Examining geographical generalisation of machine learning models in urban analytics through street frontage classification and house price regression

Stephen Law
UCL Geography, UK
The Alan Turing Institute, UK
stephen.law@ucl.ac.uk

Péter Jeszenszky
University of Bern, Switzerland
peter.jeszenszky@csls.unibe.ch

Keiji Yano
Ritsumeikan University, Japan
yano@lt.ritsumei.ac.jp

Abstract

The use of machine learning models (*ML*) in spatial statistics and urban analytics is increasing. However, research studying the generalisability of *ML* models from a geographical perspective had been sparse, specifically on whether a model trained in one context can be used in another. The aim of this research is to explore the extent to which standard models such as convolutional neural networks being applied on urban images can generalise across different geographies, through two tasks. First, on the classification of street frontages and second, on the prediction of real estate values. In particular, we find in both experiments that the models do not generalise well. More interestingly, there are also differences in terms of generalisability within the first case study which needs further exploration. To summarise, our results suggest that in urban analytics there is a need to systematically test out-of-geography results for this type of geographical image-based models.

1 Introduction

Machine learning (*ML*) methods such as convolutional neural networks (*CNN*) have achieved human-level accuracy in many computer vision tasks such as scene recognition, object detection and image segmentation [1, 16]. This level of computer intelligence has led to advances in intelligent transportation, medical imaging, robotics and in our case urban analytics. For example, these methods have been used to estimate socio-economic profiles [3], predict the perceived safety of streets [12, 20], classify street frontage quality [10] and to estimate property prices [9]. A key limitation is the lack of research on how machine learning methods on urban scenes generalise geographically. If a model trained in one context can be used in another then there is less data annotations and thus more generalisable and spatially reproducible models[7]. To address this concern, this exploratory research aims to study whether standard machine learning models (*CNN*) on urban images can generalise over vastly different geographical context on two common tasks in *ML*, namely an image-based classification task and a regression task.

1.1 Related work on the analysis of urban imagery

Diving deeper into the analysis of urban imagery, Salesses et al. [18] collected data on the perception of safety from street image, using a crowd-sourced survey to study the number of homicides in US cities. Naik et al. [12] expanded on this by fitting a regression model [20] to predict perceived safety and liveliness. Recently, Law et al. [10] have constructed a CNN model to infer whether the street has active frontages or not. While, Law et al. [9], used both street level and aerial images to estimate house price directly using a CNN-based hedonic price model for the Greater London area.

Despite the increase in research using urban imagery, studying how these models generalise geographically has been limited. Naik et al. [12] found that their urban computer vision models generalise poorly between the East and the West Coast in the United States. In an attempt to obtain a global model, [2] extended the Place Pulse dataset to 56 cities around the world. Using this dataset, Dubey et al. [2] trained a CNN model that can predict pairwise perceived safety from a pair of input StreetView images. Subsequently, they used this global model to make a similar prediction for six additional cities and found the prediction score conforms well through visual inspections. Our research main novelty is to study the concept of *ML* model generalisation from a geographical perspective; through a classification task (street frontage classification) and a regression task (real estate value prediction). For brevity, we term these case study 1 and case study 2.

2 Method and Materials

2.1 Case study 1: Street Frontage Classification

The quality of street frontages is an important factor in urban design, as it contributes to the safety and liveliness of the public space [5]. In this study, active street frontage is defined as having windows and doors on the ground floor of the building frontage, as opposed to blank walls [14]. In case study 1, we investigate the extent to which a street frontage classification model which classifies a Google StreetView image into four frontage categories; blank frontage, single-side active frontage , both-sides active frontage and non-urban frontage can generalise to different geographical contexts.

Front-facing street images were firstly collected using Google StreetView API [4] following similar procedures to [10]. In total we downloaded 109,419 front-facing StreetView images in London, 5972 images in Kyoto, 2157 images in Hong Kong, 6012 images in Tokyo, 2746 images in Barcelona, 4157 images in San Francisco, 3143 images in NYC and 4434 images in Paris. In London, 10,000 images were manually labelled in order to train the initial model, and in each of the seven cities, 350 images were labelled.

Following [10], we train a Street-Frontage-Net classifier $SFN(\cdot)$ that takes Streetview image S as input and returns a probability vector for each frontage class k. SFN uses a pretrained VGG16 architecture [19] from Imagenet as a feature extractor. These features then get pushed through a pair of fully-connected layers where a Softmax activation function is used in the final layer to estimate the probability of the four frontage class for an input image. We then split the dataset and use 60% for training, 20% for validation and 20% for testing and train the SFN using stochastic gradient descent (*lr=0.001*). We minimise the categorical cross entropy loss function; $H(y, \hat{y}) = -\sum_{k=1}^{M} y_k log(\hat{y}_k)$ where $\hat{y}_k$ is the predicted probability for class k with M classes, and y_k is the true probability for the same class. For more details of the data collection process and architecture, please see Law et al. [10].

For case study 1, we study the extent to which the SFN model trained in London can

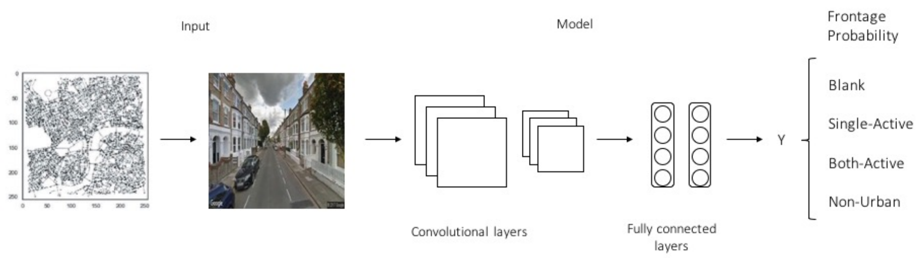

Figure 1 Case Study 1: Street frontage classification model [10]

generalise across the seven other cities. We report the classification accuracy, or the number of times the prediction of the frontage class matches the four observed frontage classes. Fig 2 shows example of the streetview images.

2.2 Case study 2: Real estate value prediction

In case study 2, we study the extent to which an urban image-based real estate value regression model can generalise between London and Kyoto. We adopt an existing end-to-end methodology akin to [9] that estimates the real estate value from both its location attributes and visual attributes from urban images. To ensure that the cases are more comparable, we construct a parsimonious hedonic price model to predict the real estate value (price per sqm) based on location and visual attributes at the street segment level.

Figure 2 Examples of Google Street images from left to right, London, Kyoto, Paris and Tokyo.

In terms of the property attributes, we use the UK Land Registry Price Paid dataset [15], coupled with detail attributes from Nationwide Housing Society [13] to form the house price data in London. For Kyoto, we used the Rosenka dataset, which is a road valuation dataset from 2012 which gives the mean land price per sqm for each street [17]. We calculate the mean house price sqm at the street-level from the London data in order to match with the Kyoto data. In terms of the location attributes, we calculate two street network accessibility measures which are commonly included in house price models [9]. Specifically, we calculate *closeness centrality*, which measures the inverse average distance to all other streets in the network as a proxy for capturing geographic accessibility, and *betweenness centrality*, which measures the number of shortest paths overlap from all streets to all streets as a proxy for street hierarchy and congestion of a city [6].

In terms of the visual attributes, we used the same front-facing streetview images from case study 1 for London. Following [9], we have also collected aerial images using Microsoft Bing Maps API [11] for both London and Kyoto. In total, the dataset consists of $39,346$ aerial image samples in London and $7,040$ in Kyoto. The output variable, price per sqm, is log transformed, which is a standard procedure in the literature [9], while all the input attributes are normalised to have a mean of 0 and a standard deviation of 1.

Following [9], we train a model $H(\cdot)$ with the streetview and aerial images while controlling for the contribution of the housing attributes. To extract visual features from the StreetView images S and aerial images A, we define two functions $F(S)$ and $G(A)$ which extract features as additional inputs into a hedonic price model. Both networks adopt a VGG-like [19] *CNN* architecture, where we take the value at the final flattened convolutional layer followed by a pair of fully-connected layers. We then concatenate the output of these two networks followed by two additional fully-connected layers in compressing the feature vectors output of $F(S)$ and $G(A)$ to a visual summary scalar response.

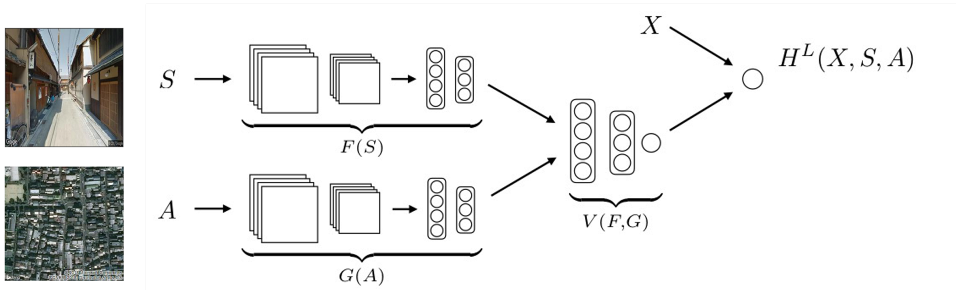

Figure 3 Case study 2: Hedonic price model architecture [9]

This visual response can then be included as an additional independent variable in an *OLS* model where we can compare a standard linear model; $H^L(X) = \beta_0 + \sum \beta X + \epsilon$, which only uses the housing attributes X, to an extended model $H^L(X, S, A)$ that includes the visual summary response as $H^L(X, S, A) = \beta_0 + \sum \beta X + \gamma V(F(S), G(A)) + \epsilon$, where β are the OLS regression weights for the location attributes, and γ as the weights for the visual summary response. We then split the dataset and use 70% for training, 15% for validation and 15% for testing and train the model using ADAM [8](learning rate=0.001) minimising the mean squared error loss function. For more details of the data collection process and architecture, please see Law et al. [9].

The aims of case study 2 are two-fold. First, to test whether the method works in a vastly different context, in this case Kyoto. Second, to test the extent to which the image features trained with the London data can be used and generalised to Kyoto and vice versa. To address both of these aims, we estimated six linear regression models on the testset, each of which are different combinations of housing attributes, and visual attributes of the two cities. Hedonic price models **M1** to **M3** deliver predictions for London, while models **M4** to **M6** for Kyoto. Model **M1** is the baseline hedonic price model for London that includes the housing attributes only. Model **M2** is the same as the London-baseline but includes both housing attributes and visual response retrieved from the London-trained-CNN model on London images. Model **M3** includes both the housing attributes and visual response retrieved from the Kyoto-trained-CNN model on London images. Model **M4** is the baseline hedonic price model for Kyoto that includes the housing attributes only. Model **M5** is the same as the Kyoto-baseline but includes both the housing attributes and the visual response retrieved from the Kyoto-trained-CNN model on Kyoto images. Model **M6** includes both the housing attributes and the visual response retrieved from London-trained-CNN model on Kyoto images. For each model, we report the adjusted R-squared measures, as a general goodness of fit metric (Table 1).

3 Results and Conclusion

Presenting the results of case study 1, Table 1 shows the accuracy of 87.5% for the baseline London model which were used to make inference for the seven other cities namely; Paris at 77.26%, New York at 73.30%, Barcelona at 70.48%, San Francisco at 69.43%, Hong Kong at 67.78%, Kyoto at 56.25% and Tokyo at 52.20%. These results confirm a naive assumption that architecturally more similar cities can achieve a higher accuracy.

Table 1 Case study 1 results

Cities	Accuracy
London	87.50%
Paris	77.26%
NYC	73.30%
Barca	70.48%
SFO	69.43%
HKG	67.78%
Kyoto	56.25%
Tokyo	52.20%

Table 2 Case study 2 results

Location	Model	adjR2
London	M1 (noVis)	63.90%
London	M2 (LonVis)	71.6%
London	M3 (KyoVis)	63.90%
Kyoto	M4 (noVis)	29.30%
Kyoto	M5 (KyoVis)	42.40%
Kyoto	M6 (LonVis)	29.90%

Table 2 shows the goodness of fit ($adjR2$) results for case study 2, comparing the six regression models. The results show that the goodness of fit improved from 63.9% (**M1** London baseline) to 71.6% for London (**M2**) and from 29.3% (**M4** Kyoto baseline)to 42.4% for Kyoto (**M5**) when including its own visual response. However, there is no improvement when using the Kyoto visual response in the London hedonic price model (**M3**) and a negligible improvement when using the London visual response in the Kyoto model (**M6**).

To summarise, this exploratory research studied whether a standard (*ML*) model such as *CNN* can generalise well geographically for two tasks, classification of street frontages and prediction of real estate values. For both tasks, we have found poor model generalisability across different geographical contexts, albeit we also noticed differences in generalisability. For example in case study 1, we found that the street frontage classification model trained using only the London StreetView images generalises better to cities that are architecturally more similar to London, such as Paris (eg. western style, bricks, stones), and poorer for cities that are architecturally dissimilar, such as Kyoto (eg. eastern style, wood, concrete). In case study 2, we confirm that response extracted from urban images can improve existing real estate value predictions for both London and Kyoto. However, we also found that the visual response learnt from one context cannot be easily generalised to another context, echoing the result of previous research [12]. A number of limitations remain, including the lack of samples and the lack of cross cities analysis. For example, whether a model trained in other cites can generalise to London and whether a model trained in a subset or all of the cities can generalise better (eg. Dubey et al. 2016 [2]). There were also a lack of case studies in the house price prediction tasks due to the difficulty in collecting comparable data in different cities. From a geographical perspective, future research could also consider how spatial dependence differs across different geographies for this type of model. To end, these results suggest that there is a need to systematically test *ML* models in different geographies as well as the need for human evaluation experiments to study these differences in detail for future research. Even though the results are not conclusive, it serves as an initial exploration on *ML* models generalisation from a geographical perspectives.

References

1 Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2016. `arXiv:1511.00561`.

2 A Dubey, N Naik, D Parikh, R Raskar, and C Hidalgo. Deep learning the city : Quantifying urban perception at a global scale. *European Conference on Computer Vision (ECCV)*, 2016.

3 T Gebru, J Krause, Y Wang, D Chen, J Deng, E Aiden, and F Li. Using deep learning and google street view to estimate the demographic makeup of neighbourhoods across the united states. *PNAS*, 2017.

4 Google. `https://www.maps.google.com/`, 2018. Google StreetView retrieved in 2018.

5 E Heffernan, T Heffernan, and W Pan. The relationship between the quality of active frontages and public perceptions of public spaces. *Urban Design international*, 2014.

6 B. Hillier and O. Shabaz. An evidence based approach to crime and urban design – Or can we have vitality, sustainability and security all at once? $http : //spacesyntax.com/wp - content/uploads/2011/11/Hillier - Sahbaz_An - evidence - based - approach_010408.pdf, visited : June2014$, 2005.

7 Peter Kedron, Amy E Frazier, Andrew B Trgovac, Trisalyn Nelson, and A Stewart Fotheringham. Reproducibility and replicability in geographical analysis. *Geographical Analysis*, 53(1):135–147, 2021.

8 Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. `arXiv:1412.6980`.

9 Stephen Law, Brooks Paige, and Chris Russell. Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–19, 2019.

10 Stephen Law, Chanuki Illushka Seresinhe, Yao Shen, and Mario Gutierrez-Roig. Street-frontage-net: urban image classification using deep convolutional neural networks. *International Journal of Geographical Information Science*, 0(0):1–27, 2018. `arXiv:https://doi.org/10.1080/13658816.2018.1555832`, `doi:10.1080/13658816.2018.1555832`.

11 Microsoft. `https://www.microsoft.com/en-us/maps/choose-your-bing-maps-api`, 2018. Bing Aerial Maps retrieved in 2018.

12 N. Naik, J. Philipoom, R. Raskar, and C.A. Hidalgo. Streetscore - predicting the perceived safety of one million streetscapes. In *CVPR Workshop on Web-scale Vision and Social Media*, 2014.

13 Nationwide. Auxiliary housing attributes. permission of use from london school of economics. `https://www.nationwide.co.uk/`, 2012.

14 Office of the Deputy Prime Minister. *Safer Places: The Planning System and Crime Prevention.* Home Office, 2005.

15 Land Registry. `https://www.gov.uk/search-house-prices`, 2017.

16 Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. `arXiv:1506.01497`.

17 Hadrien Salat, Roberto Murcio, Keiji Yano, and Elsa Arcaute. Uncovering inequality through multifractality of land prices: 1912 and contemporary kyoto. *PloS one*, 13(4):e0196737, 2018.

18 P. Salesses, K. Schechtner, and C.A. Hidalgo. Image of The City: Mapping the Inequality of Urban Perception. *PLoS ONE 8 (7): e68400.Doi:10.1371/journal.pone.0068400*, 8(7), 2013.

19 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

20 Streetscore. `http://streetscore.media.mit.edu`, 2014. Accessed: 2016-04-29.