($\alpha = 1$) and ridge ($\alpha = 0$) regression.

Note the Predictive nature of the approaches

Similarity Measure Valuation

- The SMV method is based on spatial information and a representative asset (RA).
- The RA is the "average" property derived from the database.
- The value of each property is converted to a Hedonic Value (HV) based on the characteristics of the property and the Index area.
- The role of the HV is to convert all properties into a representative property in terms of characteristics.

SVM approach

- So, each comparable in our database has each own HV based on their characteristics compared to the RA and is given by:

$$HV_i = X_S^{RA} \exp\left[\log\left(\frac{UV_i}{X_S^i}\right) + \sum_{j=1}^{J} \beta_{kj}\left(X_j^{RA} - X_j^i\right)\right] \quad (3)$$

  - where β_{kj} is the hedonic coefficient of variable j for the index area k where property i is located,
  - X_j^i is the value of variable j for the property i,
  - X_S^i is the size of the property in square meters and
  - X_j^{RA} is the value of variable j for the RA.
  - UV_i, is the updated value of property i.

SVM approach

- The UV_i is given by:

$$UV_i = V_i \frac{ind_1}{ind_3} \left(\frac{ind_1}{ind_2} \right)^{\frac{m_1 - m_2}{3}}$$

  where ind_1 is the residential index at the current quarter, ind_2 is the residential index of the previous quarter, ind_3 is the residential index of the initial quarter and m_1 and m_2 are the month of the quarter of valuation and the month of the quarter of the initial valuation respectively.
- All available properties in the database are ranked based on their similarity with the property under consideration. A metric, W_{ij}, is defined to quantify the similarity.

SVM approach

- The similarity is defined as

$$W_{ij} = \exp\left[w_1 \log \frac{c_1}{d_{ij} + c_1} + w_2 I_{ij}(X_7) + w_3 I_{ij}(X_8)\right] \qquad (4)$$

  The above formula assesses the similarity of property i to another property j from the database by considering the geographical distance between properties i and j, the administrative sector and the type of the property where:
  - d_{ij} is the geographical distance between properties i and j,
  - X_7, X_8 are some characteristics of the properties,
  - $I_{ij}(x)$ is a $0-1$ indicator which equals 1 if properties i and j are identical in terms of their characteristic x,
  - w_i are weighting coefficients, which sum up to 1;
  - c_1 is a scaling parameter for the distance.
- The weights and the scaling parameters are adjusted differently for each administrative index area and they have been defined on the basis of inputs obtained from experts.
- The higher the similarity metric W_{ij} is, the stronger the similarity between properties i and j.

Neural Networks

- Previous studies show that neural networks do not perform adequately: overfitting
- We propose a three-layer NN
- Train - Levenberg-Marquardt algorithm
- Special care for parameter tuning for neural network generalisation improvement
  - Model Identification - Alexandridis and Zapranis (2013, 2014)
  - Variable Selection (select only the statistical significant variables)
  - Model Selection (correct number of hidden units/neurons)

Further steps to avoid overfitting: Validation sample

- The in-sample data were split into two samples
  - training sample - 85computing the gradient and updating the network weights and biases
  - validation set - 15measures the generalisation ability of the network
-
- The in-sample data were split randomly
- Training stops when the validation error starts to increase

Further steps to avoid overfitting: Bayesian Regularization

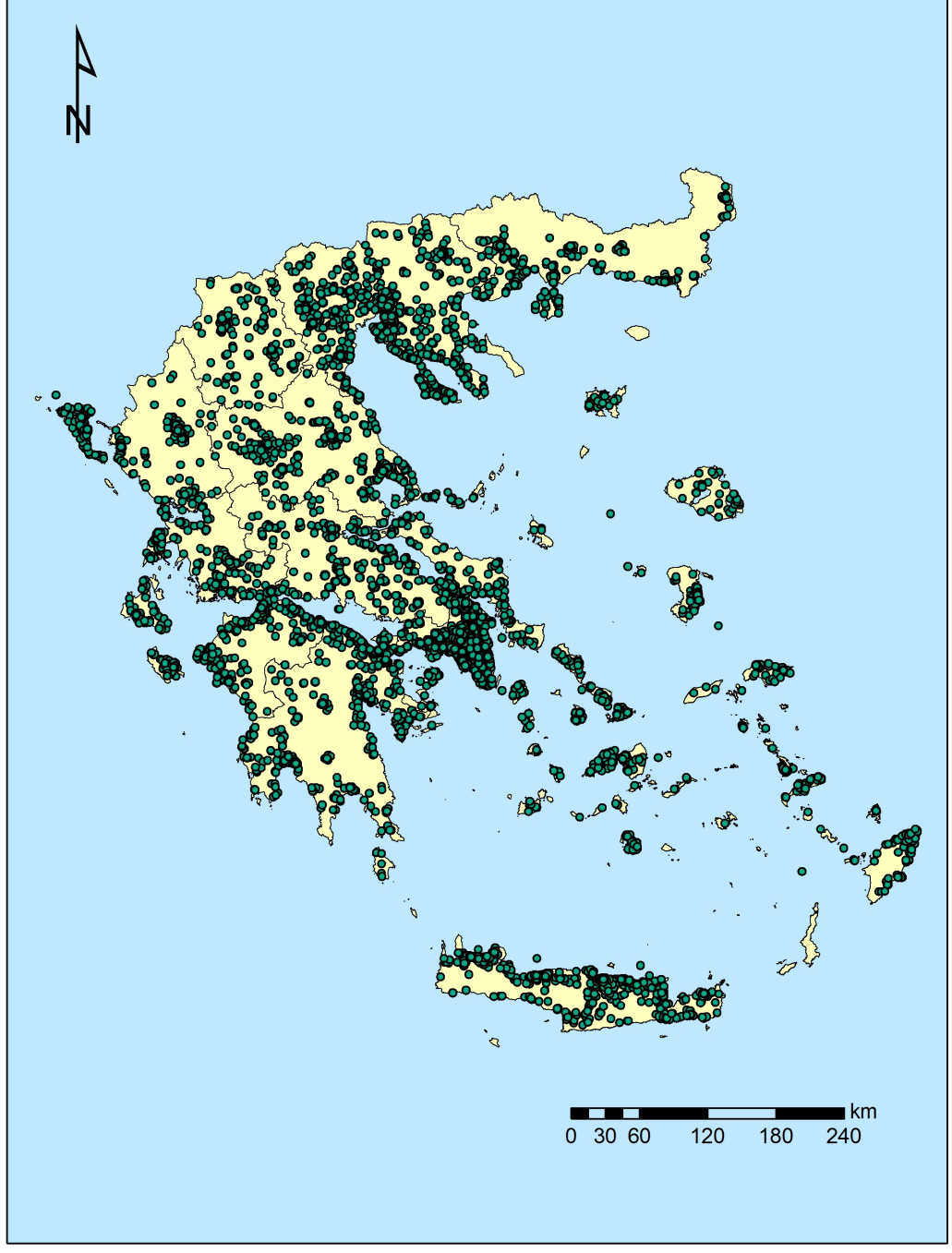
- The weights of the network are trained in order to minimize the loss function plus a penalty term
- Regularization is attempting to keep the overall growth of weights to a minimum
  - Allow only the important weights to grow
  - The rest of the weights are pulled towards zero

Data Description

- The data from a real estate agency
- Hedonic characteristics of real estate properties
- The sample consists of 36,527 properties that have been professionally evaluated in the period 2012 - 2016
- 240 different administrative sectors covering all areas in Greece
- 32 aggregated administrative areas
- The in-sample consists of 32,477 properties while the out-of-sample contains 4,050 properties

Data Description

- The majority of the properties are located in the capital or in large cities
- Around 84.5% of the properties are flats
  - 6.5% are houses
  - 5.4% maisonettes
  - 3.6% of type duplex
- Some pre-processing applied to clean the data in some extend and succeed in a certain degree of compatibility in the dataset


N
km
0 30 60 120 180 240

Initial set of variables

Code	Characteristic	Code	Characteristic
V01	ID	V12	Floor
V02	Year of valuation	V13	Total number of floors
V03	Month of valuation	V14	Existence of parking space (Y/N)
V04	Administrative sector	V15	Type of parking (In/Out)
V05	Urban classification	V16	Type of heating (categorical)
V06	Survey value (in euros)	V17	Quality of construction (categ.)
V07	Type of residence (categ.)	V18	Number of bedrooms
V08	Usable residence area (m^2)	V19	Touristic hotspot (Y/N)
V09	Land area (in sq.m)	V20	Elevator (Y/N)
V10	Year of construction	V21	View (Y/N)
V11	Distance from centre (in kms)	V22	Number of bathrooms

Measuring forecasting ability

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- $P20$, measures the percentage of the cases where the MAPE is less than 20%.

$$P20 = \frac{100}{N}\sum_{i=1}^{N} 1_{|PE_i|\leq 0.2}$$

where PE is the percentage error and it is given by

$$PE = \frac{y_i - \hat{y}_i}{y_i}$$

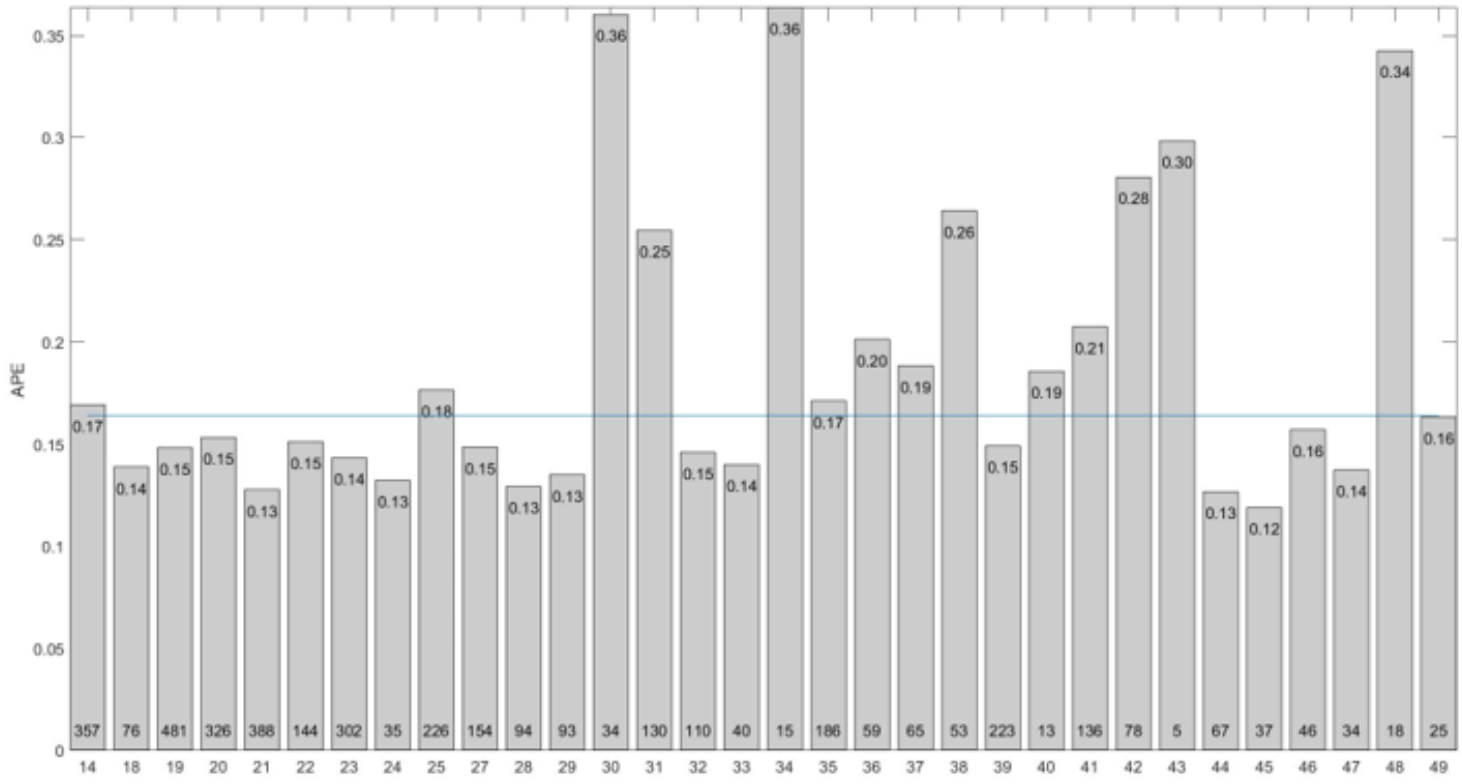
and $1_{|PE_i|\leq 0.2}$ is an indicator function where

$$1_{|PE_i|\leq 0.2} = \begin{cases} 1 & if \quad |PE_i| \leq 0.2 \\ 0 & if \quad |PE_i| > 0.2 \end{cases}$$

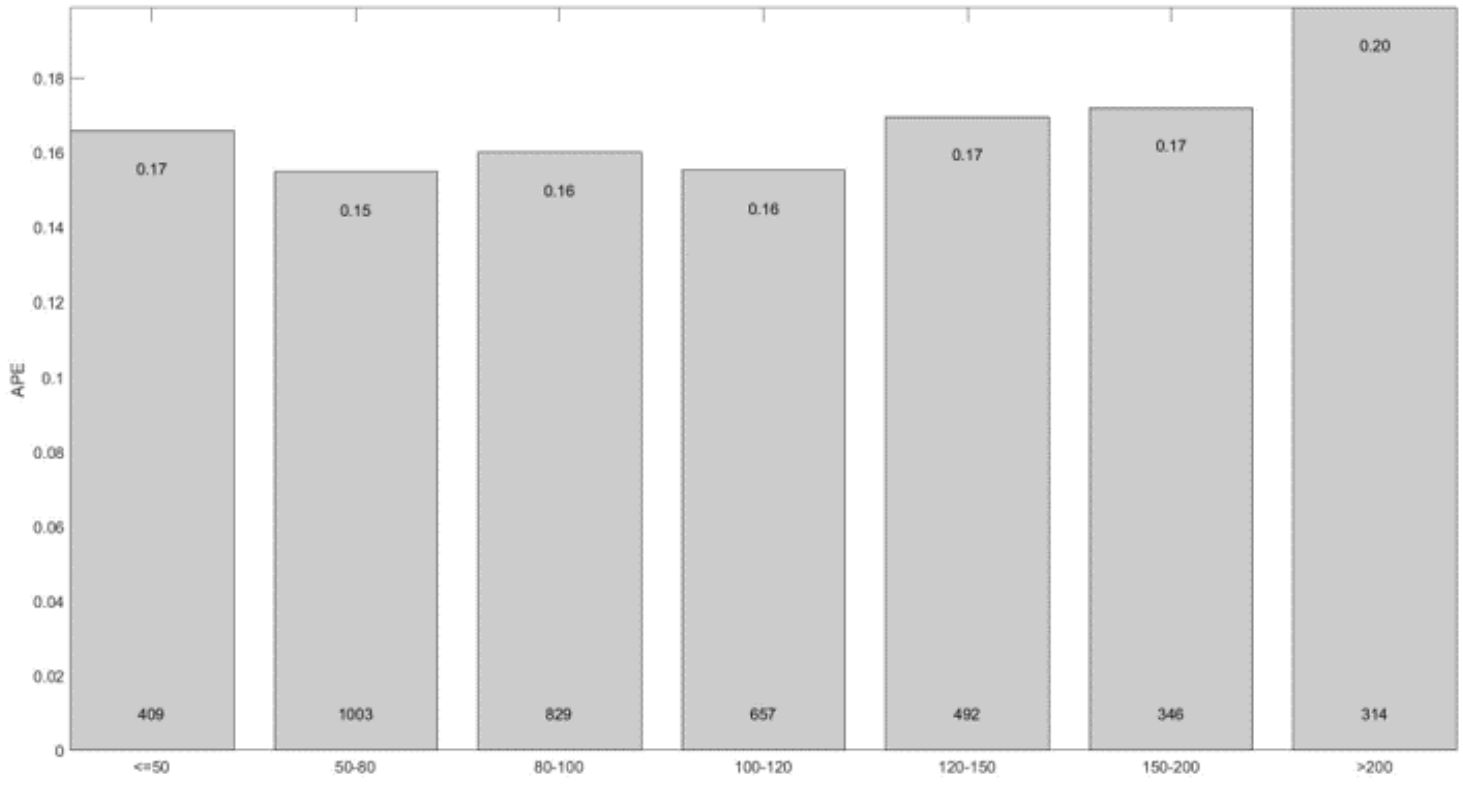
Results - Out-of-Sample Forecasting

Q1			Q2		
Method	MAD	P20	Method	MAD	P20
LM	17.85%	68.74%	LM	17.78%	68.26%
LASSO	17.57%	69.11%	LASSO	17.97%	68.07%
Elastic	17.62%	69.02%	Elastic	18.01%	68.07%
SVMOpt	16.24%	72.60%	SVMOpt	16.27%	70.97%
NN	17.10%	70.81%	NN	17.05%	67.79%
Mean	16.64%	73.45%	Mean	16.53%	70.79%
Ensemble	15.48%	72.69%	Ensemble	15.61%	72.57%
Q3			Q4		
Method	MAD	P20	Method	MAD	P20
LM	17.99%	66.92%	LM	17.46%	69.13%
LASSO	18.22%	67.16%	LASSO	17.69%	68.08%
Elastic	18.22%	67.00%	Elastic	17.67%	68.15%
SVMOpt	16.64%	70.81%	SVMOpt	16.38%	71.25%
NN	17.38%	68.57%	NN	16.88%	70.60%
Mean	16.66%	70.65%	Mean	16.05%	72.87%
Ensemble	15.62%	73.22%	Ensemble	14.17%	76.86%

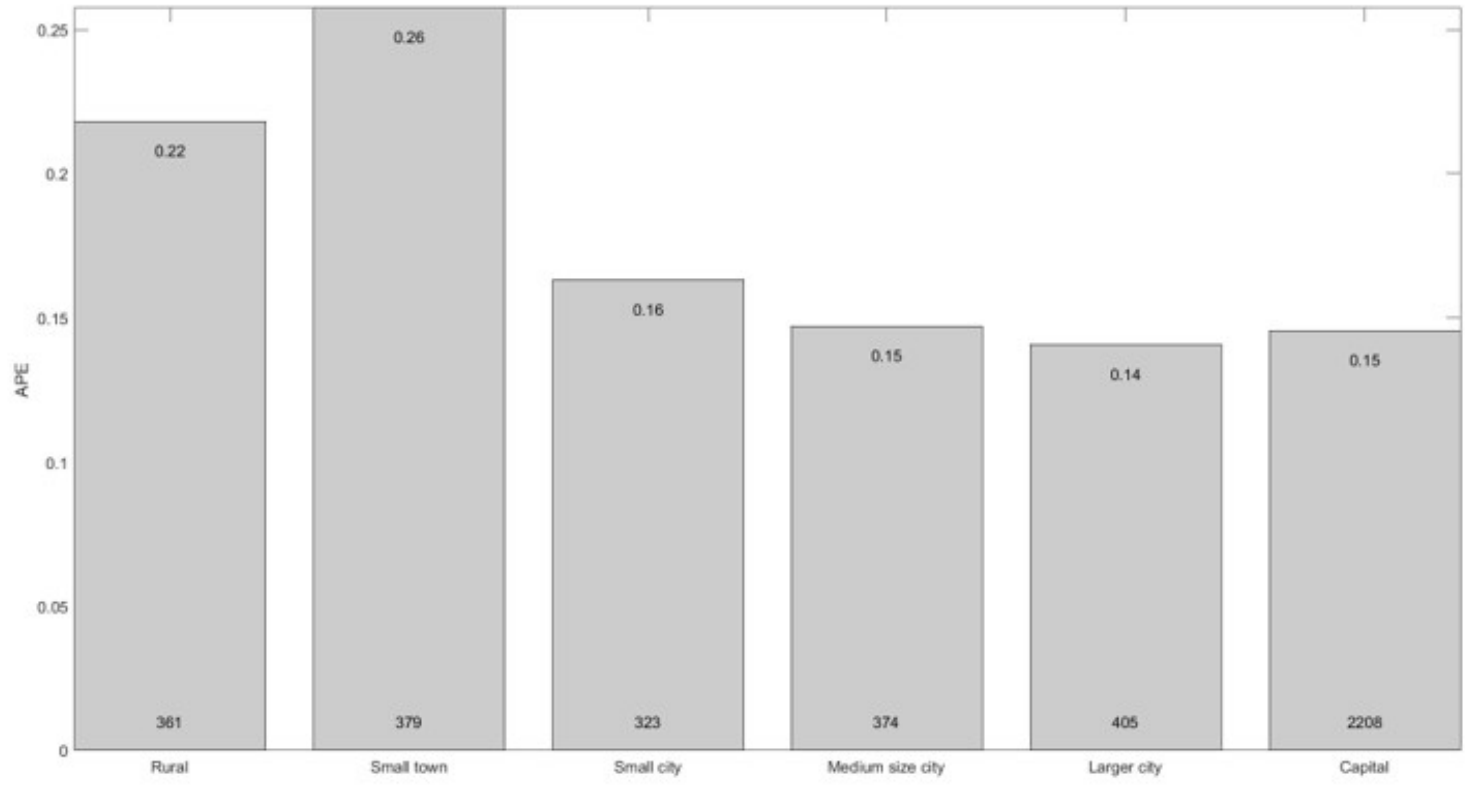
Results


APE
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.17 0.14 0.15 0.15 0.13 0.15 0.14 0.13 0.18 0.15 0.13 0.13 0.36 0.25 0.15 0.14 0.36 0.17 0.20 0.19 0.26 0.15 0.19 0.21 0.28 0.30 0.13 0.12 0.16 0.14 0.34 0.16
357 76 481 326 388 144 302 35 226 154 94 93 34 130 110 40 15 186 59 65 53 223 13 136 78 5 67 37 46 34 18 25
14 18 19 20 21 22 23 24 25 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49

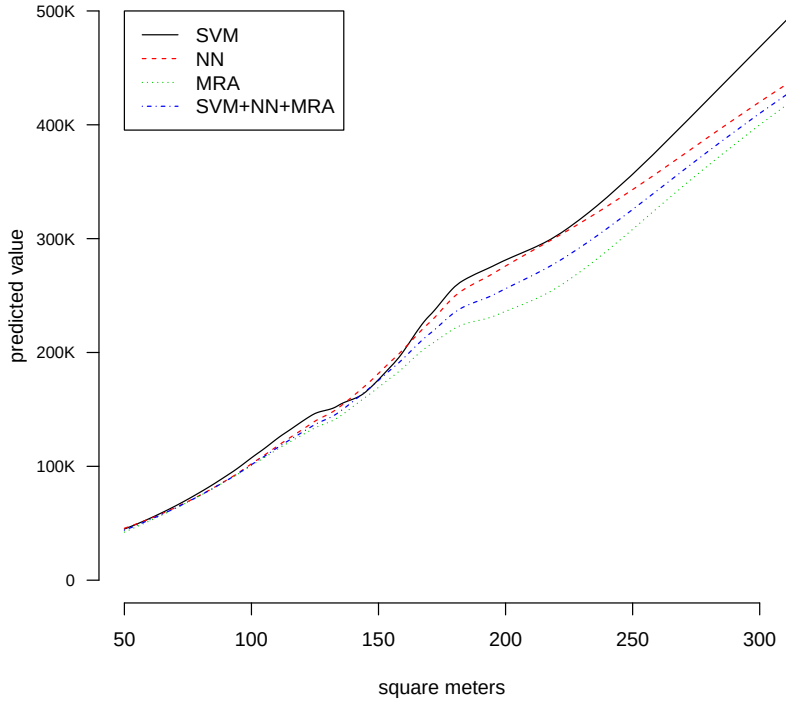
Results

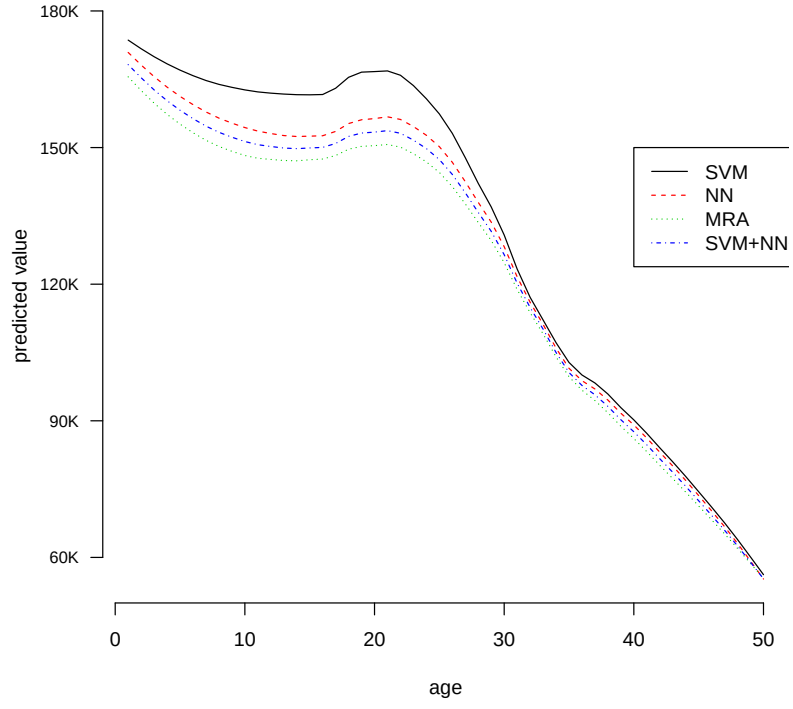

APE
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.17
0.15
0.16
0.16
0.17
0.17
0.20
409
1003
829
657
492
346
314
<=50
50-80
80-100
100-120
120-150
150-200
>200

Results


APE
0.25
0.2
0.15
0.1
0.05
0
0.22
0.26
0.16
0.15
0.14
0.15
361
379
323
374
405
2208
Rural
Small town
Small city
Medium size city
Larger city
Capital

Results


SVM
NN
MRA
SVM+NN+MRA
predicted value
500K
400K
300K
200K
100K
0
50
100
150
200
250
300
square meters


SVM
NN
MRA
SVM+NN
predicted value
180K
150K
120K
90K
60K
0
10
20
30
40
50
age

Forecasting errors: Index areas

- The MAPE is greater when only few observations are present
- The lower MAPE for all indices is obtained when the average of the three methods is considered
- The MAPE is similar across all indices for the NN, the MRA and the averaging method while is quite different for the SMV method

Forecasting errors: Urban Classification

- Higher errors for rural areas and small towns
  - while it is significantly lower for small, medium and large cities and the capital
  - the number of observations per urban classification is the same in the out-of-sample set (except the capital) this is not the case in the in-sample
- More precisely the MAPE is lower for flats while it is large for houses: (the majority of the properties are flats)

Forecasting errors: Residence Area

- For the SMV the error is minimised for properties between $50m^2$ and $80m^2$ while it is significantly larger for any other category.
- For the NN and the MRA the results are similar.
- The MAPE is lower for properties up to $120m^2$ and then it increases as the area increases
- Finally, the MAPE for the NN is smaller for every category

Forecasting errors: Land Area

- When the land area is considered all methods produce significantly higher errors.
- SMV produces significantly higher errors: when land area is included the MAPE for the SMV is 0.40 while it is only 0.29 for the remaining methods.
- When the properties do not have any land the MAPE falls 0.18 and 0.17 for the SMV and the MRA respectively while it is only 0.15 for the NN and the averaging method.
- Again the lower errors for each category are obtained by the NN and the averaging method

Forecasting errors: Age

- The MAPE is higher for properties constructed before 1970
- Also the variation for the SMV is higher compared to the other methods
- Relative stable for the remaining methods
- Again, the lower MAPE per category is obtained by the NNs

Notes

The current AVS has certain other modules, including

- Indexation per area and in total
- VaR calculations
- Visualization tools
- Usage of data from other sources
- Other methodlogies including ensemble methods
- Limited information models
- Portfolio evaluations
- Subjective variables
- Spatial characteristics

Interest lies on other aspects of real estate also

Conclusions

- We developed different mass appraisal systems for the automatic valuation of real estate properties in Greece
- We perform an extensive out-of-sample analysis in four non-overlapping data sets
- In contrast to previous studies, our results indicate that NNs constantly outperform traditional valuation methods
- averaging techniques further improve the forecasting accuracy

Conclusions

- Identify characteristics that lead to large forecasting errors
  - residence area above $120m^2$
  - the property is a house or large land area is included
  - Very old properties (built before 1970)
  - All the above indicate how the AVM may improve (if we must)
- Our results indicate that the proposed methodology constitutes an accurate tool for property valuation in non-homogeneous, newly developed markets

Future Work

The proposed Mass Appraisal System can be adapted in applications such as:

- mortgage quality control
- appraisal review
- loss mitigation analysis
- portfolio valuation
- appraisal process redesign

Acknowledgment

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH . CREATE . INNOVATE (project code:T1EDK-11293501- Real Estate Analytics REA)