Take a Look Around: Using Street View and Satellite Images to Estimate House Prices

Stephen Law
Alan Turing Institute
University College London
slaw@turing.ac.uk

Brooks Paige
Alan Turing Institute
University of Cambridge
bpaige@turing.ac.uk

Chris Russell
University of Surrey
Alan Turing Institute
crussell@turing.ac.uk

ABSTRACT

When an individual purchases a home, they simultaneously purchase its structural features, its accessibility to work, and the neighborhood amenities. Some amenities, such as air quality, are measurable whilst others, such as the prestige or the visual impression of a neighborhood, are difficult to quantify. Despite the well-known impacts intangible housing features have on house prices, limited attention has been given to systematically quantifying these difficult to measure amenities. Two issues have lead to this neglect. Not only do few quantitative methods exist that can measure the urban environment, but that the collection of such data is both costly and subjective.

We show that street image and satellite image data can capture these urban qualities and improve the estimation of house prices. We propose a pipeline that uses a deep neural network model to automatically extract visual features from images to estimate house prices in London, UK. We make use of traditional housing features such as age, size and accessibility as well as visual features from Google Street View images and Bing aerial images in estimating the house price model. We find encouraging results where learning to characterize the urban quality of a neighborhood improves house price prediction, even when generalizing to previously unseen London boroughs.

We explore the use of non-linear vs. linear methods to fuse these cues with conventional models of house pricing, and show how the *interpretability* of linear models allows us to directly extract the visual desirability of neighborhoods as proxy variables that are both of interest in their own right, and could be used as inputs to other econometric methods. This is particularly valuable as once the network has been trained with the training data, it can be applied elsewhere, allowing us to generate vivid dense maps of the desirability of London streets.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**;

KEYWORDS

real estate, deep learning, convolutional neural network, hedonic price models,machine vision, London

UrbComp'18, Aug.20, 2018, London, UK

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

House pricing remains as much art as science. The cost of a property depends not just upon its tangible assets such as the size of the property and its number of bedrooms, but also on its intangible assets such as how safe or busy a neighborhood feels, or how a house stands with relation to its neighbors. Real estate assessors face the challenging task of quantifying these effects and assigning to a property a realistic price that reflects what people are prepared to pay for these tangible and intangible assets.

From an economic perspective, it is unsurprising that people are prepared to pay for intangible assets. The urban environment directly effects people's social, economic and health outcomes. The design of a window placement can influence the amount of nature visible from within a home and also the perceived safety of a street [16]. The amount of greenery can influence both the pollutants at the street level and also its scenicness and ambiance [31]. These differences in the urban environment are reflected in the varying prices people are prepared to pay in a property market, holding other factors such as size and access to jobs constant [6, 18].

Some urban features are directly observable from photos, such as the activeness of a street frontage, the amount of greenery or the width of the pavement. Others are less directly quantifiable such as the prestige of the neighborhood or the visual aesthetics of the street. Despite the strong link between urban design attributes and economic value, there is a clear lack of research, computational tools and data that can be used to discover these attributes and inform urban planning policies. To date, the discussion regarding which urban design attributes lead to better cities or higher property values has largely been theoretical, supported quantitatively by only a few handful of studies. To measure these urban quality metrics requires many street level surveys and structured interviews with professionals.

Collecting the data required to evaluate urban quality at the city scale is both costly and time-consuming. One approach is to cast this as a problem of computer vision. This field has made great advances in image classification [17], object detection [14], image segmentation [5] and edge detection [20]. However, these advances have hinged upon the ready availability of *big data*, or in this context, hundreds of thousands of diverse images annotated with these expensive quality metrics.

Unfortunately, this is a chicken and egg scenario: to avoid the expensive and time consuming hand annotation of images, we must first perform the expensive and time consuming process of hand-labeling of thousands of such images. To avoid such issues, this research will not use machine vision methods to classify or to detect intermediate values, such as *amount of greenery*, that can be used in house price models but instead use deep learning

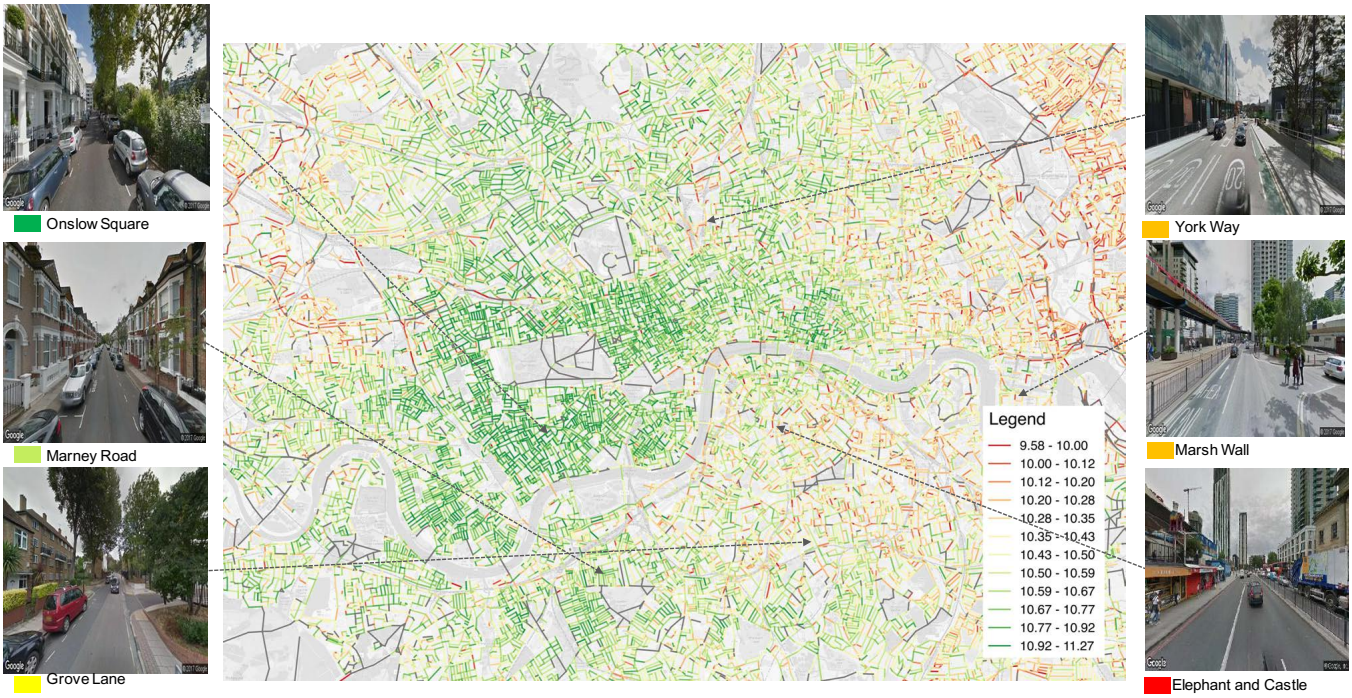

Figure 1: A map illustrating the latent visual appeal of neighborhoods across Greater London. Using a *linear hedonic model* we are able to extract the marginal effects of the visual appeal of the urban environment on house prices, as latent factors or proxy variables. The contribution of the urban environment retrieved from house prices varies from positive (green) through to negative (red). The map does not correspond to house prices, but to the visual appeal of the neighborhood which must be then combined with other housing attributes to price properties.

machine vision techniques to extract visual features in images based on the property price in an end-to-end learning model. We extract visual urban features using convolutional neural networks on urban images at both the plan and street-level which can be used in conjuncture with traditional housing features to estimate the price paid for a property in London.

A fundamental trade-off exists in econometrics between the use of tractable models [30] that are easy to analyze, and difficult to interpret black-box approaches such as [35] that often have significantly better accuracy. To handle this dichotomy, we consider two different approaches; *(i)* a full black-box model in which the a neural network implicitly integrates the cues of from standard attributes and from image data, and *(ii)* a hybrid approach in which a mapping from the image space to latent attributes is handled by a CNN and then the cues are fused by a standard linear model. This hybrid approach leads to the learning of interpretable semantic features that act as proxy variables for visual appeal of a neighborhood. Figure 1 shows a map of these features over greater London.

Our work differs markedly from previous research that has made use of images to price houses. First, we focus on using images of the urban environment at both the street and aerial level to estimate house prices rather than using interior images, and more importantly we have developed a set of interpretable proxy attributes which measure the visual desirability of neighborhoods; these variables can be used directly in existing econometric models. This concept is similar to the use of indices of multiple deprivation, crime attributes and school-performance data as proxy for neighborhood safety and prestige [13].

2 RELATED WORK

The cost of a heterogeneous good, such as housing, can be broken down into its utility-bearing components using the *hedonic price approach* [30] [6]. The principle behind the hedonic price approach is that, holding all things constant, the influence of an attribute can be discovered by observing real estate values. One can imagine this concept by comparing two properties, each with nearly identical features, except that one property has one bedroom and the other has two bedrooms. The price differential between the two is equal to the implicit price of the extra bedroom. This approach can include structural features such as the size of a house, the age of a home and the type of a home. It can also include location features such as employment accessibility or neighborhood features such as the number of shops nearby. Since its introduction, the hedonic price approach has become an established method for pricing environmental goods, constructing housing price indices, and as evidence in the development of welfare policies [25] [29] . Despite the clear improvements in accuracy, there has been limited research into the the use of machine learning methods in house price estimation as a alternative [2, 26, 35]. One reason is that the hedonic price approach can use the estimates of a OLS model to recover

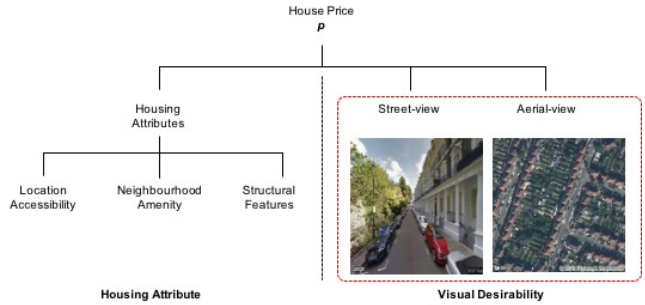

Figure 2: Conceptual model showing how urban visual features can be integrated with a traditional house pricing model.

the marginal willingness to pay for goods that are without explicit markets [30]. Despite its ease of use and interpretability, Peterson and Flanagan [26] argues such OLS model generate significant mispricing and misspecification errors. As a result, the adoption of multi-layer perceptron (MLP) in hedonic price model is logical and supported [26].

This research will also adopt machine vision methods from street images to recover the visual attributes of the urban environment. Recent studies from Naik et al. [23], Liu et al. [21] and Law et al. [19] have began to leverage on the availability of large scale street image data to extract urban knowledge. For example, both Liu et al. [21] and Law et al. [19] used machine vision techniques to retrieve geographical knowledge such as street frontage classes. In contrast, Naik et al. [23] used Street View images to estimate the perceived safety of streets [33]. This research is related to this latter study in extracting a global statistic from street images.

The following section will outline the current status of machine learning techniques in house price estimation. The first is a study from Peterson and Flanagan [26] that used a multi-layer perceptron model to estimate house price with traditional housing features such as age, size, accessibility and safety. The author compared an artificial neural network hedonic price model with two hidden layers to a standard OLS hedonic price model. The author found significant improvements in the use of an ANN model. The improvements are unsurprising due to the expected non-linear relations captured by variables in the hidden layers which can not be modeled by a linear OLS model.

These non-linear effects become more important when dealing with intangibles such as the quality of the neighborhood, as these intangibles can often have a multiplicative effect on the hedonic value assigned to tangible assets. For example, each square meter of property could cost orders of magnitude more in an exclusive neighborhood than in a run down one.

A study from Ahmed and Moustafa [2] supplemented traditional housing features with visual features extracted from property photos. The study used both property photos and traditional housing features in estimating house price. The result found objects identified using traditional machine vision methods such as Speeded Up Robust Features (SURF) [4] significantly improved the model. The research also compared a support vector regression model to a neural network model and found that the neural network one achieved better results.

Table 1: Housing attributes statistics

	mean	sd	min	max
log price	12.03	0.62	0.69	15.3
year	0.42	0.22	0.00	1.0
size	0.52	0.14	0.00	1.0
beds	0.30	0.14	0.00	1.0
age	0.62	0.14	0.00	1.0
type	0.30	0.41	0.00	1.0
park	0.76	0.15	0.00	1.0
shops	0.46	0.19	0.00	1.0
gravity	0.65	0.14	0.00	1.0

You et al. [35] also used visual features extracted from property photos to estimate a house price model. Instead of using traditional machine vision techniques in detecting image feature, this research made use of a novel recurrent neural network LSTM model to predict house price using a random walk sequence over nearby properties. This research predicted the price of a home using both the location on the random walk sequence and the visual features in an end-to-end learning model.

Gebru et al. [12] extracted car types, years and make from 50 million Google Street View images to correlate with socio-economic factors such as income and geographic demographic types across different cities in the United States. The study found that car types, years and makes can be used as features to predict accurately the income, race, education and voting patterns at both the zip code and precinct level.

Several related works do not model house prices directly, but provide further evidence that street-level photographs of a city can be used to estimate relevant features. Dubey et al. [9] collected human perception data from street images (Place Pulse 2.0) through a crowd-sourced survey [23]. They then fit a model to predict these human perception factors, such as perceived safety and liveliness, directly from the images; these factors are likely important covariates in a house price model. Arietta et al. [3] present a method for automatically identifying and validating predictive relationships between the visual appearance of locations in a city and properties such as theft rates, house price, population density, tree presence, and graffiti presence. The novelty of the study is it extracted a set of discriminative visual features [8] such as roof types and window types that corresponds to a location attribute using a support vector machine. The model successfully identified visual features that corresponds to location with higher or lower house price (binary). However the model did not generalize well across cities in the States.

We differ from these previous approaches in multiple ways. First, this study collects urban neighborhood images [12] both at the street level and aerial level rather than images of the property itself [2, 27, 35]. This allows the neighborhood features to be extracted from two perspectives, the street of the property and the neighborhood surrounding the property (Figure 2). Secondly, we compare a ANN hedonic price model [26] with only housing features to a model that is augmented with both street images collected from

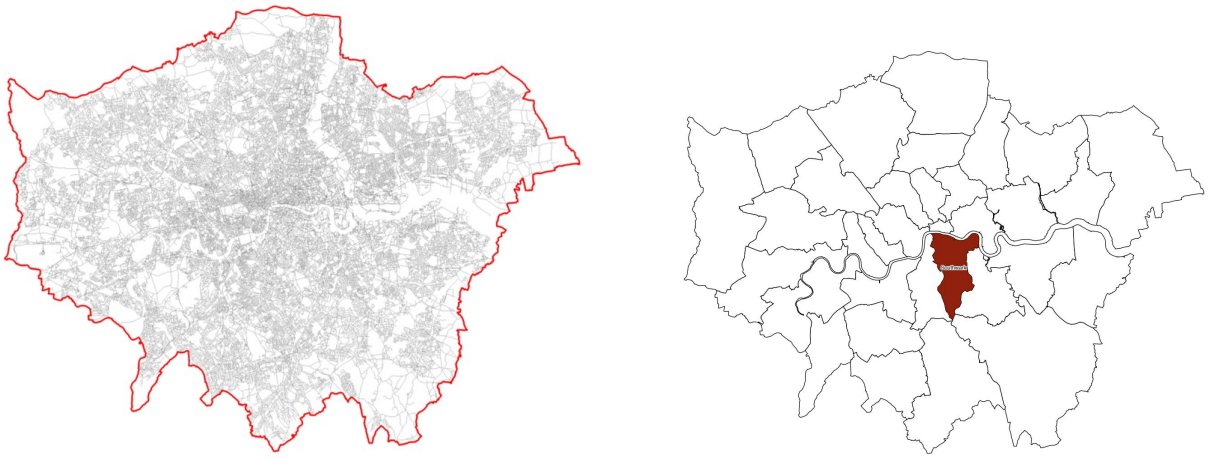

Figure 3: Left: Greater London study area. Right: The Southwark test-set used in one of the experiment. Contains Ordinance Survey data ©Crown copyright and database right ©2017.

Google Street View API [15], and aerial images collected from Bing Images API. Third, we compare a non-linear hedonic price model with a hybrid linear hedonic price model where the images gets encoded into a latent variable that achieves greater interpretability. Finally, the model is tested on multiple neural network architectures and using a spatial out-of-sample testing set, in which Southwark, an entire borough of London, was excluded from the training set, to demonstrate the generalizability of the results.

3 METHOD AND MATERIALS

We propose a model which estimates the log house price from three separate sets of input data: housing attributes, street images, and aerial images. To demonstrate the effectiveness and utility of our model, we use Greater London (Figure 3) as a case study. The proposed procedure consists of a data collection phase, a training phase and a testing phase; we begin by describing the data collection phase.

3.1 Data collection

We collected three datasets in the data collection phase. The first dataset is comprised of traditional housing attributes including structural, neighborhood and location features. House price data is taken from the UK Land Registry Price Paid dataset [28], which includes transaction details for all property sales in England, with additional property attributes from the Nationwide Housing Society [24]. The structural features, for each property transaction, include the location of the property, the price paid for the property, the type of the property, the size of the property, and the age of the property. Location features include gravitational accessibility to employment. The statistic was computed as a gravity model, where accessibility is measured as a sum of jobs divided by distance within 60 minutes $\sum e_{ij} d_{ij}^{-1}$.

Neighborhood features include the distance to the nearest parks and the number of shops and commercial uses within 800 meters. The datasets used to calculate these location features comes from

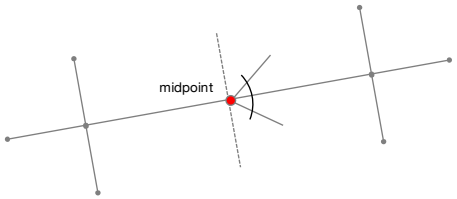

Figure 4: Street bearing diagram, and the orientation of front facing cameras.

the Ordinance Survey [34], the Office for National Statistics [11] and Historic England [10] . This dataset consists of a total of 110, 000 transactions which are then grouped by the nearest street. Descriptive statistics are shown in Figure 1. The output variable, price, is log transformed, while all the input attributes are log transformed and then linearly rescaled to have minimal values of 0 and maximal values of 1.

The second dataset is comprised of street images taken from the Google Street View API [15][1]. Following [19], one front-facing image was collected for each street in the Greater London Area using the API. (A front facing image is one which faces towards the front of the car, i.e. it typically faces away from the property at a ninety degrees angle; see figure 4 for clarification.) To collect the dataset, we first constructed a graph from the street network of London (OS Meridian line2 dataset [34]), in which every node is a junction and every edge is a street. We then took the geographic median and the azimuth of the street edge to give both the location and the bearing when collecting each image 4. This is to ensure the Street View images are constantly front-facing and are taken from the center of the road. This reduces the problem of images being too close to the junction. The field of view has been set to 120 degree in order to ensure that both sides of the building facades

[1] ©2017 Google Inc. Google and the Google logo are registered trademarks of Google Inc.

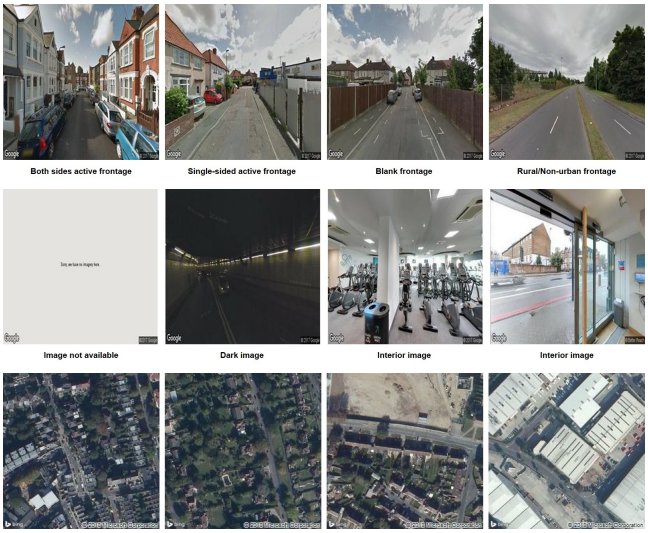

Figure 5: (Top row) Valid Google Street View images. (Middle row) Invalid images discovered using techniques in [19]. From left to right: not available image; dark image; interior image; interior image. ©2017 Google Inc. Google and the Google logo are registered trademarks of Google Inc. (Bottom row) Microsoft Bing aerial images, ©2018 Microsoft.

are captured. 110,000 images have been collected from this process of which 40,000 of them have at least one property transaction. A typical data cleaning procedures is then undertaken which includes removal of invalid images such as the interior of buildings and images that were too dark or those not available, using a series of automatic functions and manual processes [19]. Figure 5 shows examples of the valid images and invalid images. Following the cleaning process, the Street View images were then resized to a uniform resolution (256 pixels × 256 pixels).

The third dataset is comprised of aerial images extracted from the Microsoft Bing Images API [22][2]. Using the API, one aerial image has been collected for each street in the Greater London Area. To collect the dataset, we take the centroid of each street edge from the OS Meridian line 2 dataset [34]. We then download for each street an aerial image with a zoom level parameter set at 18 (roughly 150m) to get a constant aerial view of the street neighborhood. A total of 110,000 images were collected by this process of which 40,000 of them have at least one property transaction. Similarly, aerial images were then resized into the same dimension as the ground level images (256 pixels x 256 pixels). Figure 5 shows examples of these aerial images.

3.2 Model Architecture

Our architecture can be understood as a natural generalization of the hedonic perceptron model used by works such as [35]. We train a multi-layer neural network to predict log house prices on the basis of a set of normalized attributes (see Table 1). We depart from the standard model of You et al. [35] in that we also allow the input

[2]©Bing. All rights reserved.

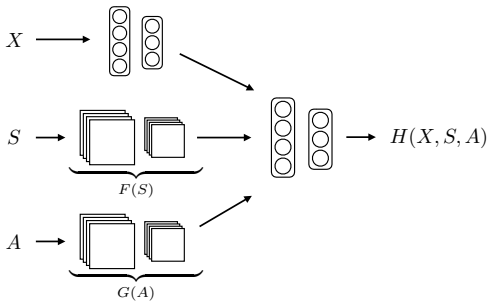

Figure 6: Full model network structure.

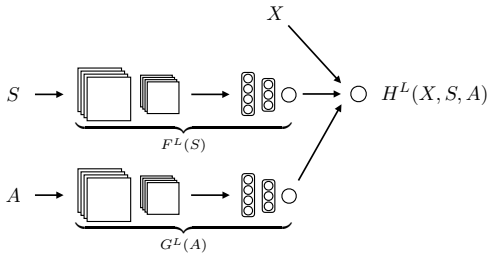

Figure 7: Linear model network structure.

of two latent attributes that can be understood as proxies for the desirability of the urban environment as captured by Street View data, and by satellite imaging.

These proxies are given by the responses of other convolutional neural networks. Importantly, as we do not have expert annotations of the desirability of the urban environment, we learn feature extractors for the Street View and satellite imagery by composing these networks with a hedonic price model $H(\cdot)$ and training the entire architecture end-to-end, while controlling for the contribution of the individual housing attributes.

There are two important uses of such hedonic models. The first lies in accurately predicting house prices as a guide for realtors and for people looking to put their house on the market; for such individuals, accurate pricing is the most important criteria, and they are happy with the use of black-box models such as neural networks providing they lead to improved accuracy. The second use of these models lies in econometrics; here interpretability and ease of analysis are more important than accuracy and the use of linear model is still favored.

Because of this, we consider two forms of the hedonic price model. The first form is designed to maximize the predictive accuracy of $H(\cdot)$ is a multi-layer perceptron (see Figure 6), capable of learning arbitrary functions, while in the second form, $H(\cdot)$ is a linear model that learns only a linearly weighted combination of features (see Figure 7). In both cases, multi-layer convolutional networks capable of learning non-linear responses are used to process the Street View and aerial images.

3.2.1 The Hedonic Price Model. We represent the overall price of the property by a function $H(W_1, \cdot)$, parameterized by a set of weights W_1 that takes as input housing attributes X and extracted image features $F(S)$ from Street View images S and $G(A)$ from aerial

photos A. For purposes of establishing baselines and quantifying the relative predictive capability of the housing attributes and the new image data, we consider a baseline model $H(W_1, X)$ which depends only on the housing attributes X, as well as models $H(W_1, \dots)$ which can additionally incorporate either or both of the Street View and aerial photos; the full combination of experimental setups is described in Section 4.

For the non-linear hedonic perceptron model, a fully connected neural network with two hidden layers is adopted. The first fully connected layer (FCL) has 60 hidden nodes, while the second FCL has 30 hidden nodes. This layer represents an extracted feature vector with a nonlinear dependence on X. In the baseline model $H(W_1, X)$, a final FCL outputs the overall response of the model; for the models which include the images S and/or A, we concatenate this vector to vector-valued output of the functions $F(\cdot), G(\cdot)$ and use this as input into an additional fully-connected network, again with two hidden layers of 60 and 30 nodes respectively. This is the model whose architecture is shown in Figure 6. These taken together yield an overall combined non-linear predictive model of the form

$$H(X, S, A) = H(W_1, X, F(W_2, S), G(W_3, A)). \tag{1}$$

The linear hedonic model can be interpreted as a network with no hidden layers, that consists of a single neuron with no non-linearity which directly outputs the response. A primary difference between the linear and non-linear models is in the handling of the images themselves. In the non-linear model, the trained sub-networks $F(\cdot)$ and $G(\cdot)$ extract a feature *vector* when used as inputs to the nonlinear model. In contrast, the sub-networks $F^L(\cdot)$ and $G^L(\cdot)$ in the linear model output scalar summaries which can be included as additional independent variables in an OLS model, where they function as proxy variables to control for visual desirability of the local urban environment.

One benefits of the interpretable econometric approach is that the feature response $\gamma_F F^L(\cdot) + \gamma_G G^L(\cdot)$, where γ_F, and γ_G are the weights learned by the linear model, can be directly interpreted as a measure of how the visual desirability of the neighborhood alters the value of the house prices. Figure 1 shows a heat plot of these responses over the whole of London.

3.2.2 Urban Environment (Street View and Satellite). To extract meaningful features from the Street View and satellite image data, we define two functions $F(W_2, S)$ and $G(W_3, A)$, with weights W_2 and W_3. Although they have different weights, both networks adopt the same convolutional neural network (CNN) architecture for the vision model. In a CNN model, the earlier layers detect the basic edges while the ladder layers detect the more complex shapes. The model follows the basic CNN architecture that uses 3x3 filters that are tested on 4, 8, and 12 convolutional layers (as in e.g. VGG[32]). We take the value at the final flattened convolutional layer as the output of the CNN. These outputs are feature vectors which summarize the Street View and aerial photo data, respectively, which can then be used as inputs into the nonlinear hedonic model.

For the linear hedonic model, we define two networks $F^L(S)$ and $G^L(A)$, which differ from the networks $F(S)$ and $G(A)$ in that their output is scalar, rather than vector. This network is defined by including two additional fully-connected layers which reduce the feature vector output by the CNN to a single scalar output $F^L(S)$. These scalar outputs can be used as proxy variables, alongside the housing attributes X, in a standard OLS model.

The linear model is described in more detail in Section 4.3 of the experiments. First, we will evaluate the predictive performance of the fully non-linear hedonic perceptron model.

3.3 Model Evaluation

The difference between the predicted log price $\hat{Y} = H(X, S, A)$ given by Equation (1) and the actual log price Y is given by the mean squared error loss function

$$L(W_1, W_2, W_3) = \frac{1}{n} \sum \big(Y - H(X, S, A)\big)^2. \tag{2}$$

This loss is a function of the weights W_1, W_2, W_3 which are optimized in the learning process.

4 EXPERIMENTAL RESULTS

We consider three sets of experiments: The first two using general neural networks to regress, and the third using a standard OLS linear regressor with neural networks as mid-level components. For all three experiments, we train the model end-to-end to minimize the mean squared error on a training set, using the ADAM optimizer with the default initial learning rate set at 0.001. We report two test set metrics: the mean squared error (MSE) and the coefficient of determination R^2 between the model prediction and the actual log-price. All the experiments are conducted with the Keras library [7] using a Tensorflow [1] back-end.

To test the importance of particular attributes with respect to the model accuracy, we constructed six different models. The first three models are individual models for each data source. The final three models are different combinations of multiple data sources.

4.1 Spatially Missing-at-Random

In the first experiment, we tested three variations of these six models by altering the architecture of the Street View network $F(S)$ and the aerial imaging network $G(A)$. We split the dataset randomly where 70% is used for training, 15% is used for validation, and 15% is used for testing, yielding an experimental setting in which the test set is spatially missing-at-random relative to the training set. We tested a 4-layer CNN, a 8-layer CNN and a 12-layer CNN model. Note that varying the architecture does not alter the attribute-only model, which has no convolutional layers.

Figure 8 shows the scatter-plots between the actual and the predicted log price for all six models, using the best-performing architecture. The result shows quite clearly that the four models which include the housing attributes X as one of the inputs achieve much higher correlation than the two models which use only Street View or aerial image data. This is to be expected, as these models only have visual information for the prediction model.

The result in Table 2 shows the mean squared error and R^2 for all six models, and across all three sizes of architectures. Of the single data source models, the housing attribute model achieve better accuracy than both the Streetview model and the aerial-image model. Models using multiple data sources achieves better accuracy than the single data source models. The model with both X and S achieves 76% accuracy, while the model with both X and A achieves 81% accuracy and the full model, including all of X,

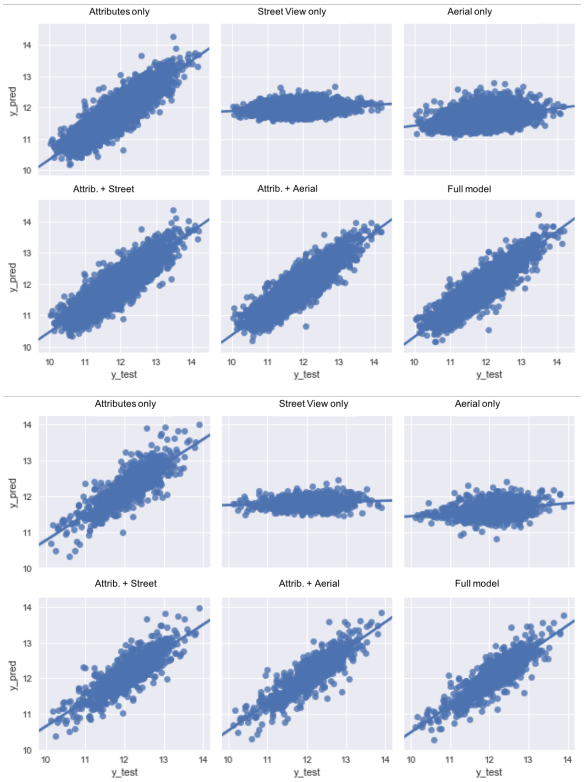

Figure 8: Top: Scatter-plots showing correlations for each model on the spatially missing-at-random experiment. Bottom: Scatter-plots showing the correlation for each model on the holding-out Southwark experiment.

S, and A achieves 82% accuracy. The results show that the model that combines housing attributes with the aerial images achieves a better result than the model without the aerial images.

4.2 Generalization: Holding out Southwark

In the second experiment, we split the dataset so the entire borough of Southwark in Figure 3 becomes a spatially out-of-sample test set. By splitting the dataset over the entire borough we show that the image network is not simply memorizing locations and recognizing neighboring streets as having similar house prices. This is a very difficult challenge, which tests the ability of the learned network to generalize to new locations which may have different visual cues indicating the desirability of neighborhoods.

Of particular note is the fact that the introduction of visual features do not just substantially improve the accuracy of the regressor, but also the stability when generalizing to unseen regions of London. Although all models exhibit a significant drop off when forced to generalize to a missing London borough rather than simply to data missing at random, this loss in accuracy is cut by two thirds — only dropping by around 5% rather than 15% — when using regressors that make use of attributes and visual features. This is remarkably successful given the challenge of the task and the high visual diversity of boroughs of London.

Table 2: Spatially missing-at-random results. MSE(top) and R^2 accuracy(bottom)

MSE	4-layers	8-layers	12-layers
Attributes only	0.10	–	–
Street View only	0.33	0.32	0.35
Aerial only	0.34	0.29	0.30
Attrib. + Street	0.08	0.12	0.09
Attrib. + Aerial	0.07	0.07	0.07
Full model	0.06	0.07	0.06
R^2	4-layers	8-layers	12-layers
Attributes only	74.85	–	–
Street View only	4.92	5.16	5.93
Aerial only	15.22	15.75	13.91
Attrib. + Street	76.08	76.46	75.59
Attrib. + Aerial	80.90	80.61	78.28
Full model	81.63	81.30	81.16

Table 3: Generalization to held-out Southwark. MSE (top) and R^2 accuracy (bottom)

MSE	4-layers	8-layers	12-layers
Attributes only	0.13	–	–
Street View only	0.42	0.40	0.34
Aerial only	0.55	0.45	0.47
Attrib. + Street	0.10	0.14	0.12
Attrib. + Aerial	0.08	0.09	0.09
Full model	0.08	0.08	0.08
R^2	4-layers	8-layers	12-layers
Attributes only	68.96	–	–
Street View only	2.65	1.66	0.72
Aerial only	6.24	5.12	5.49
Attrib. + Street	70.27	70.76	68.70
Attrib. + Aerial	75.91	75.02	72.61
Full model	76.12	75.01	75.47

4.3 Linear Hedonic Pricing Comparison

In the third experiment, we compared the linear hedonic price model which is a linear combination of both housing attributes X and the image attributes $F(S)$, $G(A)$ with the traditional linear hedonic price regression model of using only housing attributes,

$$H^L(X) = \beta_0 + \sum \beta X + \epsilon \quad (3)$$

$$H^L(X, S, A) = \beta_0 + \sum \beta X + \gamma_F F^L(S) + \gamma_G G^L(A) + \epsilon \quad (4)$$

We fit the linear hedonic price regression model both with and without proxy variables for visual urban appearance. The result shows that the linear model with proxy variables offers a significant improvement over the standard model, coming much closer to the accuracy of the more general hedonic perceptron, that does not use image data, while retaining the interpretability of the linear model. The structure of this model is shown in Figure 7.

To demonstrate how interpretable our new approach is, we plot on a map the values $\gamma_F F^L(\cdot) + \gamma_G G^L(\cdot)$ from the full model, for Street View and satellite data across the whole of central London,

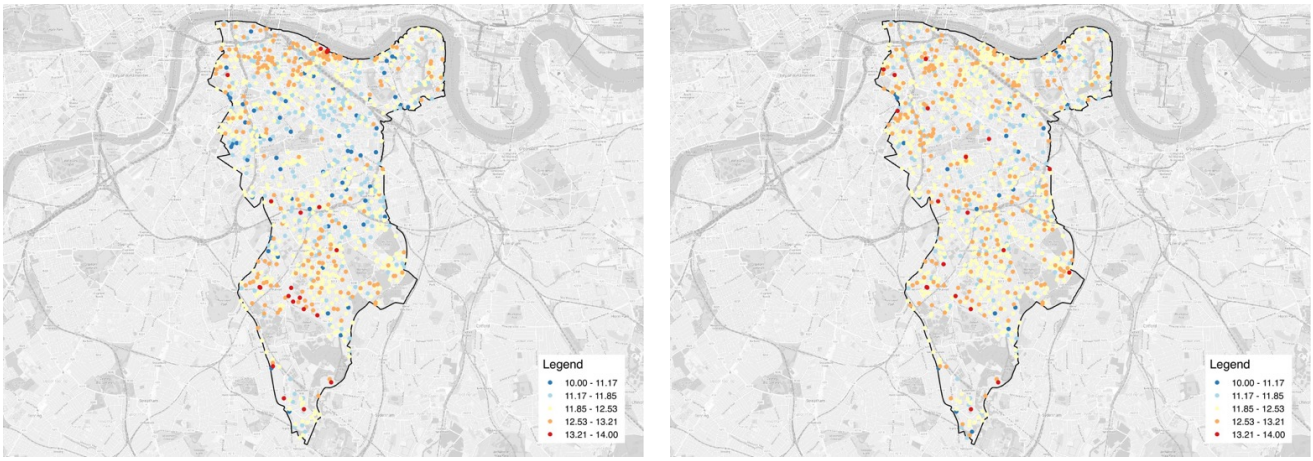

Figure 9: Test of generalization ability: predicting prices in the Borough of Southwark, using a model trained on data from elsewhere in London. *Left:* Actual log-price; *Right:* Predicted log-price. Survey data ©Crown copyright and database right ©2017.

Table 4: Linear Hedonic Model

MSE	Random-set	Southwark-set
Linear Hedonic	0.159	0.184
Linear with Images	0.103	0.148
R^2	Random-set	Southwark-set
Linear Hedonic	58.36	53.29
Linear with Images	70.29	62.31

including areas for which we have no transaction data available. This map, shown in Figure 1, contains the predicted contribution to the hedonic utility of properties based on their visual appearance.

5 DISCUSSION

This study finds encouraging results in predicting house prices in London using street images both at ground level and at aerial level. We find that the traditional housing attributes explains the majority of the variance of house price; we also find that the model augmented with features extracted from images performs better than the model without image features. This research also demonstrates that augmenting the baseline housing attribute model with aerial images perform better than the baseline model with ground-level Street View photos. This result suggests that buyers might be valuing a visually desirable neighborhood more than a visually desirable street. Importantly, we have developed a visual proxy measure that improves explainability with only minor losses in accuracy.

The focus on London as a single market reduces the extent the research can be generalized. Comparison between cities could potentially reveal differences; of which the aesthetic preferences of London and Tokyo are likely to differ.

Secondly, research is needed to better understand the less interpretable parts of the model. For example, extracting discriminative features between higher and lower house price from street images can potentially bring greater clarity to the model [3].

Thirdly, the images from Google Street View and Bing Aerial photos are not entirely reliable. Concerns can range from visual obstruction, poor lighting condition and differences in weather can affect the result.

Moreover, the work could be extended by making use of additional complementary cues, such as the images of the property interior [2, 27] and the views from within the property [31].

Another notable limitation concerns confounding environmental variables not accounted for in the hedonic price model. Additional environmental cues such as urban density and green foliage should be incorporated into the model in the future.

Even with these caveats, the results are encouraging. Developing more reliable house price model is an important topic for urban planning. The implication is that these models can be used to improve the visual quality of streets and neighborhoods through the implementation of housing policy.

6 CONCLUSION

We have presented a novel approach to house pricing that leverages visual knowledge of the urban environment to improve predictive power. In contrast to previous work [2, 26, 27] that have made use of images of the interior and exterior of the property for sale, we have focused on characterizing the neighbourhood of the property, and with the property making up only a small proportion of the aerial images; while the Street View images we make use of typically do not contain the property itself.

Our use of end-to-end training has allowed us to avoid the need for costly annotation of urban data, while still extracting meaningful proxy values from the urban environment. As well as improving the accuracy of standard models we believe that these visual proxies will be of interest to economists on estimating the willingness to pay for different levels of visual desirability. To that end we are both releasing the training code, allowing these features to be developed in new environments, and the pre-trained models, allowing the automatic generation of such proxy values in urban environments similar to London.

ACKNOWLEDGMENTS

This work was supported by The Alan Turing Institute under the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/N510129/1.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] E Ahmed and M Moustafa. 2016. House price estimation from visual and textual features. *arXiv:1609.08399[cs.CV]* (2016).

[3] S Arietta, A Efros, R Ramamooorthi, and M Agrawala. 2014. City Forensics: Using Visual Elements to Predict Non-Visual City Attributes. *IEEE Transactions on Visualization and Computer Graphics* (2014).

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *In ECCV*. 404–417.

[5] L Chen, G Papandreou, I Kokkinos, K Murphy, and A.L. Yuille. 2014. Semantic image segmentatiom with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

[6] J Cheshire and S Sheppard. 1995. On the Price of Land and the Value of Amenities. *Economica* (1995).

[7] François Chollet. 2015. keras. https://github.com/fchollet/keras. (2015).

[8] C Doersch, S Singh, C Wu, and W Hui. 2012. ACM Transactions on Graphics. *What makes Paris look like Paris* (2012).

[9] A Dubey, N Naik, D Parikh, R Raskar, and C Hidalgo. 2016. Deep Learning the City : Quantifying Urban Perception At A Global Scale. *European Conference on Computer Vision (ECCV)* (2016).

[10] Historic England. 2017. https://historicengland.org.uk/listing/the-list/data-downloads/. (2017).

[11] Office for national statistics. 2017. https://www.ons.gov.uk. (2017).

[12] T Gebru, J Krause, Y Wang, D Chen, J Deng, E Aiden, and F Li. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighbourhoods across the United States. *PNAS* (2017).

[13] Stephen Gibbons. 2003. Valuing English Primary Schools. *Journal of Urban Economics* (2003).

[14] R Girshick. 2015. Fast R-CNN. *IEEE International Conference on Computer Vision* (2015).

[15] Google. 2018. https://www.maps.google.com/. (2018).

[16] J. Jacobs. 1961. *The Death and Life of Great American Cities.* Random House Inc.

[17] A Krizhevsky, I Sutskever, and G.E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing* (2012).

[18] S Law. 2016. Defining Street-based Local Area and measuring its effect on house price using the hedonic price approach: the case study of metropolitan London. *Cities* (2016).

[19] S Law, C Seresinhe, and S Yao. 2017. An application of convolutional neural network in street image classification: the case study of london. *ACM Sigspatial'17: Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery* (2017).

[20] Y Li, M Paluri, J Rehg, and P Dollar. 2016. Unsupervised learning of edges. *CVPR* (2016).

[21] L Liu, E Silva, C Wu, and W Hui. 2017. A machine learning-based method for the large-scale evaluation of the urban environment. *Computers, Environment and Urban Systems* (2017).

[22] Microsoft. 2018. https://www.microsoft.com/en-us/maps/choose-your-bing-maps-api. (2018).

[23] N. Naik, J. Philipoom, R. Raskar, and C.A. Hidalgo. 2014. StreetScore - Predicting the Perceived Safety of One Million Streetscapes. In *CVPR Workshop on Web-scale Vision and Social Media.*

[24] Nationwide. 2012. (2012). Permission granted from LSE.

[25] R.B Palmquist. 1984. Estimating the demand for the characteristics of housing. *Review of Economics and Statistics* (1984).

[26] S Peterson and B Flanagan. 2009. Image based appraisal of real estate properties. *journal of Real Estate Research* (2009).

[27] Omid Poursaeed, Tomas Matera, and Serge Belongie. 2018. Vision-based Real Estate Price Estimation. *arXiv preprint arXiv:1707.05489* (2018).

[28] Land Registry. 2017. https://www.gov.uk/search-house-prices. (2017).

[29] R.G Ridker and J.A Henning. 1967. 'The Determinants of ResidentialProperty Values With Special Reference to Air Pollution'. *Review of Economics and Statistics* (1967).

[30] S Rosen. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* (1974).

[31] C Seresinhe, T Preis, and S Moat. 2017. Using deep learning to quantify the beauty of oudoor places. *Royal Society Open Science* (2017).

[32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[33] Streetscore. 2014. http://streetscore.media.mit.edu. (2014). Accessed: 2016-04-29.

[34] Ordnance Survey. 2017. https://www.ordnancesurvey.co.uk/opendatadownload/products.html. (2017).

[35] Q You, R Pang, L Cao, and J Luo. 2017. Neural Network hedonic pricing models in mass real estate appraisals. *IEEE Transactions on Multimedia* (2017).