NYU Center for Urban Science + Progress

Capstone Project Report
In partial fullfilment of the requirements for the
Master of Science in
Applied Urban Science & Informatics

Property Valuation and Tax Mapping from Imagery Data

CUSP capstone manager Federica Bianco [1] , Zhiao Zou [2] , Nina Nurrahmawati [3] , Te Du [4] , Bailey [5] , ss9872 [6]

ABSTRACT
Abstract content goes he

2018 Capstones: Automated Feature Detection from Imagery Data to Promote Equity and Fairness in Assessed Values

Abstract Automated tax valuation models utilize individual building features to estimate a home's value and subsequent tax liability. The records that the New York City Department of Finance (DoF) has that document the features used for tax assessment for each house currently have no quality assurance checks apart from in-person inspections done by visits to each individual home. Desktop review of high resolution street level images is an effective replacement for on-site inspections, but still requires manual labor, which, multiplied over the more than 1 million parcels in the DoF's jurisdiction, poses a large drain on resources. This study aims to 1) establish a proof-of-concept of automating the desktop review process by utilizing a state of the art image recognition algorithm, 2) compare and contrast the performance costs of classifiers trained on images of varying degrees of quality, and 3) identify the proximity label for each single family home in NYC. Four convolutional neural nets were trained on labeled images queried from Google Street View. The first had only ground truth labels and screened images, the second had all images for which there was a ground truth, the third had all images and labels, including both noisy labels and ground truth labels , and the final classifier had noisy labels and screened images. Accuracy on a test set of images was the measure of performance. The classifier trained on the fewest, but most high quality images performed the best (screened images and ground truth labels). Future work is needed to optimize the performance of the classifiers and to automate the screening of the noisy label images.

1 INTRODUCTION

Equity and fairness are paramount when determining the property tax value of a home. Deciding the appropriate home value, however, is confounded by multiple factors: homes are bought and sold on long timescales and home values are affected by multiple temporal and geographic factors which can be difficult to explain. Automated tax valuation models have been used to determine the market values of homes in an objective and

fair manner. These models use real estate transactions to predict an accurate market value for a home based on the specific characteristics of the home as well as geographic features and market trends. The valuation models produce more accurate home value estimates than traditional appraisal methods (Kok, Koponen,& Martínez-Barbosa, 2017, Carbone & Longini, 1977, Schulz, Wersing, & Werwatz, 2013). An assumption of these models is that the input characteristics of a house are accurate and reliable. For real estate valuation models, these values come from actual transaction that have taken place, and the requisite appraisal process assures that the descriptive traits of the home are factual. For property tax purposes, the records of home features cannot be assumed to be accurate and up to date, however.

The New York City Department of Finance (DoF hereafter) oversees a jurisdiction with a market value over 1.258 trillion dollars and over one million parcels. At that size, it is costly and time-intensive for the department to do in-person inspections of every property to ensure that the descriptive data is accurate and up to date. The dynamic nature of the city means that properties need to be reviewed regularly to maintain an accurate record of its physical characteristics. For high value properties in the city, the return on time invested in an on-site inspection is well worth the effort. For one, two and three family dwellings, the demand on time and personnel is too high to warrant the modest adjustments that might be garnered from an in-person inspection. Scaled over the million buildings in the city, however, and small inaccuracies add up to significant sums of missing or misappropriated revenue. A solution is to assess home features via photographs in place of an in-person inspection. While a desk-top review allows one agent to inspect many more homes, the process is still time-consuming. Training a computer vision model to identify specific features in a photograph of a house is a promising solution that can improve the accuracy of the features on file for each property, and in turn the accuracy of the tax valuation estimates.

The goals of this study are threefold. First and foremost, the goal is to test the feasibility of an automated approach of home-feature screening. The DoF must make choices within a resource constrained context, and as of now, in-person home assessments are implausibily time-consuming and resource-intense. For a city with a much smaller portfolio, a desktop review might be a reasonable solution, but at over 1 million homes, doing so at the DoF would be an expensive undertaking. Automated home-feature detection is a scalable technique, but it is only worthwhile if it is effective. Therefore, we must prove the feasibility of this

approach by training an accurate home-feature image classifier. The second goal of this study is to compare the performance of classifiers with training images of varying quantities and qualities. While image classifiers perform best when trained on a large volume of high quality images, labelling those training images is another resource cost, and requires the desktop review that this approach is meant to replace. In theory, the DoF already has labels for every address, since they have a library of features for every house. Although some of those labels are wrong, we test the performance cost of using these "noisy" labels as training labels. Finally, our last goal is to provide the DoF with an updated dataset of addresses with features identified by the classifier.

2 LITERATURE REVIEW

Automated tax valuation models have been used in an attempt to determine the market values of homes in an objective and fair manner. These models, which use real estate transactions to predict an accurate market value for a home based on its features, produce more accurate home value estimates than traditional appraisal methods (Kok, Koponen,& Martínez-Barbosa, 2017, Carbone & Longini, 1977, Schulz, Wersing, & Werwatz, 2013). A record of 25 years of real estate transactions showed that commercial real-estate appraisal was systematically over- or under valuing properties (Cannon & Cole, 2011). There is a clear need for accurate and up-to-date property valuations, as Luts was able to show that tax revenue lagged three years behind changes in housing market (2008). While it has been shown that automating the process for assessing a fair market home value leads to a more equitable valuation, these models rely on home features as inputs into the model. Missing from the literature is an evaluation of the accuracy of the characteristics that are on record for a property.

Computer vision and image processing provide a promising move towards accurately assessing property characteristics. In one study, two-dimensional images were used to recognize building materials for the purpose of monitoring construction progress (Dimitrov & Golparvar-Fard, 2014). Satellite images were used successfully to identify key property features, such as parcel boundaries, structure type, roofing material, and age, to be used in tax valuation (Jain, 2008). You, Pang, Cao, and Luo, used images from a real-estate website to train a neural network to predict home prices with features such as geodesic distance and school information (2017). Kang, Körner, Wang, Taubenböck, and Zhu trained a convolutional neural network on Google Street View images to classify facade structures and identify land use classification (2006).

Apart from identifying building materials and land use types, the existing

literature body is lacking studies that specifically use computer vision to classify building characteristics. The previously mentioned studies differ from our objective in a significant way in that they are primarily focused on identifying the land use purpose or structure types, such as commercial or residential. Our methodology focuses on a more general approach in identifying some features on the buildings. Our short term goal is to classify the proximity for one, two, or three families housing.

To attack this type of image classification problems, some of the common image classification approaches could be considered. The approach of using histograms of gradient to detect humans from images is similarly applicable to our question in identifying a specific feature (Zhu, Yeh, Cheng, & Avidan, 2006). However, sometimes object detection or recognition algorithms are interested in isolating or counting the object (Lowe, 1982, Navneet Dalal and Bill Triggs, 2012). We are more interested in a binary classification problem whether the feature present or not. Some more traditional methods such as support vector machine based and random forest and ferns have proved to be effective image classifiers (Chapelle, Haffner, & Vapnik, 1999, Bosch, Zisserman, & Munoz, 2007). A more recent advance in convolutional deep neural networks have enormous impact on image classification and neural network is emerging as the state of the art method for image classification problems (Krizhevsky, Sutskever, & Hinton, 2012). In that research, however, the datasets include Caltech 256 or e Corel stock photo collection, MNIST handwriting benchmark, and ImageNet LSVRC-2010. None of those datasets include any images for houses or buildings characteristics. Our experiment will apply all these techniques to our datasets. Furthermore, our results will benchmark the effectiveness of these techniques in identifying building features.

3 DATA

3.1 Data sources

One, two and three family dwellings imagery data were needed to train the classifier to identify particular building features. A library of building images was acquired by scraping individual building images from Google Street View. Scraping is a practice to automatically extract information from any resource existing on a remote virtual location and commonly done through Application Programming Interface (API). Google Street View API allows the users to download 25,000 images per day with a resolution of 640 x 640 pixels.

The DoF currently has 13 building physical characteristics that need to be reviewed for tax valuation purposes. Building proximity was chosen as the target variable for the identifier feature.

Building proximity is the distance between each building, whether it is attached/abutted (i.e. both sides of the building are touching or sharing a wall with a neighboring building), semi-attached (i.e. only one side of the building is touching the side or sharing a wall with a neighboring building) or detached/freestanding building. This feature was selected because it is a categorical rather than real numbered, which makes it a good classification problem. Additionally, the classes are easily distinguishable and less subjective than other features, making it easier for the team, which lacks domain knowledge, to visually assess a building's class.

The DoF provided the results of their desktop review, a dataset containing 2,520 screened addresses that gave the base of valid ground truth label for training the classifiers. From both the PLUTO and the DoF datasets, a list of addresses of each type of building proximity was generated as the source for image scraping using Google Street View API. The PLUTO dataset and the DoF desktop review results provided us with two labeled datasets, each at opposite ends of the quality and quantity spectrum. The PLUTO dataset contains all the buildings in the NYC jurisdiction, but the proximity labels are of unknown quality, since there is a presumption that a portion of those labels are incorrect. The PLUTO dataset is the source of what will be referred to as the "noisy labels", hereafter. The results of the desktop review are only a few thousand, but all can be safely assumed to be completely accurate labels, and is the source of what will be referred to hereafter as the "ground truth labels". We will compare the effect of the quality and quantity on the performance of our classifiers by training half our classifiers with only the few images with ground truth labels and the other half with the much bigger set of images with noisy labels.

In a similar vein, the images scraped via Google Street View are also of varying quality. Google Street View returns an image of the house at the address being queried. In practice, not all of those images are usable. In some cases, the image returned was empty or of an irrelevant scene. At other times, the image was taken at an odd angle, and not all sides of the house were visible. In many instances, the view of the house was obscured by trees. The varying degree of quality of the images mirrored the question of quality of the labels. Just as manually labeling the images is time consuming, screening the images to remove the images that are unusable is too time-consuming of an undertaking. For this reason, we only manually screened the images for the ground truth labeled addresses. The remaining images we did not screen, and this has resulted in another dimension on which to test the performance of the classifiers. We will train four different classifiers each with a different set of training images at varying degrees of quality and at varying quantities (see table 1). While we know the cost of screening and labeling images

(time), training four different classifiers will reveal the penality on performance of not screening or labelling.

3.2 Exploratory Data Analysis

We were able to obtain 244,133 labeled data from throughout New York City. The current image data consist of 76,310 attached buildings, 78,615 semi-attached buildings and 89,208 detached buildings that are obtained from five different boroughs.

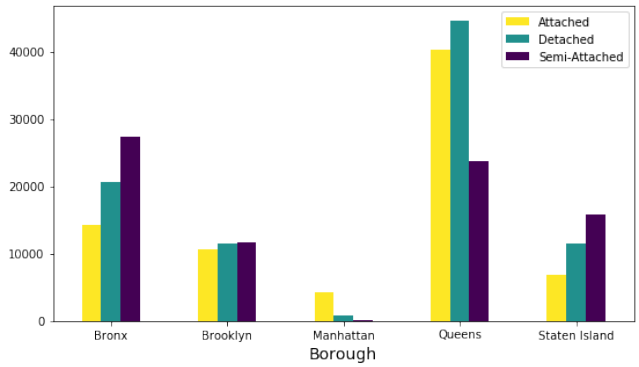

Figure 1. Borough distribution of image dataset

Proportional distribution of each proximity type in every borough is an ideal situation to be achieved for this project. However, one, two and three family housing is rare in Manhattan (Fig.1). Each of the three proximity types is represented by fewer than 5,000 images in comparison to those for Queens, which has more than 20,000 images for each of the proximity types. It is reasonable since most of the residential buildings in Manhattan are multi-family buildings, which is not included in the scope of this project.

The ground truth data has screening result of residential buildings in the Bronx

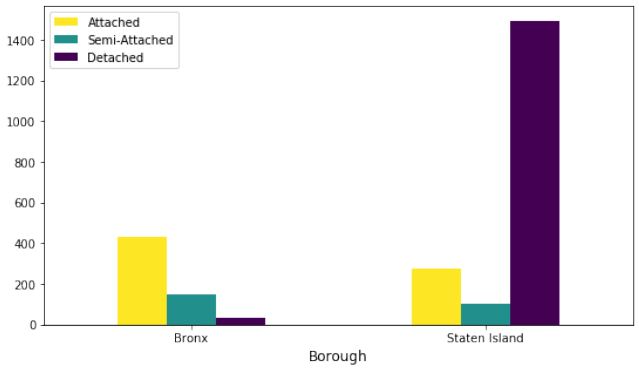

Figure 2. The proportion of ground truth data from NYC Department of Finance

and Staten Island with a disproportionate distribution of proximity classes. From 2,520 addresses, 75% are located in Staten Island. Each of the proximity types is not proportionally distributed, as 61% are detached houses (Fig. 2).

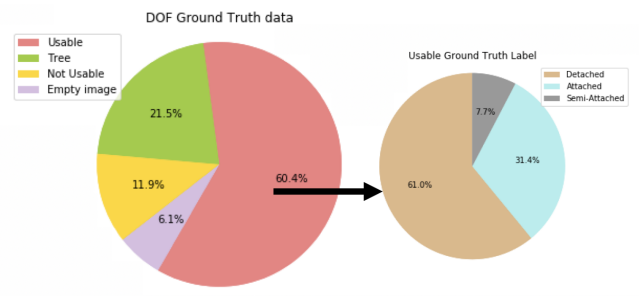

Figure 3. The percentage of usable ground truth data from NYC Department of Finance

After screening all the images of addresses with ground truth labels, only 60% are useful and can be used to train the classifiers (Fig. 3). The remaning 40% of the addresses have a Google Street View image where the proximity of the house cannot be visually assessed. Removal of unusable images further exacerbated the class imbalance problem.

Of the usuable images, 61% are detached housed while the semi-attached houses only make up 7.7%. Therefore, additional labelling was needed to reduce the severity of the class imbalance, with the added goal of limiting the bias towards Staten Island and the Bronx. Below is the example of good (Fig. 4) and bad (Fig. 5) pictures that we obtained from previous image screening. Figure 6 provides comparison the proportion of each proximity types for the ground truth data and noisy data, which shows that the noisy data has more proportionally distributed classes than the ground truth data. When comparing the ground truth labels to the noisy labels in Plutp, 14.4% of the noisy labels are erroneous.

Figure 4. Example of detached, attached, and semi-attached residential buildings

Figure 5. Example of not usable pictures from the screening result

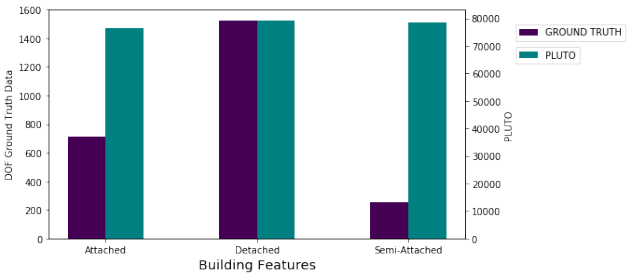

Figure 6. The proportion comparison between ground truth data and noisy (PLUTO) data

4 METHODOLOGY

The overall objective of the methods was to train a classifier that can correctly identify the proximity class of a building. We achieved that by training four image classifiers, each with different quality and quantity of training images and training labels. Additionally, we trained a baseline classifier using only the structured data (ie. the other building features associated with each address) to predict the correct proximity class.

Each image classifier was trained on a different training set of images with accompanying labels of varying quality and quantity. The first classifier, C1, was trained on the highest quality labels and images, but its training set contained the fewest images. The labels are of the highest quality because they were exclusively ground truth. The images were also guaranteed to be of high quality, because each image associated with a ground truth label was screened for quality. The training set for the C1 classifier contained 500 images. The additional screened, ground truth images were reserved for the validation and test set, which were used

for all four classifiers. A common validation and test set was used to so that the performance of the four classifiers can be compared fairly. The C2 classifier was trained on all of the ground truth labels, which included the unusable images where the house was not visible. This added an additional 1289 images to the training set for C2 (see Table 1).

Whereas C1 and C2 were only trained on ground truth labels, C3 and C4 included houses in the training set for which there was only a noisy label available. C3 was trained on all the available images, meaning that its training set was the largest with over 235,000 training images. The images are of unknown quality, but it is presumed that many are unusable, considering that only three-fifths of the screened images were usable. The C4 classifier was trained on additional noisy labels, but quality of the images was controlled for by an automated screener. The automated screening was achieved by training a usable-image classifier using the results of the hand-screening done for the C1 training set. This classifier, which had an accuracy of 85%, screened 13884 images, and of those 10769 were usable. Those images were used as the training set for the C4 classifier. Additional details regarding the specifics of the algorithms used for the image classifiers can be found in the appendix.

The baseline was trained only on structured building features, which were obtained from the PLUTO dataset. The PLUTO data consist of 84 features of geographic data from various land use classification. For predicting the building proximity, we selected nine features that are available for 1,2,3 family buildings. The features are borough, community district, building class, residential floor area, building frontage, building depth, availability of extension or freestanding structure, basement category, and when the building is built. For training and testing purpose of this dataset, we chose the same buildings that were included in the training set of the C1 image classifier. This dataset includes 1,000 training addresses and 500 test addresses which one's proximities characteristics are already verified by our DoF sponsor. The baseline was modeled using a random forest classifier out of the Sklearn package.

The image classifiers and the baseline were assessed on their performance on a test set at two tasks: proximity class detection (herein referred to as classification) and the erroneous label detection (herein referred to as ELD). In addition, the top two classifiers were externally validated on both the classification and ELD task.

The performance of each of these classifiers (the four image classifiers and the baseline classifier) was assessed by comparing the recall, precision, accuracy, and F1 score achieved on a common test set. The 500 image test set is comprised of 41 semi-attached, 113 attached, and 345 detached. The performance of the baseline provides a benchmark metric against

which to measure the advantage of using image classification.

In addition to assessing the classification performance of each model (ie, its ability to correctly predict the proximity class of each house), the utility of the classifiers was determined by assessing how well each detected erroneous labels (ie. its ability to identify a mislabeled house by predicting a class that conflicted with its noisy label). The class labels (attached, detached, semi-attached) were translated into a binary output, mislabeled or valid. The predicted class was based on the output of the classifier: a house was identified as "mislabeled" if the predicted class determined by the classifier was not consistent with the noisy label. The true condition was based on the agreement between the boisy label and the ground truth label. If the noisy label did not match the ground truth, then the true condition was determined to be "mislabeled". The success of each classifier on the ELD task was assessed by the recall, precision, accuracy, and F1 score achieved on the test set.

The top two classifiers were subjected to an external validation by a representative of the DoF. The top two classifiers were tasked with classifying the proximity label of the set of the 10769 screened images. The addresses were ranked on the reverse probability of the noisy label class, and the top 100 addresses were selected. In other words, these were the houses that the classifiers were most confident were mislabeled. The labels returned by the DoF were compared to the classifier output and the original noisy label to determine the accuracy of the classifier as well as its precision on the ELD task.

	GT LABELS		GT + NOISY LABELS	
SCREENED IMAGES	Train	500	Train	10769
	Validation	500	Validation	500
	Test	500	Test	500
ALL IMAGES	Train	1789	Train	235341
	Validation	500	Validation	500
	Test	500	Test	500

Table 1. Cross-validation of four classifiers

5 RESULT AND DISCUSSION

Reported below are the performance of each classifier on the classification task and the ELD task as well as the results of the external validation for the top two performing image classifiers. C1 was the best performer on the classification task and the ELD task when assessed via performance on the test set. C4, however, performed best at the ELD task when assessed via the external validation. The performance of the classifiers on the test set indicates that labelling ground truth labels improves overall classification performance. The confounding results of the disparate performance on the external validation may have various explanations, which are discussed below.

5.1 Proximity Classification Performance

On the classification test, C1 performed best. C1 has a high accuracy (91%) and an F1 score of nearly 80%, which far surpasses the performance of the other classifiers, as seen in Table 2. It maintains

a high recall (83%) while still maintaining a good precision (77%). Specifically, the C1 was most accurate relative to the other classifiers at identifying detached and attached classes, as shown in the confusion matrix (Figure 7). The C1 classifier identified 337 of the 345 detached houses, and 101 of the 114 attached houses, making it the best at identifying those classes out of all the classifiers. Additionally, C1 is the only classifier to outperform the baseline, which had an accuracy of 85%, as shown in Table 3.

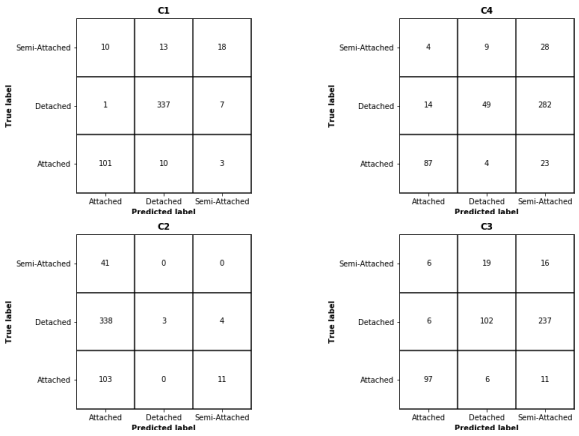

Figure 7. Classification Confusion Matrix

The C4 classifier comes in the second best. Although its ability to identify the detached falls off a notch compared to C1 (correctly identified 28 out of the total 41 semi-attached houses), its improved ability to identify the semi-attached class lift its overall performance. Despite its superior performance in the semi-attached class, C4's overall performance is low relative to C1 since semi-attached is the smallest group in the test set. While C4 does well at identifying the semi-attached houses, it has may false positives for the semi-attached class, as shown by the 282 detached houses it mis-classed as semi-attached. C4 has a precision rate 53%, recall rate 57%, and F1 score 55%. C2 and C3 perform poorly because they are performing poorly for at least two out of the three classes. C2 has a particularly poor ability to identify attached houses correctly and C3 is weakest in identifying the detached class. They have a F1 scores of only 35% and 55%, respectively. C2, C3, and C4 all perform worse than the baseline.

	GT LABELS		GT + NOISY LABELS	
SCREENED IMAGES	Accuracy	0.91	Accuracy	0.33
	Recall	0.83	Recall	0.57
	Precision	0.77	Precision	0.53
	F1-score	0.8	F1-score	0.55
ALL IMAGES	Accuracy	0.21	Accuracy	0.43
	Recall	0.41	Recall	0.49
	Precision	0.3	Precision	0.51
	F1-score	0.35	F1-score	0.55

Table 2. Classification Performance

	RANDOM FOREST
Accuracy	0.85
Recall	0.6
Precision	0.68
F-1 Score	0.63

Table 3. Baseline Classification Performance

5.2 Erroneous Label Detection Performance

The C1 was the most successful at detecting erroneous labels. Similarly to its superior performance on the classification task, C1 had the highest accuracy (93%) and the highest F1 score (71%). The next best performer, C4, has a much

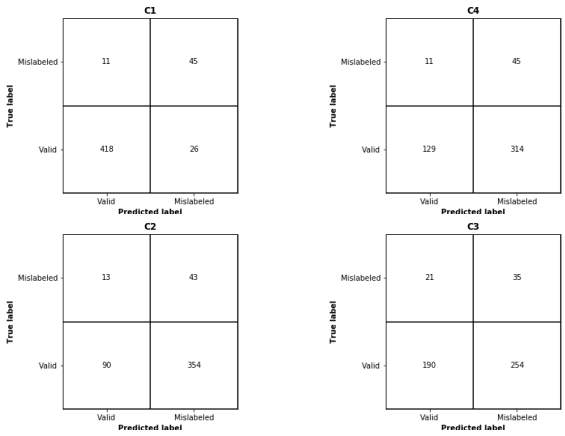

Figure 8. Erroneous Label Detection Confusion Matrix

	GT LABELS		GT + NOISY LABELS	
SCREENED IMAGES	Accuracy	0.93	Accuracy	0.35
	Recall	0.8	Recall	0.8
	Precision	0.63	Precision	0.13
	F1	0.71	F1	0.22
ALL IMAGES	Accuracy	0.27	Accuracy	0.45
	Recall	0.77	Recall	0.63
	Precision	0.11	Precision	0.12
	F1	0.19	F1	0.2

Table 4. Erroneous Label Detection Performance

	RANDOM FOREST
Accuracy	0.73
Recall	0.68
Precision	0.18
F-1 Score	0.61

Table 5. Baseline Erroneous Label Detection Performance

lower accuracy of 35% and F1 Score of 22%. The confusion matrix (Fig. 8) shows that C1 correctly identifies most of the mislabeled labels (45 out of 56) and most of the valid labels (418 out of 444). Additionally, the C1 is the only classifier to outperform the baseline random forest model (Table 4). All the other classifiers over-detect erroneous labels, resulting in many more false alarms, and subsequently lower precision. None of the latter three classifiers has a precision above 13%, whereas C1 has a precision of 63%. C4 and C3 are the next best erroneous label detectors, with nearly equal F1 scores (21% and 20%), but C4 outperforms C3 in recall, which is achieved by it catching 45 of the 56 truly mislabeled addresses. While it is able to detect those erroneous labels, it also over-corrects the noisy labels, as shown by the poor precision, and the 314 false positives.

External Validation Performance

When externally validated, C1 is the better classifier. On the external validation set, C1 has a classification accuracy of 93% compared to C4, which has an accuracy of 35%, as shown in Table 3. C4, however, is more successful at detecting erroneous labels. Of the top 100 properties that C4 was most confident were mislabeled (as measured by the classifier's reverse noisy label probability), 98 were returned with a label from the DoF. Of those 98, 96 truly had been mislabeled. For the top 100 properties that the C1 classifier was most confident were erroneously labeled, 95 were returned, and of the 95, 81 truly were mislabeled. The precision of each classifier is shown as a function of the top k most likely erroneous labels in Figure 9. A perfectly performing erroneous label detector would present in this graph as a horizontal line at a precision of 1. C4 clearly outperforms C1, when considering the precision curve. The top 20 most likely to be labeled erroneously, were in fact mislabeled, as shown by the precision of 1 for the first 20-K ranked predictions. C1 shows a large drop-off, indicating that for several houses, it was very confident that they were mislabeled when in truth, they had valid labels. As C1's confidence decreases, so does its precision, as shown by the gently downward sloping curve. C4 also loses precision as its confidence declines, but a slower rate (as indicated by the flatter slope of the curve). This indicated that on the external validation, C4 is more successful at its job of detecting erroneous labels.

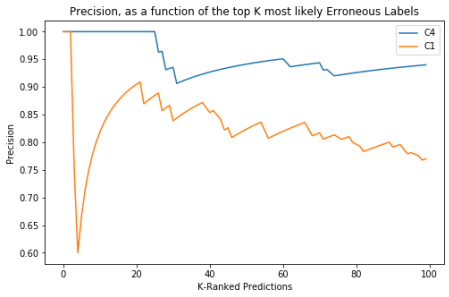

Figure 9. External Validation Precision

Comparing External Validation Results between C1 and C4 Classifier

	C1
Classification Accuracy	0.79
ELD Precision	0.81
External Validation n	95

Table 6. External Validation Results Comparison between C1 & C4

5.3 Discussion of Results

The superior performance of the C1 classifier relative to the others can be attributed more to the certainty that all the images it was trained on were usable. The effect of low-quality images can be seen in the impact on performance between the C1 and C2 classifiers, which both had entirely ground truth labels. C2 was the lowest performing classifier. That C3 and C4 performed better than C2 suggests that additional images will improve the performance of the classifier, even if many of those images are unusable and of low quality.

That C3 and C4 perform relatively equally (C4 does slightly better) is more

surprising. Presumably, the C4 classifier should outperform C3, since it should be trained on a much higher proportion of usable images (but not an entirely perfect set, since the image screener would not be perfect). One would expect a difference in performance akin to that between C1 and C3. The difference in expected performance could be explained by a few reasons. One could be the advantage of the higher training set size for C3. If most of the images are usable, and most of the images are good, then in a set of over 200,000, the classifier should pick up on the signal in those images, and learn to ignore the noise. By design, the C4 classifier was meant to be trained on a greater sample size (the entire C3 training set minus the screened out images). In practice, screening the images took longer than the scope of the project timeline allowed for. We were forced to cut off the screener early after it processed only 13,000 of the images. This resulted in the much smaller than planned training set for the C4 classifier. Additionally, this test set had a disparate class distribution than the other classifiers. While C3 was trained on an evenly spread distribution, the training set for C4 was dominated by semi-attached houses as shown in Table 7. For that reason, the C4 classifier does not perform well on the test set, which has only a small proportion of semi-attached houses, as shown in Table 8. Another confounding result is the superior performance of the C4

	GT LABELS		GT + NOISY LABELS	
SCREENED IMAGES	Attached	250	Attached	1
	Detached	664	Detached	1
	Semi-Attached	86	Semi-Attached	7
ALL IMAGES	Attached	312	Attached	7
	Detached	1659	Detached	7
	Semi-Attached	125	Semi-Attached	7

Table 7. Training Sets Distribution

	Test Set Distribution
Attached	114
Detached	345
Semi-Attached	41

Table 8. Test Set Distribution

classifier compared to the C1 when externally validated. Again, this could be a product of the disparate distribution in the training, test, and external validation sets. The C4 training set, which made up the images were classified for the external validation set, was overrepresented in the semi-attached class. Therefore, the C4 classifier would expectedly do worse on a test set that is underrepresented in semi-attached houses. The C1 classifier would similarly do better (as it does) on a test set that was representative of the distribution of its training set. Additionally, since the entire set of screened images without ground truth labels were used both as the training set and then also as the set to classify, then C4 may also be overfitting, since it has seen all those images before. Finally, the performance of C4 could be explained by survivor bias since only 100 of the results were validated by DoF. These explanations all assume that C4 is doing better than it should on the external set, judging by its performance on the test set. An alternative

explanation could be that the C1 classifier does better than it should on the test set, judging by its performance on the external validation. Since the team manually labelled additional images, which were added to the C1 test, validation, and training set. But those google street view images often contained multiple houses in one image when the house was semi-attached or attached, and it was not clear which house in the picture corresponded to the address we were labelling for. It could be that the C1 classifier learned the bias of our labelling, which would also be present in the test set. Correcting the distribution of classes in all training and test sets, reserving a separate set of images for the external validation, and ensuring the validity of our labels by using better images and stricter standards would all be solutions to correct for these confounding results.

6 CONCLUSION

Our goal in this paper is to explore the possibility of using computer vision classification to determine building features for tax mapping purposes. Based on the results of the external validation, the image classification is an effective tool for highlighting properties in the DoF portfolio that are likely to have an erroneous label that is in need of correcting. This demonstrates that an image classifier, such as the C1 classifier, can improve the desktop review process by prioritizing properties in need of label correction.

The other goal for the project is to explore the classifiers performances with limited validated labels. We constructed four classifiers with various combination of the screen labels, noisy labels, screened images and noisy images. Our classifiers with ground truth labels and screened images outperform our other models by a large margin. So as our preliminary results demonstrated, the quality of labels and images are more important than the volume of the training sets in our classification problem.

6.1 Future Works

Our future works should consider few possibilities. First, one primary constraint in our current project is that the image quality we have downloaded from Google Street View is poor in resolutions and quality compared to other sources of data Open Street maps or Cyclomedia dataset. In the future, we would like to perform our experiment on more reliable images to see if the model will improve. Second, our ground truth training set is limited both in its accuracy and also in number. Due to time constraint, we can only label a limited number of images and due to our lack of domain knowledge and poor quality of images, our labels might be incorrect. For future work, we would want to create a larger and more precise training set. A few adjustments can be made on the model as well. As time constrained, we did not manage to carefully manipulate our training dataset to have same proximity distribution match exactly our test set or

what the actual distribution of properties . As our results suggested, the distribution of training sets can have some impact on the prediction performances. So it is worthwhile to think about possibility to fine-tune or customize classifiers for even specific borough by adjusting the proximity distributions in the training set. Lastly, we utilized the structure dataset as a way to establish a baseline for evaluating the performance for the image classifiers. There might also be value in using the structured data in addition to the image classifier to make a composite model for our classification problem.

7 REFERENCES

Bosch, A., Zisserman, A., & Munoz, X. (2007). Image Classification using Random Forests and Ferns. 2007 IEEE 11th International Conference on Computer Vision. doi:10.1109/iccv.2007.4409066

Cannon, S. E., Cole, /& R. A. (2011). How Accurate are Commercial Real Estate Appraisals? Evidence from 25 Years of NCREIF Sales Data. SSRN Electronic Journal. doi:10.2139/ssrn.1824807

Carbone, R., and Longini, R. L. (1977). A Feedback Model for Automated Real Estate Assessment. Management Science, 24(3), 241248. doi:10.1287/mnsc.24.3.241

Chapelle, O., Haffner, P., /& Vapnik, V. (1999). Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks, 10(5), 1055-1064. doi:10.1109/72.788646

Dalal, Navneet and Triggs, Bill. (2005). Histograms of Oriented Gradients for Human Detection. Comput. Vision Pattern Recognit.. 1. 886-893. 10.1109/CVPR.2005.177.

Dimitrov, A., & Golparvar-Fard, M. (2014). Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections. Advanced Engineering Informatics, 28(1), 37-49. doi:10.1016/j.aei.2013.11.002

Jain, Sadhana, "Remote sensing application for property tax evaluation, International Journal of Applied Earth Observation and Geoinformation", Volume 10, Issue 1, February 2008, Pages 109-121

Kang, J., Körner, M., Wang, Y., Taubenböck, H., & Zhu, X. X. (2018). Building instance classification using street view images. ISPRS Journal of Photogrammetry and Remote Sensing. doi:10.1016/j.isprsjprs.2018.02.006

Kok, N., Koponen, E., and Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. The Journal of Portfolio Management, 43(6), 202-211. doi:10.3905/jpm.2017.43.6.202

Krizhevsky, A., Sutskever, I., and Hinton, G.E. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and

K. Q. Weinberger (Eds.), Vol. 1. Curran Associates Inc., USA, 1097-1105.

Lowe, D. (1999). Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision. doi:10.1109/iccv.1999.790410

Luts, Byron F., "The Connection Between House Price Appreciation and Property Tax Revenues", National Tax Journal, Vol. 61, No. 3, NOW FOR SOMETHING COMPLETELY DIFFERENT: TAX POLICY AT THE CHANGE IN THE PRESIDENCY (September, 2008), pp. 555-572

Schulz, R., Wersing, M., and Werwatz, A. (2013). Automated valuation modelling: A specification exercise. Journal of Property Research, 31(2), 131-153. doi:10.1080/09599916.2013.846930

You, Q., Pang, R., Cao, L., and Luo, J. (2017). Image-Based Appraisal of Real Estate Properties. IEEE Transactions on Multimedia, 19(12), 2751-2759.

Zhu, Q., Yeh, M., Cheng, K., and Avidan, S. (2006). Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR06). doi:10.1109/cvpr.2006.119

8 APPENDIX

From above, it can be concluded that in future works, we might well change the characteristic "proximity level" to a clearer one like "exterior wall material" or any feature that a learning algorithm can extract features of more easily. Second, we should find a way to deal with the situation that there could be 3 to 4 houses in a single image. Additionally, we can try object detection which contains both object localization and objects classification since it will make the classification part more accurate. Adding additional features to the classifier, such as borough, is another strategy we will attempt to improve the performance of our classifier.

Based on these initial results as they stand, however, it appears that the best classifier is the one trained on the screened, GT labeled training images which is provided by DOF since it achieved the highest accuracy 91.2% and F1-score 79.6% compared to other 3 classifiers. This suggests that even we tried to screen out all of the unusable images, with more training samples, the performance of the classifier actually diminished so chances are that the biggest problem is the algorithm is struggling to extract features of a specific house in the image as in most cases there are at least two houses in an image and they can be located in any places.

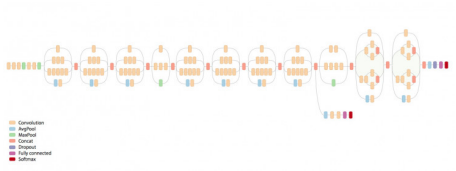

Figure 10. Based on the fact that the training sets are quite small, we decided to finetune the Inception Resnet V2 classifiers instead of training a new one using Tensorflow Slim module which is developed for finetuning pre-trained models. By finetuning, we mean only updating the weight parameters in the final softmax layer which is used to output the probabilities of each class that the classifier predicts. Due to a lack of time, we trained all of the three classifiers in no more than 20000 iterations using an initial learning rate of 1 with exponential decay factor of 0.76, cyclical learning rates, an Adam optimizer and a batch size of 32.