

Working Paper 176

Using Machine Learning to Create a Property Tax Roll: Evidence from the City of Kananga, D.R. Congo

Augustin Bergeron,
Arnaud Fournier,
John Kabeya Kabeya,
Gabriel Tourek,
Jonathan L. Weigel

November 2023

ICTD Working Paper 176

Using Machine Learning to Create a Property Tax Roll: Evidence from the City of Kananga, DR Congo

Augustin Bergeron, Arnaud Fournier, John Kabeya Kabeya and
Gabriel Tourek

November 2023

Using Machine Learning to Create a Property Tax Roll: Evidence from the City of Kananga, DR Congo

Augustin Bergeron, Arnaud Fournier, John Kabeya Kabeya, Gabriel Tourek and Jonathan L. Weigel

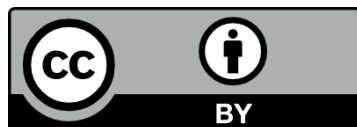
ICTD WORKING PAPER 176

First published by the Institute of Development Studies in NOVEMBER 2023

© Institute of Development Studies 2023

ISBN: 978-1-80470-153-9

DOI: [10.19088/ICTD.2023.053](https://doi.org/10.19088/ICTD.2023.053)



This is an Open Access paper distributed under the terms of the Creative Commons Attribution 4.0 International license (CC BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original authors and source are credited and any modifications or adaptations are indicated. <http://creativecommons.org/licenses/by/4.0/legalcode>

Available from:

The International Centre for Tax and Development at the Institute of Development Studies, Brighton BN1 9RE, UK

Tel: +44 (0) 1273 606261

Email: info@ictd.ac

Web: www.ictd.ac/publication

Twitter: [@ICTDTax](https://twitter.com/ICTDTax)

Facebook: www.facebook.com/ICTDTax

IDS is a charitable company limited by guarantee and registered in England

Charity Registration Number 306371

Charitable Company Number 877338

Using Machine Learning to Create a Property Tax Roll: Evidence from the City of Kananga, DR Congo

Augustin Bergeron, Arnaud Fournier, John Kabeya Kabeya, Gabriel Tourek and Jonathan L. Weigel

Summary

Developing countries often lack the financial resources to provide public goods. Property taxation has been identified as a promising source of local revenue, because it is relatively efficient, captures growth in real estate value, and can be progressive. However, many low-income countries do not collect property taxes effectively due to missing or incomplete property tax rolls.

We use machine learning and computer vision models to construct a property tax roll in a large Congolese city. To train the algorithm and predict the value of all properties in the city, we rely on the value of 1,654 randomly chosen properties assessed by government land surveyors during in-person property appraisal visits, and property characteristics from administrative data or extracted from property photographs. The best machine learning algorithm, trained on property characteristics from administrative data, achieves a cross-validated R^2 of 60 per cent, and 22 per cent of the predicted values are within 20 per cent of the target value. The computer vision algorithms, trained on property picture features, perform less well, with only 9 per cent of the predicted values within 20 per cent of the target value for the best algorithm.

We interpret the results as suggesting that simple machine learning methods can be used to construct a property tax roll, even in a context where information about properties is limited and the government can only collect a small number of property values using in-person property appraisal visits.

Keywords: property tax; machine learning; Democratic Republic of Congo; computer vision; property valuation; state capacity.

Augustin Bergeron is an Assistant Professor of Economics at the University of Southern California. His research lies at the intersection of development economics, public economics and political economy.

Arnaud Fournier is CEO and Co-founder at Bastion Technologies.

John Kabeya Kabeya is the head of the tax revenue and tax base division at the Provincial Tax Ministry of the Kasai-Central Province (DGRKAC).

Gabriel Tourek is an Assistant Professor of Economics at the University of Pittsburgh. His research focuses on the development of fiscal and state capacity, and the equity of taxes and transfers in low-income countries.

Jonathan L. Weigel is an Assistant Professor in the Business and Public Policy Group at the Haas School of Business at the University of California, Berkeley. His research interests lie at the intersection of political economy, development and public economics.

Contents

	Summary	3
	Acknowledgements	6
	Acronyms	6
1	Introduction	7
2	Background on property valuation and taxation	9
3	Setting	11
4	Data	11
	4.1 Property values	11
	4.2 Property features	14
	4.3 Geographic features	15
	4.4 Missing values	17
	4.5 Property photographs	17
5	Machine learning	17
	5.1 Methodology	18
	5.2 Algorithms	18
	5.3 Performance evaluation metric	19
	5.4 Results	20
	5.5 Feature importance	21
6	Computer vision	22
	6.1 Convolutional neural networks	22
	6.2 Convolutional neural networks with a small training sample	23
	6.3 Image pre-processing	23
	6.4 Results	23
7	Comparison with other studies	24
8	Conclusion	25
	References	27
Tables		
Table 4.1	Features used to train machine learning modules	15
Table 5.1	Performance of machine learning and computer vision models	20
Figures		
Figure 4.1	Document measuring property characteristics	12
Figure 4.2	Document with property valuation given to respondents	13
Figure 4.3	Training sample – distribution of property values	14
Figure 4.4	Location of important infrastructure	16
Figure 4.5	Example of property photographs	17
Figure 5.1	Feature importance – number of times a feature is used to split the data across all trees	22

Acknowledgements

We thank Edward Glaeser, Nikhil Naik and Wilson Prichard for their encouragement and advice. For outstanding research assistance, we thank Elie Kabue Ngindu and Michele Bernadine. We thank our fantastic team of enumerators for excellent programme management – Augustin Nikebele Nzambi, Costantin Tshimanga Mukendi, Junior Milongo Mushidi and Théodore Kalamba Muepu. We are very grateful for the collaboration with the provincial government of Kasai-Central. We are particularly grateful for the assistance of government land surveyors and managers from the cadastral service of the Kasai-Central Province in Kananga who worked on this project – Ambroise Kankonde Kanunseki, Oscar Tshibambe Mfuamba, Noé Tshitenge Kalamba, Jean Balanganyi Mpenge, Jean-Marie Lubula Kasadi and Patrice Mukenge Makolo. We gratefully acknowledge funding from the International Centre for Tax and Development (ICTD), the Harvard Lab for Economic Application and Policy (LEAP), and the IPA Research Methods Initiative (IPA RMI). This study has been approved by the Harvard Institutional Review Board: Protocol IRB17-0724.

Acronyms

CNN	Convolutional neural networks
DRC	Democratic Republic of the Congo
GAN	Generative adversarial network
MAE	Mean absolute error
MAPE	Mean absolute percentage error
OLS	Ordinary least squares
RBF	Radial basis function
SVM	Support vector machine
SVR	Support vector regression
VGG	Visual geometry group

1 Introduction

Governments in the world's poorest countries face severe revenue constraints. They typically collect less than 10 per cent of Gross Domestic Product (GDP) in taxes, compared to the 25-50 per cent collected in high-income countries (Pomeranz and Vila-Belda 2019). The literature on state capacity and development argues that inability to collect taxes is at the heart of why low-income countries are as poor as they are (Kaldor 1965; Besley and Persson 2009, 2013; Besley et al. 2013; Dincecco and Katz 2016; Mayshar et al. 2022). This research suggests that the path to economic prosperity may begin with investment in the capacity of governments to collect the tax revenue necessary to provide productivity-enhancing public goods.

At the local level, property taxation is often the primary source of government revenue, and is essential for provision of local public goods (Rosengard 1998; Collier 2017; Fjeldstad et al. 2017). Property taxation has many advantages. First, it is economically efficient, because it is hard to avoid, and easily enforceable if evaded. Second, it is socially equitable – it is often progressive, a relatively good proxy for a wealth tax, and a way for the public sector to capture a share of private sector gains from real estate value appreciation. Beyond the economic efficiency of property taxation, research suggests that the political salience of property taxes may lend itself to the development of a fiscal social contract between citizens and the state (Prichard 2015; Weigel 2020). However, property taxation remains one of the most under-utilised taxes in developing countries (Brockmeyer et al. 2023), partly because taxing properties requires mapping and assessing property values, which is complex and expensive. Only 39 per cent of the 159 non-OECD countries in the World Bank's *Doing Business* surveys have mapped their largest city's private plots. In sub-Saharan Africa, only 14.6 per cent of the main cities' private plots have been mapped (Lall et al. 2017).¹

To improve their capacity to collect property tax, low-income countries need to develop and adopt methods to map and value properties in a cost-effective way. Several approaches have been proposed (see Zebong et al. (2017) for a review).² One possibility is to rely on property valuation by experts during in-person appraisal visits. These visits typically result in a precise valuation of properties, but have the disadvantage of being time-consuming, costly, and prone to collusion between land surveyors and property owners.³ Given these limitations, many countries have instead adopted simplified valuation methods, such as area-based valuation, which consist of using the area of land and buildings to assess a property's value. Area-based valuation has the advantage of being transparent, easy to verify, and equitable in terms of the size of land or properties. However, it has the disadvantage of failing to adequately incorporate qualitative aspects of buildings, which significantly limits its fairness, and can lead to regressive property tax rates. Given the drawbacks of both approaches there has been an increased interest in implementing simplified hybrid methods – methods that are designed to be both practically feasible and equitable, such as points-based valuation. Points-based valuation methods account for the limitations of area-based methods by assigning points. These are based on the surface area of land and buildings;

¹ Governments often collect property taxes even when they lack an accurate and complete property tax roll. Property taxes in these settings are often beset by inefficiencies, such as a narrow tax base (Casaburi and Troiano 2016) or simplified tax categories (presumptive taxation), which result in a regressive tax schedule (Fjeldstad et al. 2017).

² The choice between different approaches typically depends on their cost, feasibility, speed, and how prone they are to corruption and leakages (Ali et al. 2018).

³ Khan et al. (2015) document that tax collectors frequently collude with taxpayers to reduce the tax assessment in exchange for a bribe in Punjab, Pakistan.

additional points are awarded for positive features and deducted for negative features of the property.⁴ Because they combine feasibility, simplicity and equity, points-based methods have been adopted in many low-capacity countries, such as Pakistan, Sierra Leone and Malawi.

In this study we take inspiration from these simplified hybrid approaches to property valuation. But we propose instead to predict the value of properties by training machine learning and computer vision algorithms on a small dataset of property values assessed by expert land surveyors, and property characteristics available for all properties through administrative data or extracted from property photographs. This approach is especially appropriate in contexts where information about properties is limited, and the government can only collect property values using assessor valuation for a small set of properties due to limited capacity and resources. We implement this approach in the city of Kananga in the Democratic Republic of the Congo (DRC). We worked with six land surveyors from the cadastral service of the provincial government of Kasai-Central to evaluate the value of 1,654 randomly chosen properties, and had access to data from a property census containing property characteristics and property photographs, as described in Balan et al. (2022) and Bergeron et al. (2023).⁵

We first predict the value of all properties in Kananga by training several machine learning algorithms on the sample of property values assessed by land surveyors, using property characteristics from an administrative property census. The machine learning model with the highest out-of-sample accuracy relies on ensemble modelling, and is a combination of boosting models. Using 10-fold cross-validation to assess its out-of-sample prediction accuracy, we find that it achieves an R^2 of 60 per cent, a Mean Absolute Percentage Error (MAPE) of 70 per cent, and that 22 per cent of the predicted values are within 20 per cent of the target value. This approach is simple, but it requires access to property characteristics from administrative data, which is rarely available in settings where there is low state capacity.

For this reason, we also predict the value of all properties in Kananga by training computer vision algorithms using characteristics extracted from property photographs, and the same sample of property values assessed by land surveyors. The advantage of this approach is that it does not require property characteristics from administrative data, and instead leverages existing information about properties contained in property photographs. However, we find that computer vision algorithms trained on features extracted from photographs of properties perform less well than machine learning methods trained on property characteristics from administrative data. The best algorithm uses a deep convolutional neural network (CNN) for feature extraction. Using 10-fold cross-validation, we find an R^2 of 40 per cent, a MAPE of 99 per cent, and only 9 per cent of the predicted values within 20 per cent of the target value. While there is interest in using property pictures (Glaeser et al. 2018) and satellite imagery (Bency et al. 2017; Bachofer et al. 2020) to predict property value, our results suggest that this approach works less well when a training sample is small and little information can be extracted from property photographs.⁶

⁴ The assignment of points to property characteristics can be done manually or using quasi-regressions or regressions (Zebong et al. 2017), which is usually more accurate.

⁵ The provincial government of Kasai-Central has repeatedly shown interest in increasing tax revenue using evidence-based policy (Balan et al. 2022; Bergeron et al. 2023), and is interested in evaluating different ways of constructing a property tax roll in order to improve the progressivity of the property tax schedule.

⁶ As we discuss in Section 6, it is easier to extract informative features from Google Street View photographs (Naik et al. 2014; Glaeser et al. 2016; Naik et al. 2017; Glaeser et al. 2018; Law et al. 2019) than from photographs of properties taken in Kananga.

Our approach contributes to the literature that gives machine learning a prominent place in the economics toolbox (Mullainathan and Spiess 2017; Athey and Imbens 2019).⁷ It is closely related to studies that have used property market prices and machine learning and computer vision methods to predict property values in high-income settings (Chopra et al. 2007; Bency et al. 2017; Glaeser et al. 2018; Law et al. 2019). In our context, machine learning and computer vision algorithms perform less well than in these settings, which is hardly surprising given their larger training datasets and richer features. However, our results are comparable to studies that estimate property value using machine learning in low-income settings (Bachofer et al. 2020; Knebelmann et al. 2023). This is despite having a smaller training sample and more limited property characteristics. We, therefore, view our contribution as showing that machine learning methods can be successfully used to construct a property tax roll – even in contexts where the government has access to a small number of assessor valuations, and limited information about property characteristics.

The rest of the paper is organised as follows. Section 2 provides background on property valuation and taxation. Section 3 describes the setting, and Section 4 presents the data used to predict property value in Kananga. Section 5 describes the machine learning algorithms, and Section 6 the computer vision models used to predict the market value of properties. Section 7 compares our results with other studies that have used machine learning methods to predict property value in high and low-income settings. Section 8 concludes.

2 Background on property valuation and taxation

Low-income countries have adopted several methods to map and value properties for property tax collection. Many countries rely on direct assessment of property value by experts, often through in-person appraisal visits. These visits typically result in a precise valuation of properties, but have the disadvantage of being time-consuming, costly and prone to collusion between land surveyors and property owners. Given these limitations, many countries have instead adopted simplified valuation approaches, like area-based valuation, which consist of using the area of land and buildings to assess a property's value. Area-based valuation is transparent, easy to verify and equitable regarding the size of land or properties. However, it has the disadvantage of failing to adequately incorporate qualitative aspects of buildings, which can lead to regressive and unfair property tax rates.

Due to limited capacity and the drawbacks of both approaches, there has been an increased interest in implementing simplified hybrid methods that are designed to be practically feasible and equitable. Points-based valuation, for example, accounts for the limitations of area-based methods by assigning points based on the surface area of the land and buildings, and additional points are awarded for positive features and deducted for negative features of a property.⁸ Points-based methods initially require the manual assignment of points to property

⁷ Machine learning methods have been used to predict local economic outcomes using satellite data (Jean et al. 2016), large-scale phone network data (Blumenstock et al. 2015; Blumenstock 2016), and neighbourhood safety, income, and change in appearance using images from Google Street View in New York City and Boston (Naik et al. 2014; Glaeser et al. 2016; Naik et al. 2017).

⁸ A related approach consists of allocating properties to bands of values that are taxed differently and defined based on the property's observable characteristics, such as land area and buildings (Lim et al. 2008). McCluskey et al. (2002) and Davis et al. (2012) analyse the performance of property tax banding relative to *ad valorem* property tax

characteristics (Jibao and Prichard 2015), which is often time-consuming (Jibao 2017), inequitable (Jibao 2017; Manwaring and Regan 2019), and politically challenging to change (Jibao and Prichard 2015, 2016; Jibao 2017). To circumvent these issues points-based methods have increasingly relied on regressions to estimate property value, by calibrating a formula for property value using property characteristics and a sub-sample of property values (Fish 2018). This regression approach has been shown to perform significantly better than manually assigning values to different characteristics of properties (Manwaring and Regan 2019). Because they combine simplicity, feasibility and equity, points-based systems have been adopted in many low-capacity countries, including Pakistan, Sierra Leone and Malawi.

Inspired by these simplified hybrid approaches, recent studies have instead proposed to predict the value of properties using machine learning algorithms, which are typically trained using a small dataset of property values – obtained from transaction or market prices, or direct assessment by land assessors – and property characteristics available for all properties through administrative data or extracted from property photographs. The trained machine learning algorithms are then used to predict property values outside the training sample using property characteristics. Proponents of machine learning methods have argued that they are typically less costly, more accurate, and have been associated with less revenue leakage than direct valuation or simplified valuation approaches.

Machine learning methods perform well at predicting property value in high-income settings, characterised by large training datasets that rely on property transaction data and a rich set of property characteristics obtained from market data, property photographs and satellite imagery. For example, Chopra et al. (2007) use a non-parametric latent manifold model trained on a large dataset of houses from Los Angeles County to predict property value, Bency et al. (2017) use satellite images and a convolutional neural network framework to predict house prices in London, Birmingham and Liverpool. Glaeser et al. (2018) use Street View, house price data and computer vision methods to predict house prices in Boston, and Law et al. (2019) use Street View, satellite Images and computer vision methods to predict house prices in London.

Similar methods have also proved helpful in predicting property values in low-income settings, despite these settings being typically characterised by smaller training samples with more limited information about property characteristics.⁹ Bachofer et al. (2020) use maximum relevance and minimum redundancy models trained on several sources of information, including geospatial data, remote sensing data and house price data, to predict property value per m² in Kigali, Rwanda. Similarly, Knebelmann et al. (2023) use elastic-net regression models trained on property market values provided by assessors, combined with information on property sections, total built area (measured using high resolution satellite and drone images), number of floors and other observable property characteristics, to predict the annual rental value of properties in Dakar, Senegal.

Our study closely relates to Bachofer et al. (2020) and Knebelmann et al. (2023), and asks whether machine learning methods can predict property value for taxation in a country with less capacity and data, like the DRC. Machine learning methods might offer more accurate

valuation, and find that banding scores well in terms of simplicity, valuation costs and taxpayer comprehensibility, but performs less well with respect to fairness and progressivity.

⁹ Earlier studies have also used simple hedonic regressions of property characteristics on property value to predict property value outside the training sample. For example, Ali et al. (2018) combine high resolution satellite imagery with information on sales prices, targeted surveys and routine statistical data to predict property values in Rwanda. Similarly, Franklin (2019) uses hedonic regressions of observed rent on apartment characteristics to predict property values in Addis Ababa, Ethiopia.

predictions than area-based or points-based methods (Manwaring and Regan 2019), and might be associated with less collusion between tax assessors and taxpayers during in-person appraisal visits (Ali et al. 2018). It is worth mentioning that in a context like the DRC, where increasing tax revenue and tax compliance is a high priority, policymakers need to properly weigh the benefits, in terms of accuracy and reduced leakage, against the complexity and opacity of these models – which might reduce taxpayer understanding and lower tax morale (Manwaring and Regan 2019).

3 Setting

The DRC is one of the most populous countries in Africa, and also one of the poorest. Kananga, the capital of Kasai-Central Province and the setting for this study, is a city with 1-2 million inhabitants and an average monthly household income of US\$106 (PPP US\$168). The DRC is a low-capacity state, with a tax-to-GDP ratio ranking that is 188th out of 200 countries. Tax revenue is extremely low in the DRC. In the years before this study, the provincial government of Kasai-Central had tax revenue equal to roughly US\$0.3 per person per year. The government has turned to property tax to raise revenue – this currently accounts for about 26 per cent of provincial tax revenue. It has begun to extend the property tax net by launching annual city-wide collection campaigns (Weigel 2020; Balan et al. 2022; Bergeron et al. 2023).

As in many developing countries, tax authorities in the DRC tend to rely on simplified tax instruments (Fjeldstad et al. 2017). For example, when designing its property tax, the provincial government of Kasai-Central decided to rely on a flat property tax rate schedule (Balan et al. 2022; Bergeron et al. 2023). Properties built with non-durable materials are assigned to the low-value tax band and taxed FC3,000¹⁰ annually. Properties built with durable materials are assigned to the high-value tax band, and are taxed FC13,600 annually. While less resource-intensive to administer, simplified tax instruments can be regressive (Bergeron et al. 2023), and perceived as unfair (Robinson 2022). This paper explores the provincial government's attempts to rely on machine learning and computer vision models to create a property tax roll that could be used to improve the progressivity of the property tax.

4 Data

4.1 Property values

To train our machine learning and computer vision algorithms, we use information on the value of 1,654 properties randomly chosen from the 4,246 properties in the baseline sample of Balan et al. (2022). Six land surveyors from the cadastral services of the provincial government of Kasai-Central were in charge of assessing the market value of each of these properties, based on information collected during in-person appraisal field visits conducted between August and September 2019.¹¹

¹⁰ Or roughly US\$2 (as of 2019).

¹¹ The relatively small number of land surveyors involved in the appraisal field visits prevents us from studying heterogeneity by land surveyors.

Four land surveyors from the Cadastral Division of the provincial government of Kasai-Central conducted the in-person appraisal visits. Only the outside of the property was visited to avoid bias due to property owners refusing the visit. During appraisal visits the land surveyors collected information on the neighbourhood of the property (commune,¹² quartier (neighbourhood), localit  (locality), street), size of the plot, size of each construction, materials used for each construction, depreciation of these materials, and number of fruit trees on the property. This information was recorded by the land surveyors on an administrative form, *Proc s Verbal de Mesurage et de Bornage*, which the land surveyors filled out during the appraisal visits (see Figure 4.1).

Figure 4.1 Document measuring property characteristics (*Proc s Verbal de Mesurage et de Bornage*)

FORMULAIRE UNIQUE A UTILISER LORS DE LA VISITE 1

Pr nom, Nom, Post nom du propri taire :
 Polygone du r pondant (demander   l'enqu teur) :
 Code du r pondant (demander   l'enqu teur) :
 Nom de l'enqu teur :
 Nom de l'agent du cadastre :
 Nom du stagiaire des titres :

PROCES VERBAL DE MESURAGE ET DE BORNAGE

Province : _____ Parcelle cadastr e sous le num ro : _____
 District : _____ Ville : _____ Superficie : _____
 Quartier : _____
 Localit  : _____
 Commune : _____
 Lieu dit : _____

CROQUIS ORIENTE DE LA PARCELLE

Echelle: _____

BORNAGE

Dimension des bornes r glementaires moyennes :
 Bornes r glementaires moyennes de dimensions :

Sommets	Longueur	Angles

ELEMENTS DU MESURAGE

Sommets du polygone	Longueurs des cot�s r�duites � l'horizon (indiquer si apr�s mesures brutes R ou apr�s compensation d'angles calcul C)	Angles aux sommets grades (�) degr�s (�) (indiquer l'inscription qui ne convient pas)	Autres renseignements permettant le calcul de la superficie: r�tablissement et l'orientation du plan, ainsi que le r�glage �ventuel des bornes	Description
			Y X	

Erreur num rique de fermeture angulaire Lin aire
Instruments employ s pour le mesurage

La parcelle est situ e  
 Elle est cadastr e sous le n  Et enregistr e. Vol. folio
 Elle provient du morcellement de la parcelle cadastr e sous le n 
 Enregistr e volume folio
 Elle provient de la r unification des parcelles n  cadastraux.....
 Ha Volume folio

La recherche des limites   donn  lieu aux constatations suivantes :
 Superficie : Ha   ca

Nous avons plac  des bornes, r glementaires moyennes des dimensions... Marqu es
 ... des bornes moyennes de dimensions ... existaient aux sommets marqu es

Les constructions suivantes existaient au moment du mesurage :

Servitudes et autres mentions :

Combien de borne l gale la personne dispose :
 Combien de borne le r pondant doit il acheter pour obtenir son titre :

Sign  par

 Stagiaire de la Direction de Cadastre

Source: Authors own from project. Notes: This form was used by land surveyors to collect information on property characteristics. Land surveyors filled in the information based on the responses of property owners, and their own measurement of the dimensions of the property. The land surveyors collected information on the neighbourhood of the property (commune, quartier (neighbourhood), localit  (locality), street), size of the plot, size of each construction, materials used for each construction, depreciation of these materials, and number of fruit trees on the property.


¹² A commune is a level of administrative division in Kananga. There are five communes in Kananga - Kananga, Katoka, Lukonga, Ndesha and Nganza.

The two principal managers of the Cadastral Division of the provincial government, the general director and technical director, supervised the work of the four land surveyors. They were also in charge of estimating the market value of each property, based on the information collected by land surveyors during appraisal visits and reported on the *Procès Verbal de Mesurage et de Bornage*. To facilitate their work, they were also shown pictures of the property taken by enumerators who accompanied the land surveyors during the in-person appraisal visits. The general director and technical director of the Cadastral Division are experts in assessing the market value of properties in Kananga, and, due to their expertise, are often hired by banks to estimate the property value of clients interested in applying for a mortgage, using their property as collateral.

Properties were randomly selected from the baseline sample of Balan et al. (2022). Hardly any of the owners selected for an in-person appraisal visit by a land surveyor refused the visit. Property owners were informed that they would receive a document attesting to the value of their property following the appraisal visit, which might have incentivised them to accept the visit and partly explains the low refusal rates. After the land surveyors' appraisal visit and the managers' subsequent estimate, the owners of visited properties received a *Procès Verbal d'Expertise*, shown in Figure 4.2. This document was signed by the technical director of the Cadastral Division, and officially attested the value of the property estimated by the agents of the Cadastral Division. The attestation of the property's value was valuable to property owners who were interested in selling, or applying for a bank mortgage using their property as collateral.

Figure 4.2 Document with property valuation given to respondents

République Démocratique du Congo
 Province du Kasai-Central
 Circonscription Foncière de Kananga
 Division du Cadastre
KANANGA



Procès-Verbal d'Expertise n°.....

L'An deux mil vingt, le jour du mois de,

Nous arpenteur du Cadastre de Kananga et y résidant, déclarons ce jour procéder aux travaux d'expertise de la parcelle portant le numéro cadastral SU :, d'une superficie de a ca %, occupée par Monsieur/Madame suivant volume folio

La parcelle détient maison(s) :

- dont la 1^{ère} en superficie bâtie
- 2^{ème} en superficie bâtie
- 3^{ème} en superficie bâtie
- 4^{ème} en superficie bâtie
- 5^{ème} en superficie bâtie

Total :

Vétusté de la maison :

Valeur de la parcelle
 Valeur vénale estimative est de :

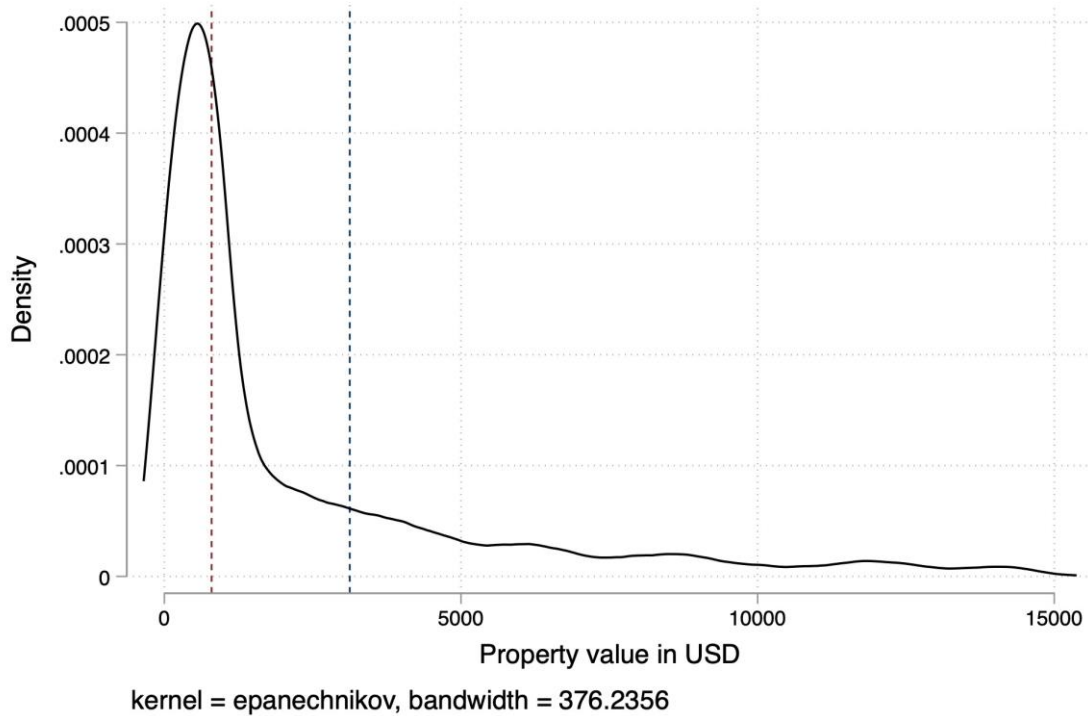
En foi de quoi ce P.V. d'expertise est établi aux jour, mois et an que dessus.

L'Arpenteur du Cadastre
 Oscar Tshubambe Mfiamba

Source: Authors own from project. Notes: This document was provided to respondents whose properties were surveyed after their properties had been evaluated by the two expert committees.

Figure 4.3 gives the distribution of the 1,654 properties in the sample, and shows substantial variation in property values in the city of Kananga. The minimum property value in the sample is US\$28, and the maximum US\$158,401. The mean property value is US\$3,450 (blue dotted line), and the median value is lower at US\$940 (red dotted line). The substantial difference in property value between the mean and the median is explained by the fact that distribution of property values is skewed to the left.

Figure 4.3 Training sample - distribution of property values



Source: Authors own from collected data. Notes: This figure shows the distribution of property values for the 1,654 properties in the training sample. The value of each property is estimated by a team of land surveyors during an in-person field visit. The estimation of the property value by the land surveyors is based on six criteria: (i) neighbourhood, (ii) property size, (iii) home size, (iv) materials used in the construction of the home, (v) home depreciation, and (vi) number of fruit trees on the property. The median property value in the training sample (US\$797) is represented by the vertical red dotted line. The mean property value in the training sample (US\$3,125) is represented by the vertical blue dotted line.

4.2 Property features

We train the machine learning algorithms using property characteristics from administrative data described in Balan et al. (2022) and Bergeron et al. (2023). During the property registration that preceded the 2018 property tax campaign, tax collectors, accompanied by enumerators from our research team, collected information on the materials used to construct the walls, roof and fence of the main house of each compound in the city. They also recorded the quality of the road on the nearest street, and whether the property was threatened by erosion.¹³ These variables are described in detail in Panel A of Table 4.1.

¹³ This survey was conducted with every property owner in Kananga 2-4 weeks after the 2018 tax collection campaign ended in each neighbourhood.

Table 4.1 Features used to train machine learning models

Category	Description
	Property latitude and longitude Communes (1-5 indicator) Geographic stratum (1-12 indicator)
<i>Panel A:</i> <i>Property</i>	Materials of the fence (1-4 scale) Roof materials (1-4 scale) Roof quality (1-4 scale) Wall quality (1-7 scale) Road quality (1-5 scale) Erosion threat (1-3 scale)
	Distance of the property to the city centre Distance of the property to the nearest commune building Distance of the property to the nearest gas station Distance of the property to the nearest health centre Distance of the property to the nearest hospital
<i>Panel B:</i> <i>Distance to infrastructure</i>	Distance of the property to the nearest market Distance of the property to the nearest police station Distance of the property to the nearest private school Distance of the property to the nearest public school Distance of the property to the nearest university Distance of the property to the nearest government building Distance of the property to the nearest road Distance of the property to the nearest ravine
	K-Fold target encoded geographic stratum property value K-Fold target encoded neighbourhood property value
<i>Panel C:</i> <i>Value of nearby properties</i>	Average property value in a 200 m radius Average property value in a 500 m radius Average property value in a 1 km radius Average price of the 3 nearest properties Average price of the 5 nearest properties

Source: Authors own from collected data. Notes: This table shows the features used to train the machine learning models. The property features in Panel A come from the registration and midline surveys of Balan et al. (2022) and Bergeron et al. (2023). They also leverage administrative data about the boundaries of the five communes of Kananga and the geographic strata used in Balan et al. (2022), which are smaller geographical units than the communes. Distance to infrastructure in Panel B is computed as the crow flies between the GPS location of the compound and the nearest infrastructure of each type collected during the city-wide infrastructure census conducted by the enumerators in September 2019. The value of nearby property features in Panel C are computed using the sample of 1,654 property values assessed by government land surveyors during in-person visits conducted in August and September 2019.

4.3 Geographic features

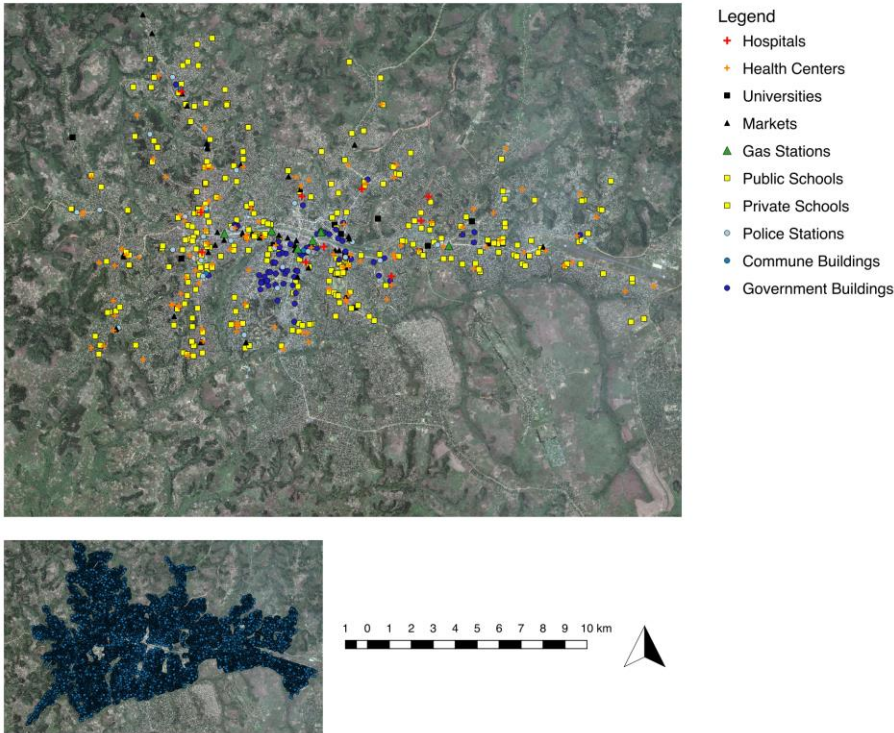
We also train our machine learning algorithms using geographic information:

Distance to infrastructure. In September 2019 staff from the provincial government, accompanied by enumerators from our research team, conducted a city-wide infrastructure survey. This recorded the GPS location of the following types of infrastructure in Kananga: hospitals, health centres, public schools, private schools, universities, markets, gas stations, police stations and government buildings (communal, provincial and national).¹⁴ Figure 4.4 shows a satellite map of Kananga containing the GPS location of each type of infrastructure. Additionally, administrative maps of the city's main roads, and areas subject to severe erosion, were digitised. For each type of infrastructure, we compute the distance of each

¹⁴ One limitation of this data is that it only records the location of each type of infrastructure, and does not contain any information on the quality of each type of infrastructure.

compound in the city to the nearest infrastructure using the GPS location of every compound collected during the property registration preceding the 2018 property tax campaign, described in Balan et al. (2022) and Bergeron et al. (2023).¹⁵ These variables are described in detail in Panel B of Table 4.1.

Figure 4.4 Location of important infrastructure



Source: © Google Maps 2016, adapted by Augustin Bergeron, CC-BY. Please ask for permission to re-use adapted material. Notes: This figure shows a satellite map of the city of Kananga with the GPS location of the main infrastructure (top panel) and the GPS location of all the properties (bottom left panel) in the city of Kananga.

Value of nearby properties. We also train our machine learning algorithms using information about the market value of nearby properties. This information comes from the previously mentioned sample of 1,654 properties, whose market value was directly assessed by land surveyors during in-person appraisal visits. We compute the average property value assessed by land surveyors in each geographic stratum used in Balan et al. (2022), and in each neighbourhood of the city.¹⁶ Because property values can vary at a very local level, we go beyond neighbourhood information and compute the average property value assessed by land surveyors within a 200 m, 500 m, and 1,000 m radius, and the value of the three or five nearest properties assessed by land surveyors. The importance of these features in predicting property value (see Figure 5.1) implies that relying on finer geographic information to predict property values, such as machine learning or area or points-based methods, could considerably improve the predictiveness of property valuation methods. These different measures of the value of nearby properties are summarised in Panel C of Table 4.1.

¹⁵ The goal of the door-to-door property registration visits was to construct a complete property tax roll for the city of Kananga, as described in Balan et al. (2022) and Bergeron et al. (2023).

¹⁶ This approach is in line with simplified property valuation methods, such as points-based valuation, which often add or subtract points based on neighbourhood.

4.4 Missing values

Missing data is a recurrent issue in classical econometrics models (Tobin 1958). With missing values, non-parametric machine learning models perform better than parametric models. While we tried different solutions to account for missing values, such as k-nearest neighbours imputation, multiple correspondence analysis imputation, and out-of-range imputation, we report results that rely on the ‘missingness incorporated in attributes’ method (Twala et al. 2008), which has been shown to outperform other methods (Josse et al. 2019).

4.5 Property photographs

We rely on pictures of all the properties in Kananga to train our computer vision algorithms. Staff from the provincial government, accompanied by enumerators from our research team, took these pictures during the property registration that preceded the 2018 property tax campaign. Examples of property photographs are shown in Figure 4.5. Panel A shows the picture of a property built with non-durable material (e.g. mud brick), representing 89 per cent of properties in Kananga. Panel B shows the picture of a property built with durable material (e.g. cement), corresponding to the remaining 11 per cent of properties in Kananga.

Figure 4.5 Example of property photographs

A Low-value band property



Source: Authors own photo.

B High-value band property



Source: Authors own photo. Notes: This figure shows pictures of a property built with non-durable materials, such as mud bricks (Panel A), and a property built with durable materials, such as cement (Panel B).

5 Machine learning

In this section, we use machine learning methods trained on the market value of a sample of 1,654 properties (section 4.1), and property, geographic and nearby property features of all properties (sections 4.2-4.3), to obtain predictions of the market value of all properties in Kananga with the highest possible out-of-sample accuracy.

5.1 Methodology

Machine learning (or rather ‘supervised’ machine learning) revolves around the problem of **prediction**: we want to predict the value of a target variable y using a set of features x . For example, in our case, y is the market value of a property, and x is a vector of property characteristics. Machine learning aims to find functions of x that predict y well out-of-sample. Out-of-sample accuracy is a different objective than parameter estimation, which aims to find the relationship between y and x that maximises in-sample accuracy. To maximise out-of-sample accuracy, a machine learning algorithm will take a loss function $L(\hat{y}, y)$ as input, and look for a function \hat{f} that has low expected prediction loss $E_{x,y}L(\hat{y}, y)$ on a new data point from the same distribution.

Machine learning is appealing for its high dimensionality (flexible functional forms can fit varied data structures). But high dimensionality also means many possibilities. To find functions that work well out-of-sample, machine learning requires **regularisation** (sometimes called ‘hyperparameter optimisation’, ‘hyperparameter tuning’, or ‘tuning the algorithm’).¹⁷ This is typically done by splitting the data into a training set and a test set. The training set is used to train the algorithms’ parameters, and the test set is used to assess the out-of-sample performance of the trained algorithm.

A k-fold cross-validation approach is often used with small datasets (Stone 1974; Geisser 1975; Schaffer 1993). With k-fold cross-validation, the data is first partitioned into k equally-sized subsamples or folds. Subsequently k iterations of training and validation are performed, so that within each iteration a different fold of the data is held out for validation, while the remaining k-1 folds are used for learning. When selecting a model, k-fold cross-validation aims to reduce the amount of bias caused by a particular choice of validation set (Refaeilzadeh et al. 2009). Given that our sample of property value is small, we use k-fold cross-validation to tune our machine learning algorithms, and compare the out-of-sample performance of different machine learning algorithms.

5.2 Algorithms

Each machine learning model has well-known advantages and drawbacks (Hastie et al. 2001). However, the advantage of using machine learning is that it allows us to compare the performance of different models by assessing their out-of-sample accuracy. We trained the following type of algorithms, and tested their out-of-sample accuracy using 10-fold cross-validation:

1. Penalised linear models (LASSO, Ridge, elastic net). Penalised linear models, such as LASSO (Tibshirani 1996), Ridge (Hoerl and Kennard 1970), and elastic net (Zou and Hastie 2005), are widely used by econometricians. They rely on constructing a linear model that is penalised for having too many variables in the model. These models are also known as shrinkage or regularisation methods.
2. Kernel models (Support Vector Machine (SVM) and Support Vector Regression (SVR)). SVM, and its regression equivalent, SVR, usually perform well on small datasets due to their non-parametric nature and the flexibility of kernel functions (Bierens 1987). A kernel is a feature map of the input data to a higher dimension space. While data may not be

¹⁷ Regularisation is a complex procedure that requires relying on existing research with each algorithm, but also on experience and intuition - especially for complex models, such as boosting, which typically have more than 70 parameters to tune.

linear on the original input space, moving to a higher dimension space may help find a linear line with best fit. With SVR, the linear regression function is fit into the kernel space, and often turns out to be a non-linear function in the original input space. We test the most commonly used kernels: Linear and Radial Basis Function (RBF).

3. Regression trees and forests. Regression trees (Breiman et al. 1984), and their extension, random forests (Breiman 2001), are effective at flexibly estimating regression functions in settings where out-of-sample predictive power is essential. They are considered to have good out-of-the-box performance without requiring much regularisation.
4. Boosting. Boosting is a general-purpose technique that aims to improve the performance of simple supervised learning methods. In the context of tree-based models, boosting works as tree ensembles that are grown sequentially, with a new tree fitted on residuals of the previous model. Trees are not fully grown, and are considered 'weak learners'. The combination of multiple rounds of sequential weak learners has been shown to deliver a 'strong learner', characterised by high predictive performance (Schapire and Freund 2012).
5. Ensemble modelling. Another key feature of the machine learning literature is that it is possible to use model averaging and ensemble methods (Dietterich 2000). In many cases, a single model or algorithm does not perform as well as a combination of different models, averaged using weights obtained by optimising out-of-sample performance. We use a combination of boosting models with loss functions that depend on the property type.

5.3 Performance evaluation metric

Panel A of Table 5.1 assesses the out-of-sample accuracy of each machine learning model using the usual statistics from the property valuation literature (MAE, MAPE and percentage of the predicted values that fall within 20 per cent or 50 per cent of the land surveyor value) computed using 10-fold cross-validation. Column 1 reports the Mean Absolute Error (MAE), the average absolute difference between the target and the predicted value. It is a standard evaluation metric for regression models, and has the advantage of penalising large errors and being robust to outliers. Column 2 shows the Mean Absolute Percentage Error (MAPE), defined as the average absolute difference between the target and the predicted value, expressed in percentage of the actual value. The MAPE is also commonly used as an evaluation metric for regression models, due to its scale-independency and interpretability. However, its main drawback is that it can produce infinite or undefined values when the predicted value is close to zero. Columns 3-4 report results when using the share of predictions within a 20 per cent or 50 per cent band of the target value as additional performance evaluation metrics commonly used in the literature that aims to predict property value using machine learning (Bachofer et al. 2020).

Table 5.1 Performance of machine learning and computer vision models

Model	MAE	MAPE	Within 20%	Within 50%
	(1)	(2)	(3)	(4)
Panel A: Machine learning algorithms				
Linear regression	2,687.95	241.33%	11.30%	26.96%
Elastic net	2,871.15	265.33%	10.87%	27.20%
SVR - Linear kernel	2,687.95	241.33%	11.30%	26.96%
SVR - RBF kernel	2,567.45	154.49%	6.40%	21.86%
Random forest	2,259.19	154.31%	17.83%	41.30%
Boosting - MAPE loss	2,227.29	55.95%	17.64%	48.88%
Boosting - MAE loss	1,983.13	116.13%	18.88%	43.23%
Ensemble modelling	1,912.23	69.57%	22.11%	53.54%
Panel B: Computer vision algorithms				
Imagenet pre-training	6,062.46	678.49%	8.29%	15.12%
Imagenet & pre-training using Glaeser et al. (2018)	5,804.09	1,007.03%	7.32%	16.10%
Deep feature extraction	2,733.03	98.50%	8.78%	24.88%

Source: Authors own from collected data. Notes: This table assesses the out-of-sample accuracy of each machine learning model used to predict property values in Kananga. In Panel A, we examine the following machine learning algorithms: penalised linear model (LASSO, Ridge and elastic net), kernel models (SVR), regression trees and forests (random forest), and boosting models. In Panel B, we examine computer vision methods that rely on deep convolutional neural networks pre-trained on Imagenet. Column 1 reports the cross-validated mean absolute error (MAE), the average absolute difference between the target and predicted values. Column 2 reports the cross-validated absolute percentage error (MAPE), the average absolute difference between the target and the predicted value, expressed in percentage of the actual target value. Columns 3 and 4 report the cross-validated share of predictions that are within 20 per cent or 50 per cent of the target value, respectively. All the performance evaluations reported in Columns 1–4 rely on 10-fold cross-validation.

5.4 Results

For each machine learning model, we present the results of these performance evaluation metrics using 10-fold cross-validation in Panel A of Table 5.1. The first two rows present the results for penalised linear models: linear regression (row 1) and elastic net (row 2). The next two rows present the results for kernel models: SVR with a linear kernel (row 3) or an RBF kernel (row 4).¹⁸ Row 5 presents the results for a random forest model, and the next two rows show the results when using a boosted tree model with a MAE loss function (row 6) or a MAPE loss function (row 7). The last row shows the result for an ensemble model that uses a boosted-tree model with a MAPE loss for low-value properties (below US\$1,000), and a MAE loss for high-value properties (above US\$1,000).¹⁹ We test many combinations of hyperparameters for each type of algorithm, and only present the model with the highest out-of-sample performance when using 10-fold cross-validation.

We find that boosted tree models outperform penalised linear models, kernel models and tree models. This is in line with recent studies that have found that boosting models tend to perform better than other machine learning algorithms in a wide range of settings (Schapire and Freund 2012). The model that performs best is an ensemble modelling method that uses a boosted tree model with a MAPE loss function for low-value properties

¹⁸ We report results using two different kernels because the choice of kernel typically affects the performance of the SVR algorithm.

¹⁹ To differentiate between low- and high-value properties, we fit a random forest classifier that predicts whether a house is worth more than US\$1,000 or not.

(below US\$1,000), and a MAE loss function for high-value properties (above US\$1,000).²⁰ As shown in Panel A of Table 5.1, this ensemble modelling approach achieves a 10-fold cross-validated MAE of US\$1,912, a MAPE of 70 per cent and an R^2 of 60 per cent.^{21,22} Additionally, 22 per cent (respectively 54 per cent) of its predicted property values are within 20 per cent (respectively 50 per cent) of the target value. We will refer to the ensemble modelling approach as our preferred machine learning model in the rest of the paper.

While some studies argue in favour of simple machine learning models instead of complex ones, especially in contexts where properties in the training sample are relatively homogenous (Zurada et al. 2011), it is interesting to highlight that in our context the more complex machine learning algorithms (e.g. random forest or boosting) outperform simpler algorithms (e.g. linear regression and elastic net), plausibly due to the heterogeneity of property quality, disrepair and overall value in Kananga.

5.5 Feature importance

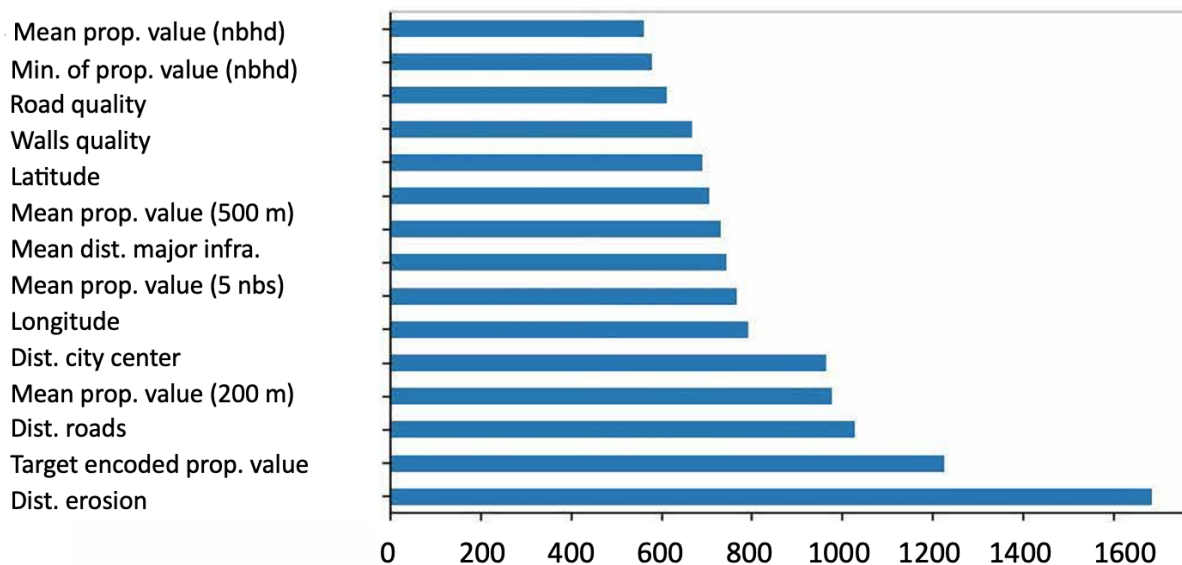
While machine learning models' predictive performance typically comes at the cost of explainability, we can describe how our preferred machine learning model based its prediction by looking at the features used most often for prediction. Figure 5.1 reports the importance score by attribute for the 15 attributes with the highest importance score. A feature's importance score measures the number of times a feature is used to split the data across all trees. We find that the value of neighbouring properties (average value of the 3 or 5 nearest properties, average value of properties within 200 m or 500 m, average, minimum and maximum value of properties in the neighbourhood), the location of the property (distance to erosion, distance to nearest road, to city centre or any major infrastructure), and the characteristics of the property (quality of the walls, roof and nearby road) are essential for the construction of the boosted decision trees within the model.

²⁰ Our results also show that the performance of boosting algorithms is greatly affected by the choice of loss function. While the penalty applied to a US\$200 error is constant across property value for a MAE loss function, with a MAPE loss function it is high for low-value properties and low for high-value properties. As a result, the prediction procedure with a MAPE loss function will push predictions downwards for low-value properties and upwards for high-value properties.

²¹ Table 5.1 reports the 10-fold cross-validated MAE (Column 1) and MAPE (Column 2). It also reports the share of predicted properties within 20% of the target value (Column 3) and within 50% of the target value (Column 4). It does not report R^2 , which is a common performance evaluation metric in machine learning, but less frequently used than MAE or MAPE when predicting property value. We find an R^2 of 60% for the ensemble modelling approach.

²² It is important to distinguish between evaluation metric and loss function. The evaluation metric is the function used to estimate the out-of-sample performance of the trained algorithm. The loss function is the function used by the algorithm to decide tree splits (in the case of random forest or tree-based boosting). Often, choosing a loss function similar or close to the evaluation metric will lead to the best results, but this may not always be the case. In our sample, for example, properties with a value below US\$1,000 are predicted more accurately when a MAPE loss function is used according to the MAE evaluation metric. This is due to the way the penalty is applied using the loss function, hence choosing the right loss is critical.

Figure 5.1 Feature importance – number of times a feature is used to split the data across all trees



Source: Authors own from collected data. Notes: This figure shows the importance of each feature for our preferred machine learning model, which is a combination of boosting models. It displays the number of times each feature is used to split the data across all trees. We only report the 15 most important features for this measure of feature importance. nbhd = neighbourhood; nbs = neighbours.

6 Computer vision

In this section, we use computer vision methods trained on the market value of a sample of 1,654 properties (section 4.1), and property features extracted from pictures of all properties in Kananga (section 4.5), to obtain predictions of the market value of all properties in Kananga with the highest possible out-of-sample accuracy.

6.1 Convolutional neural networks

Computer vision is a sub-field of machine learning that seeks to develop techniques to help computers understand the content of digital images, like photographs. Images are a complex form of data for an algorithm to handle – they have large dimensionality and are challenging to interpret. A human eye sees forms and context, but a computer processes a coloured image as a superposition of matrices – one for each primary colour (red, green and blue) – with each element being the strength of the colour.

Most computer vision algorithms first create filters based on the images – a vast literature on signal processing discusses the extensive range of existing filters (Oliva and Torralba 2001; Dalal and Triggs 2005) – to maximise the signal to extract from each image. A computer vision algorithm then learns the important features from these filters in the training sample, and predicts the label in the test sample.

Until recently, Support Vector Machine or SVM (Vapnik 1998; Scholkopf and Smola 2001), was considered the state-of-the-art methodology when analysing high-dimensional images. However, deep learning methods have been shown to outperform SVM methods for computer vision problems (Voulodimos et al. 2018). In particular, Convolutional Neural Networks algorithms or CNNs (LeCun et al. 1989, 2010) can learn the essential filters to

process the image before predicting the label based on the image. As a result, they tend to outperform SVM algorithms for most computer vision problems.

6.2 Convolutional neural networks with a small training sample

CNNs typically have more than 1 million hyperparameters, and require large-scale training data. Our dataset is small, with 1,654 property photographs in the training sample. While directly training a deep CNN on such a small training sample is not feasible, several alternative approaches can be used:

1. Shallow CNN. In principle, we could train a shallow CNN model with a small number of layers. The network would be trained solely on our data, and would be tailored to the task of predicting property value. However, the strength of CNNs resides in their ability to extract information from a large number of layers, and shallow CNNs perform poorly in our context due to the small number of property photographs.²³
2. Transfer learning on a deep CNN. Another solution is to use transfer learning, which consists of selecting a deep CNN that has already been trained on a similar task, and fine-tuning the hyperparameters model using our data (Zhuang et al. 2021). We select the VGG16 (Simoyan and Zisserman 2015) CNN, which loads weights pre-trained on ImageNet (Deng et al. 2009). We implement a version that fine-tunes the last layer with our property photographs. We also implement a version that fine-tunes the 16 top layers using large-scale property prices and pictures data for the city of Boston from Glaeser et al. (2018), and then fine-tunes the last layer with our property photographs.
3. Feature extraction using a deep CNN. Another solution for small training samples is to use a CNN as a ‘feature extractor’ (Dara and Tumma 2018). With CNNs, the last layer of the regression network takes input from many neurons of the previous layer and outputs the target value. Instead, we can use the features created by the neural network from the previous layers, de-noise these features using a principal component analysis, and use them in a standard SVR algorithm. This approach leverages that CNNs are good at extracting features from images, which can be captured at a high-level layer of the network and used in another model to predict the target variable of interest.

6.3 Image pre-processing

Before getting pictures through the networks, it is necessary to pre-process them. First, we resize the picture, so they have the same format as the network input size.²⁴ When an image is rectangular instead of square, we reformat it to be square-shaped along the smallest image axis. We also normalise the images to fit the distribution of images on which the network is trained. Image data augmentation is performed by introducing a random horizontal flip of the picture as it goes through the network.²⁵

6.4 Results

We present the results for the CNN algorithms in Panel B of Table 5.1. Algorithms that use deep CNNs for feature extraction outperform algorithms that leverage transfer learning on a

²³ For example, when we train a shallow CNN using Generative Adversarial Network (GAN) unsupervised pre-training, about half the predicted property values are negative, and the 10-fold cross-validated MAE is US\$15,000.

²⁴ For example 224x224 for the VGG16 network model.

²⁵ A horizontal flip does not have much effect to the human eye. However, it changes the disposition of artifacts on the picture, which helps the computer learn better.

deep CNN. As shown in Panel B of Table 5.1, CNNs that use transfer learning achieve at best a 10-fold cross-validated MAE of US\$5,804, a MAPE of 678 per cent. In contrast, the algorithm that relies on a deep CNN for feature extraction achieves a MAE of US\$2,733, a MAPE of 98.50 per cent. As a consequence, we refer to the later approach as our preferred computer vision model in the rest of the paper.

Overall, computer vision algorithms perform less well than machine learning models in our context. Our preferred machine learning model achieves a MAE of US\$1,912 and a MAPE of 70 per cent, which is lower than the MAE of US\$2,733 and MAPE of 98.50 per cent obtained by our preferred computer vision model. The lower performance of computer vision algorithms could be explained by two factors. First, the training sample might be too small to make the most out of computer vision algorithms. High performing networks are typically trained on millions (Deng et al. 2009), or at least tens of thousands, of images (Naik et al. 2014; Glaeser et al. 2016; Naik et al. 2017; Glaeser et al. 2018). Second, property photographs in Kananga have a lower resolution, and it is possible that less information can be extracted from these pictures than from Google Street View photographs, which are typically used to predict property value (Glaeser et al. 2018; Law et al. 2019), neighbourhood safety (Naik et al. 2014), neighbourhood income (Glaeser et al. 2016), or changes in neighbourhood appearance (Naik et al. 2017) in high-income countries.²⁶

Overall, we interpret the results as suggesting that more complex computer vision methods trained on features extracted from property pictures underperform relative to simpler machine learning methods trained on property characteristics from administrative data, especially when the training sample is small, and little information can be extracted from the property photographs.

7 Comparison with other studies

Machine learning and computer vision algorithms perform less well in our context than in high-income settings. For example, Bency et al. (2017) predict house prices in London, Birmingham and Liverpool using geospatial data combined with data extracted from satellite imagery and a large house price dataset. The CNN algorithm with the highest performance achieves an R^2 of 90 per cent. By contrast, our preferred machine learning model achieves a much lower R^2 of around 60 per cent in our setting. This is hardly surprising, since studies that use machine learning methods to predict property value in high-income settings typically have access to more extensive training datasets and richer features (Chopra et al. 2007; Bency et al. 2017; Glaeser et al. 2018; Law et al. 2019).

However, our results are comparable to studies that estimate property value in low-income settings. Ali et al. (2018) combine high-resolution satellite imagery with information on sales prices, targeted surveys and routine statistical data to predict property values in Rwanda, and find an R^2 of around 56 per cent when using a simple hedonic regression of property value on property characteristics using Ordinary Least Squares (OLS).²⁷ Similarly, Franklin

²⁶ The pictures of properties taken in Kananga have a lower resolution, they are typically not centred on the property, and might contain trees, objects or people. Additionally, it is mechanically harder to infer property value from pictures in low-income contexts. Because properties have lower values, fewer characteristics can be learned from property photographs.

²⁷ Property characteristics include district, area in m^2 by district, and volume of buildings in m^3 by district.

(2019) uses hedonic regressions of observed rent on apartment characteristics in Addis Ababa, Ethiopia, and finds a higher R^2 of around 85 per cent.²⁸

More specifically, our results are similar to other studies that use machine learning to estimate property values in low-income settings. Bachofer et al. (2020) use maximum relevance and minimum redundancy model (Peng et al. 2005) to predict property values per m^2 in Kigali, Rwanda. They train their algorithms using several sources of information, including geospatial data, remote sensing data and house price data. Their best algorithm results in a cross-validated R^2 of 45.8 per cent, and 24.5 per cent of the predicted values per m^2 are within 20 per cent of the target value. Knebelmann et al. (2023) use elastic net regression models trained on a sample of 4,448 property market values provided by assessors, combined with information on property sections, total built area and number of floors, to predict the annual rental value of properties in Dakar, Senegal. The best performing algorithm results in a cross-validated R^2 of 87 per cent, and 54.2 per cent of the predicted values per m^2 are within 30 per cent of the target value.²⁹

Despite our sample of property values being much smaller – 1,654 property values vs. 7,445 in Bachofer et al. (2020) and 4,448 in Knebelmann et al. (2023), and our label being harder to predict – property value in US\$ vs. log price per m^2 in US\$ in Bachofer et al. (2020) and log annual rental value in Knebelmann et al. (2023), we find results that are similar to Bachofer et al. (2020) and slightly underperform Knebelmann et al. (2023), with an R^2 of 60 per cent and 22.1 per cent of the predicted values that are within 20 per cent of the target value (Table 5.1). We interpret these results as suggesting that machine learning methods can be successfully used to construct a property tax roll, even in contexts where the government has access to a small number of assessor valuations and limited information about property characteristics.³⁰

8 Conclusion

Developing countries often lack the financial resources to provide public goods. Property taxation has been identified as a promising source of local revenue, because it is relatively efficient, captures growth in real estate values, and can be progressive. However, many governments of low-income countries do not collect property taxes effectively due to absent or incomplete property tax rolls.

We propose to use machine learning and computer vision to remedy this issue, and, as an illustration, construct a property tax roll for the city of Kananga in DRC, where information about properties is limited, and the government can only collect a small number of property values through direct assessor valuation. We rely on a training sample of 1,654 property values estimated by government land surveyors during in-person property appraisal visits.

²⁸ Apartment characteristics include indicators for number of bedrooms (studio, 1,2,3+), site fixed effects, block has shop space, block has communal slaughter area, distance from roads, basic quality of finishing.

²⁹ The cross-validated R^2 increases to 90%, and 59.6% of the predicted values per m^2 are within 30% of the target value, when adding a large vector of property characteristics visible from the outside, such as usage (residential, commercial or mixed), type of fence, state of the fence, type of cladding, state of the cladding, cement wall, presence of decorative tiles, quality of doors and windows, landscape improvement, architectural improvement, presence and type of garage, balcony, location with respect to main road, type of road, presence of sidewalk, whether the property is at an angle, and presence of street lights.

³⁰ We only have values for about 3.7% of properties in the city (1,654 out of 45,162 properties), which is less than the 5% of properties threshold recommended by Manwaring and Regan (2019) for calibration to be effective.

We combine this data with property characteristics from an administrative property census and features extracted from property photographs, and use machine learning and computer vision algorithms to predict the value of all properties in the city.

Our best machine learning algorithm results in a cross-validated R^2 of 60 per cent and 22.1 per cent of the predicted values within 20 per cent of the target value. Computer vision methods trained using features extracted from property photographs perform less well, with only 9 per cent of the predicted values within 20 per cent of the target value. We interpret the results as suggesting that simple machine learning methods trained on property characteristics from administrative data can outperform more complex computer vision methods trained on features extracted from property pictures, especially when the training sample is small and little information can be extracted from the property photographs.

While our best machine learning algorithm performs less well than machine learning algorithms trained on larger training datasets and richer features in high-income settings (Chopra et al. 2007; Bency et al. 2017; Glaeser et al. 2018; Law et al. 2019), it compares favourably to other studies that have used machine learning to predict property values in other low-income settings, such as the city of Kigali in Rwanda (Bachofer et al. 2020) and Dakar in Senegal (Knebelmann et al. 2023).

Overall, we interpret our results as suggesting that machine learning methods can be successfully used to construct property tax rolls in contexts where information about properties is limited, and where the government has limited capacity and resources to measure property values using assessor valuation. While machine learning methods are typically less costly and offer more accurate predictions than other property valuation methods, the added benefit of these methods in terms of accuracy needs to be weighed against the complexity and opacity of these models for taxpayer understanding (Manwaring and Regan 2019). In particular, in contexts where raising tax morale and compliance is important, policymakers should properly trade off accuracy vs. having a tax system that citizens fully understand. We view the effect of different property valuation methods on tax morale, tax compliance and revenue, and tax incidence, as fruitful lines of future inquiry.

References

- Ali, D., Deininger, K. and Wild, M. (2018) *Using Satellite Imagery to Revolutionize Creation of Tax Maps and Local Revenue Collection*, World Bank Policy Research Working Paper 8437, Development Research Group, World Bank, <https://documents1.worldbank.org/curated/en/347231526042692012/pdf/WPS8437.pdf>
- Athey, S. and Imbens, G. (2019) 'Machine Learning Methods Economists Should Know About', *Annual Review of Economics* 11: 685-725 Working Paper
- Balan, P., Bergeron, A., Tourek, G. and Weigel, J. (2022) 'Local Elites as Tax Collectors: Experimental Evidence from the D. R. Congo', *American Economic Review* 112(3): 762-97
- Bency, A., Rallapali, S., Ganti, R., Srivatsa, M. and Manjunath, B. (2017) *Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery*, Working Paper, https://vision.ece.ucsb.edu/sites/default/files/publications/bency_wacv_17.pdf
- Bergeron, A., Tourek, G. and Weigel, J. (2023) *The State Capacity Ceiling on Tax Rates: – Evidence from Randomized Tax Abatements in the DRC*, Working Paper
- Besley, T. , Ilzetki, E. and Persson, T. (2013) 'Weak States and Steady States: The Dynamics of Fiscal Capacity', *American Economic Journal: Macroeconomics* 5(4): 205-235
- and Persson, T. (2013) 'Taxation and development', *Handbook of Public Economics* 5: 51-110
- (2009) 'The Origins of State Capacity: Property Rights, Taxation and Politics', *American Economic Review* 99(4): 1218-1244
- Bierens, H. (1987) 'Kernel Estimators of Regression Functions', *Advances in Econometrics: Fifth World Congress*, 1: 99-144
- Blumenstock, J. (2016) 'Fighting Poverty with Data,' *Science* 353(6301): 753-54
- Cadamuro, G. and On, R. (2015) 'Predicting Poverty and Wealth from Mobile Phone Metadata', *Science* 350(6264): 1073-76
- Breiman, L. (2001) 'Random Forests', *Machine Learning* 45(1): 5-32
- Friedman, J., Stone, C. and Olshen, R. (1984) *Classification and Regression Trees*, CRC Press
- Brimble, P., McSharry, P., Bachofer, F., Bower, J. and Braun, A. (2020) 'Using machine learning and remote sensing to value property in Rwanda,' *IGC Working Paper, C-38315-RWA-1*
- Brockmeyer, A., Estefan, A., Serrato, J. and Ramirez, K. (2023) *Taxing Property in Developing Countries: Theory and Evidence from Mexico*, NBER Working Paper 28637, National Bureau of Economic Research, <https://www.nber.org/papers/w28637>

- Casaburi, L. and Troiano, U. (2016) 'Ghost-House Busters: The Electoral Response to a Large Anti Tax Evasion Program', *Quarterly Journal of Economics* 131(1): 273-314
- Chopra, S., Thampy, T., Leahy, J., Caplin A., and LeCun, Y. (2007) 'Discovering the Hidden Structure of House Prices with a Non-Parametric Latent Manifold Model', *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*
- Collier, P. (2017) *Land and Property Taxes for Municipal Finances*, International Growth Center Working Paper 07/18, <https://www.theigc.org/sites/default/files/2017/08/Land-and-Property-Taxes-for-Municipal-Finance-06.07.18.pdf>
- Dalal, N. and Triggs, B. (2005) 'Histograms of Oriented Gradients for Human Detection', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005(1): 886-893
- Dara, S. and Tumma, P. (2018) *Feature Extraction By Using Deep Learning: A Survey*, 2018 Second International Conference of Electronics, Communication and Aerospace Technology (ICECA), <https://ieeexplore.ieee.org/xpl/conhome/8466240/proceeding>
- Davis, P., McCluskey, W., Grissom, T. and McCord, M. (2012) 'An empirical analysis of simplified valuation approaches for residential property tax purposes', *Property Management* 30(3): 232-254
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009) *ImageNet: A large-Scale hierarchical image database*, 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), <https://ieeexplore.ieee.org/document/5206848>
- Dietterich, T. (2000) 'Ensemble methods in machine learning', *International workshop on multiple classifier systems*, pp. 1-15, https://link.springer.com/chapter/10.1007/3-540-45014-9_1
- Dincecco, M. and Katz, G. (2016) 'State Capacity and Long-run Economic Performance', *The Economic Journal* 126(590): 189-218
- Fish, P. (2018) 'Practical guidance note: training manual for implementing property tax reform with a points-based valuation, African Tax Administration Paper 2, Brighton: International Centre for Tax and Development, <https://www.ictd.ac/publication/practical-guidance-note-training-manual-for-implementing-property-tax-reform-with-a-points-based-valuation/>
- Fjeldstad, O-H., Ali, M. and Goodfellow, T. (2017) *Taxing the urban boom: property taxation in Africa*, CMI Insight No. 1, Chr. Michelsen Institute, <https://www.ictd.ac/publication/practical-guidance-note-training-manual-for-implementing-property-tax-reform-with-a-points-based-valuation/>
- Franklin, S. (2019) *The demand for government housing: Evidence from lotteries for 200,000 homes in Ethiopia*, Working Paper, https://www.tse-fr.eu/sites/default/files/TSE/documents/sem2019/Jobmarket2019/franklin_jmp.pdf
- Geisser, S. (1975) 'The predictive sample reuse method with applications', *Journal of the American statistical Association* 70(350): 320-328

- Glaeser, E., Kincaid, M. and Naik, N. (2018) Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks, NBER Working Paper 25174, https://www.nber.org/system/files/working_papers/w25174/w25174.pdf
- Kominers, S., Luca, M. and Naik, N. (2016) 'Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life', *Economic Inquiry* 56(1): 114-136
- Hastie, T., Tibshirani, R. and Friedman, J. (2001) *Element of statistical learning*, Massachusetts Institute of Technology
- Hoerl, A. and Kennard, R. (1970) 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics* 12(1): 55-67
- Jean, N., Burke, M., Xie, M., Davis, M., Lobell, D. and Ermon, S. (2016) 'Combining Satellite Imagery and Machine Learning to Predict Poverty,' *Science* 353(6301): 790-94
- Jibao, S. (2017) 'Sierra Leone', in R. Franzsen and W. McCluskey, *Property Tax in Africa: Status, Challenges and Prospects*, Cambridge MA: Lincoln Institute of Land Policy
- and Prichard, W. (2016) 'Rebuilding local government finances after conflict: Lessons from a property tax reform programme in post-conflict Sierra Leone', *The Journal of Development Studies* 52(12): 1759-1775
- ——— (2015) 'The political economy of property tax in Africa: Explaining reform outcomes in Sierra Leone', *African Affairs* 114(456): 404-431
- Josse, J., Prost, N., Scornet, E. and Varoquaux, G. (2019) 'On the consistency of supervised learning with missing values,' *arXiv:1902.06931*
- Kaldor, N. (1965) 'The Role of Taxation in Economic Development', in E. Robinson (ed), *Problems in Economic Development*, Springer
- Khan, A., Khwaja, A. and Olken, B. (2015) 'Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors', *The Quarterly Journal of Economics* 131(1): 219-271
- Knebelmann, J., Pouliquen, V. and Sarr, B. (2023) *Bureaucrat Discretion versus Algorithms: Implications for Property Tax Equity in Senegal*, Working Paper
- Lall, S., Henderson, V. and Venables, A. (2017) *Africa's Cities: Opening Doors to the World*, Washington DC: World Bank
- Law, S., Paige, B. and Russell, C. (2019) 'Take a Look Around: Using Street View and Satellite Images to Estimate Houses Prices', *ACM Transactions on Intelligent Systems and Technology* 10(5): 1-19
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1989) 'Backpropagation Applied to Handwritten Zip Code Recognition', *Neural Computation* 1: 541-551
- Farabet, C. and Kavukcuoglu, K. (2010) 'Convolutional Networks and Applications in Vision', *IEEE International Symposium on Circuits and Systems*, <https://ieeexplore.ieee.org/document/5537907>

- Lim, L., McCluskey, W. and Davis, P. (2008) 'Area-Based Banding for Property Tax Assessment in Transitional Countries: An Empirical Investigation', *Journal of Real Estate Literature* 16(2): 201-215
- Manwaring, P. and Regan, T. (2019) *Enhancing property tax in Kampala: Successes, challenges, and next steps for increasing municipal revenue*, International Growth Centre Policy Brief, <https://www.theigc.org/publications/enhancing-property-tax-kampala-successes-challenges-and-next-steps-increasing>
- Mayshar, J., Moav, O. and Pascali, L. (2022) 'The Origin of the State: Land Productivity or Appropriability?', *Journal of Political Economy* 130 (4): 1091-1144
- McCluskey, W., Plimmer, F. and Connellan, O. (2002) 'Property tax banding: a solution for developing countries', *Assessment Journal* 9(2): 37-47
- Mullainathan, S. and Spiess, J. (2017) 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspectives* 31(2): 87-106
- Naik, N., Kominers, S., Raskar, R., Glaeser, E. and Hidalgo, C. (2017) 'Computer Vision Uncovers Predictors of Physical Urban Change', *PNAS* 114 (29): 7571-7576
- Philipoom, J., Raskar, R. and César, H. (2014) 'Streetscore - Predicting the Perceived Safety of One Million Streetscapes', *IEEE Computer Vision and Pattern Recognition Workshops*
- Oliva, A. and Torralba, A. (2001) 'Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope', *International Journal of Computer Vision*, 42: 145-175
- Peng, H., Long, F. and Ding, C. (2005) 'Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8): 1226-1238
- Pomeranz, D. and Vila-Belda, J. (2019) 'Taking State-Capacity Research to the Field: Insights from Collaborations with Tax Authorities', *Annual Review of Economics* 11: 755-781
- Prichard, W. (2015) *Taxation, responsiveness and accountability in Sub-Saharan Africa: the dynamics of tax bargaining*, Cambridge University Press
- Refaeilzadeh, P., Tang, L. and Liu, H. (2009) 'Cross-validation', in L. Liu and M. Ozsu (eds), *Encyclopedia of database systems*, Springer
- Robinson, J. (2022) *Tax Aversion and the Social Contract in Africa*, NBER Working Paper 29924, <https://www.nber.org/papers/w29924>
- Rosengard, J. (1998) *Property Tax Reform in Developing Countries*, Kluwer Academic Publishers
- Schaffer, C. (1993) 'Selecting a classification method by cross-validation', *Machine Learning*, 13 (1): 135-143
- Schapire, R. and Freund, Y. (2012) *Boosting: Foundations and Algorithms*, Springer

- Scholkopf, B. and Smola, A. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press
- Simoyan, K. and Zisserman, A. (2015) 'Very Deep Convolutional Networks for Large-Scale Image Recognition', <https://doi.org/10.48550/arXiv.1409.1556>
- Stone, M. (1974) 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2): 111-133
- Tibshirani, R. (1996) 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society: Series B*, 58(1): 267-288
- Tobin, J. (1958) 'Estimation of Relationships for Limited Dependent Variables', *Econometrica* 26(1): 24-36
- Twala, B., Jones, M. and Hand, D. (2008) 'Good Methods for Coping with Missing Data in Decision Trees', *Pattern Recognition Letters* 29(7): 950-956
- Vapnik, V. (1998) *Statistical learning theory*, vol. 1, New York: Wiley
- Voulodimos, A., Doulamis, N., Doulamis, A. and Protopapadakis, E. (2018) 'Deep Learning for Computer Vision: A Brief Review', *Computational Intelligence and Neuroscience* 2018: 7068349
- Weigel, J. (2020) 'The Participation Dividend of Taxation: How Citizens in Congo Engage More with the State when it Tries to Tax Them', *Quarterly Journal of Economics*, 135(4): 1849-1903
- Zebong, N., Fish, P. and Prichard, W. (2017) *Valuations for Property Tax Purposes*, International Centre for Tax and Development Summary Brief Number 10, <https://www.ictd.ac/publication/valuation-property-tax/>
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2021) 'A Comprehensive Survey on Transfer Learning', *Proceedings of the IEEE* 109(1)
- Zou, H. and Hastie, T. (2005) 'Regularization and Variable Selection via the Elastic Net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301-320
- Zurada, J., Levitan, A. and Guan, J. (2011) 'A comparison of regression and artificial intelligence methods in a mass appraisal context', *Journal of real estate research* 33(3): 349-388



www.ictd.ac