# Using Computer Vision and Machine Learning to Unlock Historical Data

**Jun Tao Luo**

CMU-CS-23-111

April 2023

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Matthew Gormley, Chair
Rayid Ghani

*Submitted in partial fulfillment of the requirements for the degree of Master of Science.*

**Keywords:** Machine Learning, Computer Vision, OCR, Historical Records

*For my late grandfather who instilled a love of learning in all of us.*

## Abstract

Historical data, especially those recorded in tables and forms, have significant value for contemporary research and industry applications. However, such data is rarely digitized or available in readily usable formats such as Excel sheets and database tables. Using historical property appraisals as a case study, we demonstrate how machine learning and computer vision methods can help address this data gap in a cost-effective way.

The earliest standardized property appraisal records in the United States were typically handwritten on physical cards. Using scanned cards from Ohio in the 1930s, we test approaches to digitize a property's earliest appraised value. We find that image processing and Optical Character Recognition (OCR) deep learning models can retrieve this value accurately with a Mean Absolute Percentage Error (MAPE) of 14.72%. For cases where OCR cannot be applied, such as when scanned documents are not available, our machine learning model can use contemporary data to estimate this value with a reduced accuracy of 17.48% MAPE. Both methods present a substantial saving over manually digitizing the same data, with OCR achieving a cost reduction of 81% and the machine learning model achieving a cost reduction of 89%.

## Acknowledgments

I would like to thank Matt Gormley and Junia Howell for giving me this opportunity to work on this exciting subject and learn a wealth of concepts and skills. Without their valuable insights and advice, this work would not have been possible. I would also like to thank my family, friends, advisors and who supported me along the way.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Across disciplines, there are several academic and practical use cases for historical data. For example, historical economic data on trade records has informed research into how globalization has developed over time [10]. Often, this data is stored in physical formats with handwritten information that is not easily accessible or usable by data analysts and scientists in its original form. This is especially challenging for data recorded in tables and forms where the structure of the document, such as the relationship between rows and columns, is critical to correctly parsing the document. Current methods for obtaining usable information from these historical records involves manually scanning documents and manually entering data which can be prohibitively costly for large numbers of documents.

As an alternative to existing methods, we propose a cost effective process for digitizing such records using computer vision, OCR and machine learning techniques (see Figure 1.1).

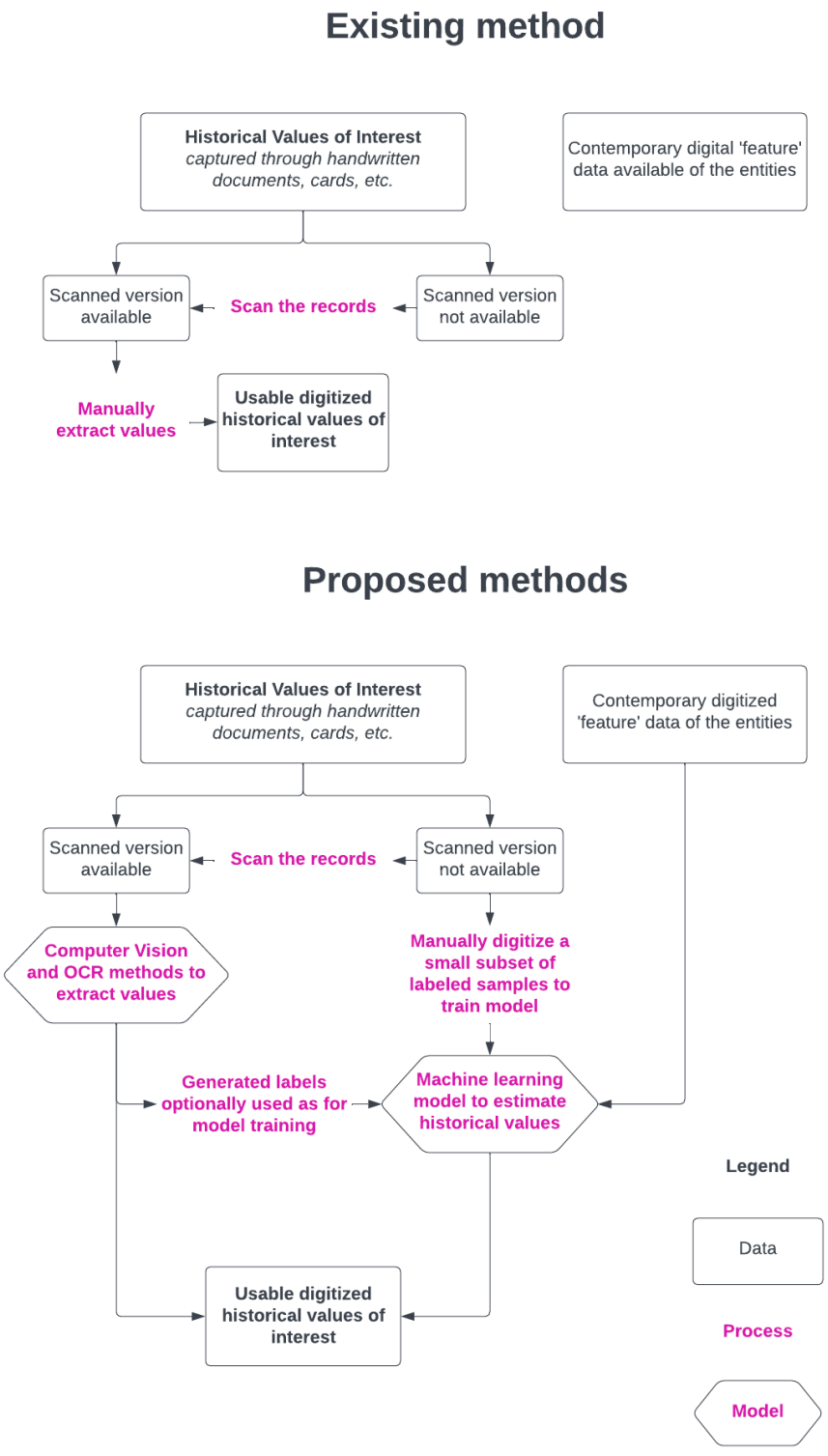



Figure 1.1: Methodology for Digitizing Records

We begin with the premise that the historical data of interest was recorded in a relatively consistent format across entities. This holds true for many documents such as Census surveys and property appraisals. While a computer vision and OCR approach can be applied when there are scanned versions of the documents, we recognize that there are cases where only physical copies exist, and thus explore using a machine learning model to estimate the values of interest in these settings. Our goal is to present a holistic approach that contains a menu of options for digitization with varying cost and accuracy trade-offs.

To test and evaluate this approach, we use historical property appraisal cards from Hamilton County (Cincinnati) as a case study. Historical property appraisals are of particular interest for equitable housing research which explores the connection between race and property values [13]. In this field of research, housing equity scholars have proposed fairer alternative appraisal systems which value properties using data on the actual costs of construction rather than comparable sales [12]. This system is untested because it relies on digitized construction cost data, and in Hamilton County, approximately 55% of all buildings were constructed before such data exists (1960). Researchers argue that early historical property appraisals are a good proxy for the original construction costs of these old buildings because of appraising practices at the time.

In summary, the goal of this case study is to digitize the earliest available building appraisal value from the historical property cards, such that it can serve as a usable measure of original construction costs for buildings of this vintage.

We demonstrate that machine learning and computer vision methods can help bridge this data gap and produce reliable estimates of these costs for residential buildings constructed before 1930. We use computer vision techniques and Optical Character Recognition (OCR) models to extract the building appraisal value from scanned property ownership cards in 1933, which is the earliest year available. For cases where this value cannot be extracted using OCR, we combine manually labeled ownership cards with contemporary information about the land parcels[1]

[1]Parcels are the main administrative unit for properties

and building characteristics to construct a training set. We then build machine learning models to estimate the original construction cost using only the contemporary data. To evaluate the extensibility of our models on other counties, we use data collected from Franklin County (Columbus) to measure our generalization performance.

## 1.1 Related Work

Recent interests in digitizing data from historical records have turned to OCR technologies to extract data from scanned documents including balance sheets [5], newspapers [3], and other historical texts [17]. Like ourselves, some of this work has primarily focused on the challenges of numeric extraction from historical census [19] and church documents [27]. However, the historical documents we are digitizing are in tabular form which requires segmenting the table into cells before applying OCR techniques similar to approaches in described in tabular OCR works [20], [9], [22]. We also incorporate techniques from previous OCR works including TesseractOCR [24] and transformer based TrOCR [16] in developing our models.

To the best of our knowledge, there is no literature using machine learning models to estimate data from historical documents. Machine learning models have been used to link families and individuals across historical census records, but not estimate specific values from these records [8], [21].

There is a rich literature using machine learning to predict contemporary property value and sale price given its industry applications to real estate valuation and transactions [28], [11], [26], [2], [25]. However, we could not find any papers that attempt to estimate historical property or building values.

# Chapter 2

# Experimental Setting

This section details how we translate the broad approach defined in Figure 1.1 to our use case. We use Hamilton County (Cincinnati), Ohio as a starting point because of the public availability of both scanned historical appraisal cards and contemporary property information. Sections 2.1 and 2.2 describe the available data, processing needed to apply our methods, and how we arrived at a subset of parcels for analysis.

Section 2.3 describes the metrics we measure performance against, with Section 2.4 proposing a simple baseline method of estimating building construction cost to benchmark our results against.

## 2.1 Data Sources

We obtain administrative data made publicly available by the Hamilton County Auditor on their website. [1] The records listed below are linked by a unique identifier at the land parcel level. [2]

[1]The Auditor is the County's Chief Fiscal Officer and Property Assessor. Their website is: https://hamiltoncountyauditor.org/

[2]We use "parcels" and "properties" interchangeably, with "buildings" used to refer to specific constructions.

### 2.1.1 Target Data: Historical Property Ownership Cards

These documents are scanned images of the historical property details for a parcel. This includes ownership and transfer information as well as land and building valuations. While dates are missing for most of the valuations, process documents suggest that the assessments that generated these values followed the same three year cycle used today, with the first values generated in 1933. An example of this document can be seen in Appendix A. Our main target for digitization comes from this body of documents: the initial value of the building, as recorded in 1933, which serves as a proxy for its original construction cost. The distribution of this variable, based on 10,452 randomly selected hand-labeled samples, is shown in Figure 2.1.

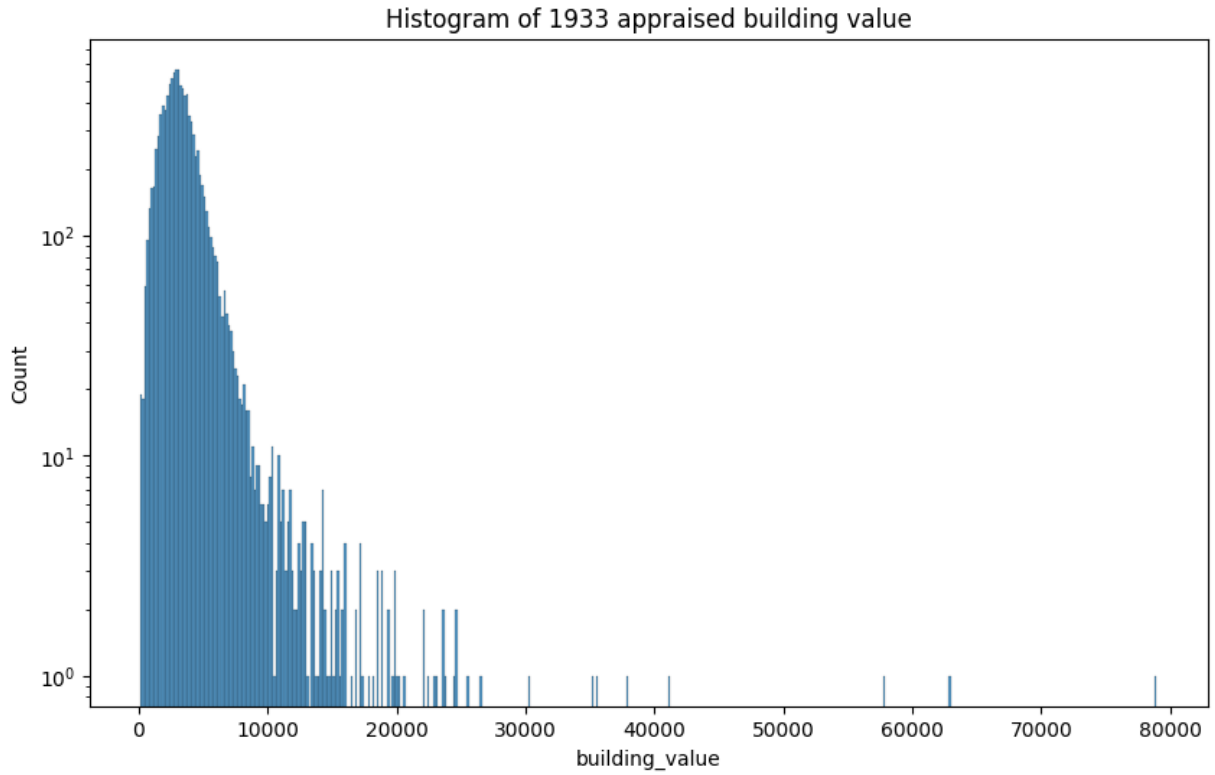



Figure 2.1: Histogram of Target Value

### 2.1.2 Feature Data: Contemporary Tax Assessment Information

Every three years, the County Auditor updates all property assessments for tax purposes [1]. We use data from the latest assessment, in 2020, as the most updated and comprehensive set of parcels. The data includes information about the parcel's administrative status, information about the physical characteristics of any buildings on the parcel, and information about valuations and sales. A full set of features[3] used from this data is listed in Table 2.1.

Table 2.1: Features built from contemporary data

| | |
|---|---|
| Square footage | attic, basement, floor 1, floor 2, half-floor, total livable area |
| Building characteristics | stories, style, grade/condition of building, exterior wall type, basement type, heating type, air conditioning type, total number of rooms, total full and half bathrooms, number of fireplaces, garage type and capacity |
| Parcel characteristics | land use code, neighborhood, number of subparcels |

## 2.2 Data Processing

The raw data for contemporary tax assessment information could be downloaded directly as structured Excel files from the source. Several data cleaning steps were needed to process the data into a format ready for analysis, including handling nonsensical values, grouping categories, and creating consistency in formats across source tables. These are detailed in Appendix C.

The 353,973 parcels found in the contemporary data were subset based on the following criteria:

[3]We use one-hot encoding for categorical features

1. **Parcels defined as residential**: building characteristics such as rooms and bathrooms are not captured for commercial buildings
2. **Parcels with one single, finished building**: the data does not contain a building identifier, making it impossible to know which building the building characteristics pertain to. Hence, we focus on only parcels with one building.
3. **The building was constructed before 1930**: since the target value of interest is the appraised building value recorded in 1933, we only consider parcels that had a construction before 1930.
4. **No data inconsistencies between tables**: to ensure all the feature information could be used, we ignored parcels that did not match across source tables or contained inconsistent information about building characteristics between source tables.

The resulting set of 59,378 parcels forms the overall sample of interest for Hamilton County. Of this set, we were only able to successfully retrieve 56,037 scanned documents, which indicates 5.6% of the parcels are missing their ownership card documents. Next, we perform basic preprocessing on the documents including cropping, rotating, and conversion to grayscale.

In order to create labels for the machine learning models, we randomly sampled 12,423 of the retrieved ownership cards for manual labeling. See Appendix B for additional details about the manual labeling process. For a breakdown of the number of samples after each processing step, see Figure 2.2.

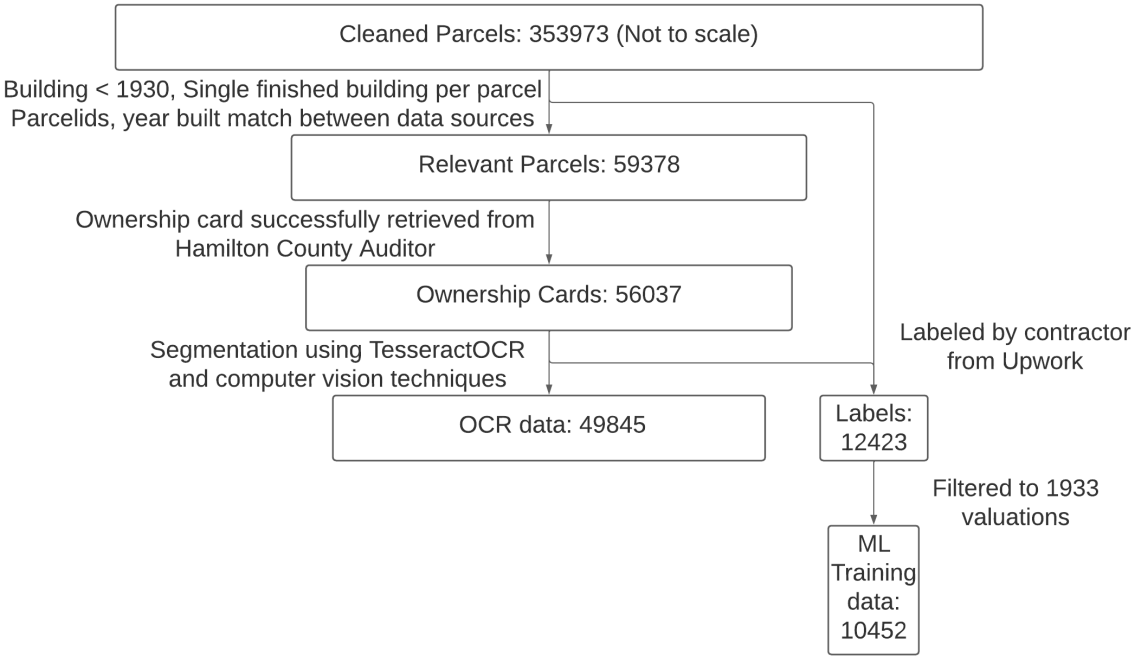



Figure 2.2: Data Processing Flow

## 2.3 Evaluation metrics

To assess the performance of our methods, we consider a variety of standard metrics used in regression problems including Mean Absolute Error and Root Mean Squared Error. A common metric in real estate value estimation is the mean absolute percentage error (MAPE), and thus we use this as our primary metric for reporting results.

We report results on buildings in the middle 90% of properties based on appraised value, to reduce the effect of outliers.

## 2.4 Baseline Model

Our baseline model was created with the following question in mind: in the absence of machine learning and OCR, what would an investigator do to estimate the original construction cost of

buildings built before 1930?

We use a national model for estimating construction costs based on the American Housing Survey, that takes into consideration region and urban/rural classification. Specifically, we use the following equation for Midwest urban regions which contains both Hamilton County.

$$\begin{aligned}
&\text{Estimate Construction Cost} \\
&= -231522.85458 + 116.894831 * \text{Year} \\
&+ 1222.49492 * \text{Basement} + 6267.1243 * \text{Heating} \\
&+ 1769.41924 * \text{Central AC} + 1579.6329 * \text{Total rooms} \\
&+ 991.89863 * \text{Bathrooms} - 4.9088314 * \text{Half Bathrooms} \\
&+ 2373.59715 * \text{Garage}
\end{aligned} \tag{2.1}$$

In Equation 2.1, the variables are defined as follows:

- Year - The year the home was built
- Basement - A dichotomous variable indicating the presence of any basement
- Heat - A dichotomous variable denoting the presence of heating system
- Central AC - A dichotomous variable denoting the presence of central air conditioning
- Total Rooms - The total number of rooms in the building
- Bathrooms - The number of bathrooms in the building
- Half Bathrooms - The number of half bathrooms in the building
- Garage - The number of cars the garage can hold

# Chapter 3

# Methods

This chapter described the details regarding the models and experiments that were performed as part of this work.

Section 3.1 describes the computer vision and OCR workflow: that is, given the availability of scanned cards, the methodology we use to extract the earliest appraised value. Section 3.2 describes how we build the machine learning models to estimate the value using contemporary feature data. Section 3.3 details how we combine these two workflows by incorporating predictions from OCR into the training data for the ML models.

Finally, Section 3.4 describes an experiment to test how well the model generalizes to Franklin County (Columbus), Ohio. A model that could generalize across cities and regions would unlock significantly more historical appraisal data for use by researchers and practitioners at lower cost than building city-specific models.

## 3.1 Computer vision and OCR workflow

### 3.1.1 Tabular Data Segmentation

Although there are many existing solutions for OCR, no individual method accomplishes our task completely. The first challenge is to recognize the tabular structure of the ownership card documents and locate the relevant information. Given that we use the building value of the first recorded appraisal in the Hamilton ownership documents as a proxy for our target variable, this involves obtaining the first entry of the ”BUILDINGS” column. To accomplish this, we use a customized process for segmentation which involves using TesseractOCR to locate the column header ”BUILDINGS” then using Hough Transform to locate surrounding row and column divisions for cropping individual table cells. The individual cropped cells are then used for OCR. For more details, see Appendix D.

### 3.1.2 Optical Character Recognition (OCR) Models

Given that our task involves recognizing only numerical values, it is challenging to use off the shelf OCR solutions or pretrained models such as TesseractOCR or TrOCR since these models predicts all characters, including digits, punctuation and letters, and performs poorly on our dataset. Initial experiments shows that these pretrained models would often confuse letters and digits including recognizing the digit 0 with the letter O and the digit 1 with lowercase L or uppercase I. To address this type of error, we perform additional find tuning using a mixture of different datasets including CAR-B (handwritten digit strings from scanned checks) [6] and DIDA (historical handwritten digit dataset) [15]. For detailed descriptions of our OCR experiments, see E.

## 3.2 Machine Learning (ML) model workflow

We approach the machine learning component as a standard regression problem, where the target value to be predicted is the labeled 1933 building appraisal value. As mentioned in Section 2.2, we have 10,452 parcels with labels collected by hand, which we merge with the contemporary feature data outlined in Table 2.1 to create the training and test matrices. We use an 80-20 train-test split, and employ 5-fold cross validation within the training set for hyperparameter tuning.

We use a stepwise approach to model selection. First, using a single set of default hyperparameters, we train many different model classes and observe performance on a validation set. The results of this are in Appendix F. We then select the best performing model classes, and conduct a more extensive hyperparameter grid search, selecting the best model using the 5-fold cross validation root mean squared error (RMSE).

This approach leads us to choose the following random forest regressor as the best machine learning model:

Table 3.1: Chosen ML model

| | |
|---|---|
| Model class | Random forest regressor |
| Number of estimators | 2500 |
| Max depth | 200 |
| Minimum samples for split | 4 |
| Max features | sqrt |

## 3.3 Augmented ML Models

While OCR methods and ML methods are two separate approaches for predicting the same target variable, they accomplish their task using different inputs and techniques. These two methods are complementary to each other in that they can be combined in various ways to improve performance. In this work, we use the trained OCR model to create annotated labels for training the

ML model. This allows the use of all 56,037 retrieved canned documents for training and testing of the ML model instead of only the 12,423 manually labeled samples. This approach has been shown to improve the performance of the ML models, see Section 4.5.

## 3.4 Model generalization

To test whether our ML model generalizes to a different city, we collect test data from Franklin County (Columbus), Ohio. Similar to Hamilton County, Franklin County has publicly available data on both historical scanned appraisal cards as well as contemporary feature data from recent appraisals. Since the appraisal cycles of both counties did not exactly align, we chose the closest appraisal year to 1933 that we could find data on in Franklin County. For most properties, this was the 1931 appraised value.

Using the same logic as in Hamilton, we subset the universe of parcels in Franklin to a sample of 42,100 parcels that have one residential building built before 1930. We manually hand-labeled a randomly drawn subset which provides us with a small test set of 506 observations.

To apply the trained model to make predictions in Franklin County, we had to ensure that the features in Franklin County were comparable to those used in the Hamilton model. While some of the important features were common (e.g., square footage of floor 1), there were several features not available in Franklin County or captured in a different format (e.g. presence of attic captured rather than specific square footage). To test generalization, we train the model with only the subset of features that were comparable across both counties, and use this limited model to report performance on the Franklin County test set.

See Appendix C.1 for more details on the feature subset used.

# Chapter 4

# Results

The statistics of the best performing models from our experiments described above is shown in Table 4.1. We sampled 20% of the manually generated labels and filtered the samples to values within the 5-95 percentile range to remove outliers.[1] The resulting test set is used to report the metrics in this section. We use several common statistical evaluation metrics for regression tasks including coefficient of determination ($R^2$), Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), the Median Percentage Error (MPE) for which half of the predictions has a percentage error that is smaller and the percentage of test cases where we predicted a value that is within 5%, 10%, or 20% of the true value.

[1]The generalization model reports performance on the smaller test set of 506 observations in Franklin County.

Table 4.1: Prediction performance of evaluated models

| Metrics | Baseline | OCR | ML | Augmented | Generalization |
|---|---|---|---|---|---|
| $R^2$ (higher is better) | 0.0195 | 0.6264 | 0.6177 | 0.7428 | 0.0521 |
| MAPE (lower is better) | 33.89% | 14.72% | 17.48% | 16.12% | 40.19% |
| RMSPE (lower is better) | 52.10% | 40.04% | 27.73% | 24.01% | 63.30% |
| MPE (lower is better) | 21.49% | 0% | 10.60% | 11.27% | 28.31% |
| Within 5% of True Value (higher is better) | 11.84% | 85.36% | 25.81% | 24.39% | 9.112% |
| Within 10% of True Value (higher is better) | 25.44% | 85.39% | 48.06% | 45.85% | 17.99% |
| Within 20% of True Value (higher is better) | 47.32% | 85.40% | 73.44% | 74.55% | 35.98% |

## 4.1 Baseline Model

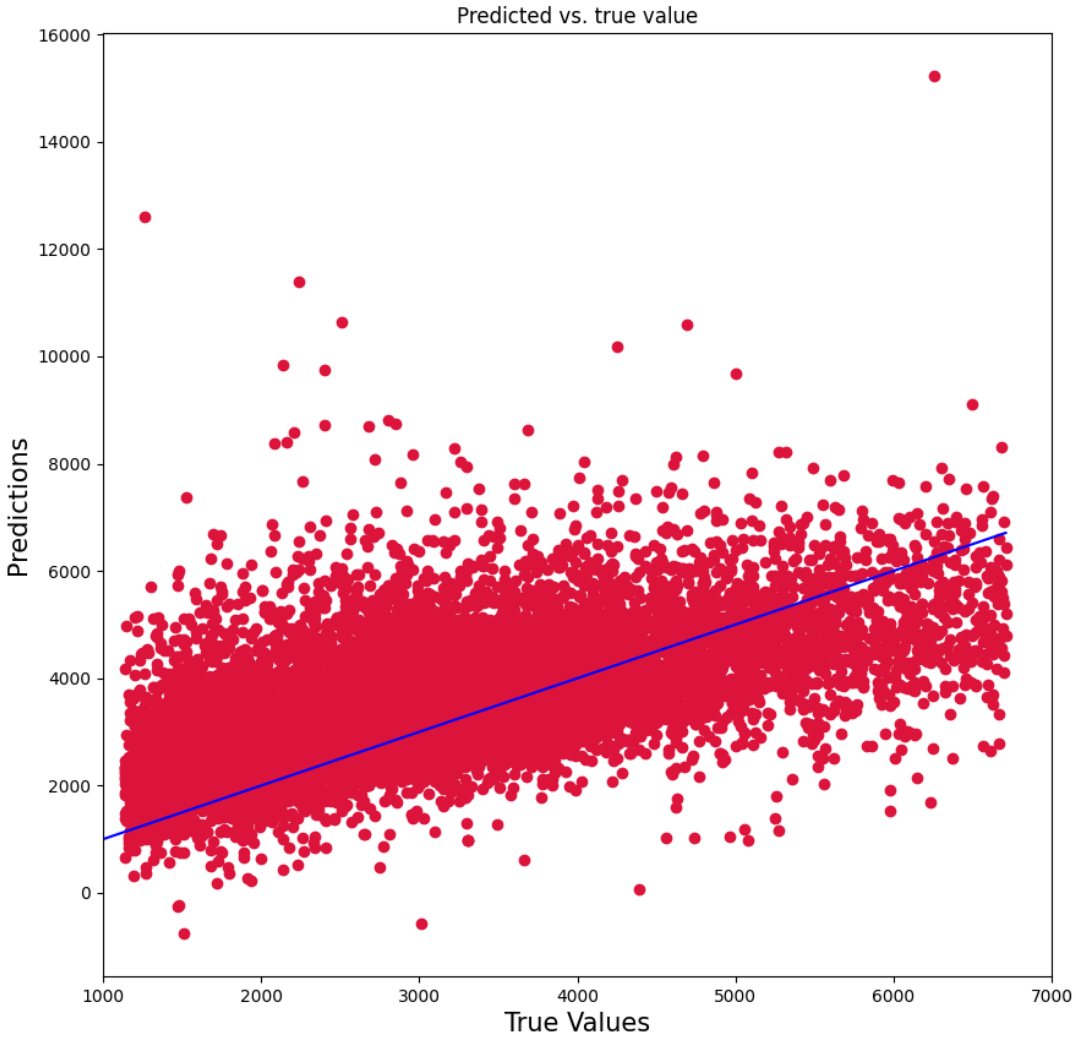



Figure 4.1: Baseline Model Predictions

We use the data for Hamilton county parcels and compare the estimates generated by Equation 2.1 with the hand labeled ground truth values. The predictions of this model are shown in Figure 4.1. Analysing the errors between the predictions and true values shows that the model makes

large errors in its prediction, as illustrated by large MAPE, RMSPE and MPE values, and these errors are common with only 11.84% of the predictions falling within 5% of the true values. This model performs relatively poorly since it is developed using regional level data (e.g. Midwest Urban) but does not account for trends existing at the county level like the other methods proposed in this work.

## 4.2 Tabular Data Segmentation

Since this component does not directly produce estimates for the construction cost of the building it is not included in Table 4.1 even though it contributes to the performance of the OCR model. There are two metrics of interest when evaluating the segmentation method: the success rate of extracting a segment and the accuracy of extracting the correct segment. For the success rate, we use our segmentation algorithm on 56,037 documents and are able to successfully extract segments for 49,845 of them giving a success rate of 89.0%. To evaluate the accuracy, we randomly sample 831 ownership documents and examine the tables to compare if the extracted table segment is correct. We find only 1 error case where the segment represents the second cell in the column instead of the first, giving an accuracy of 99.9%.

## 4.3 Optical Character Recognition

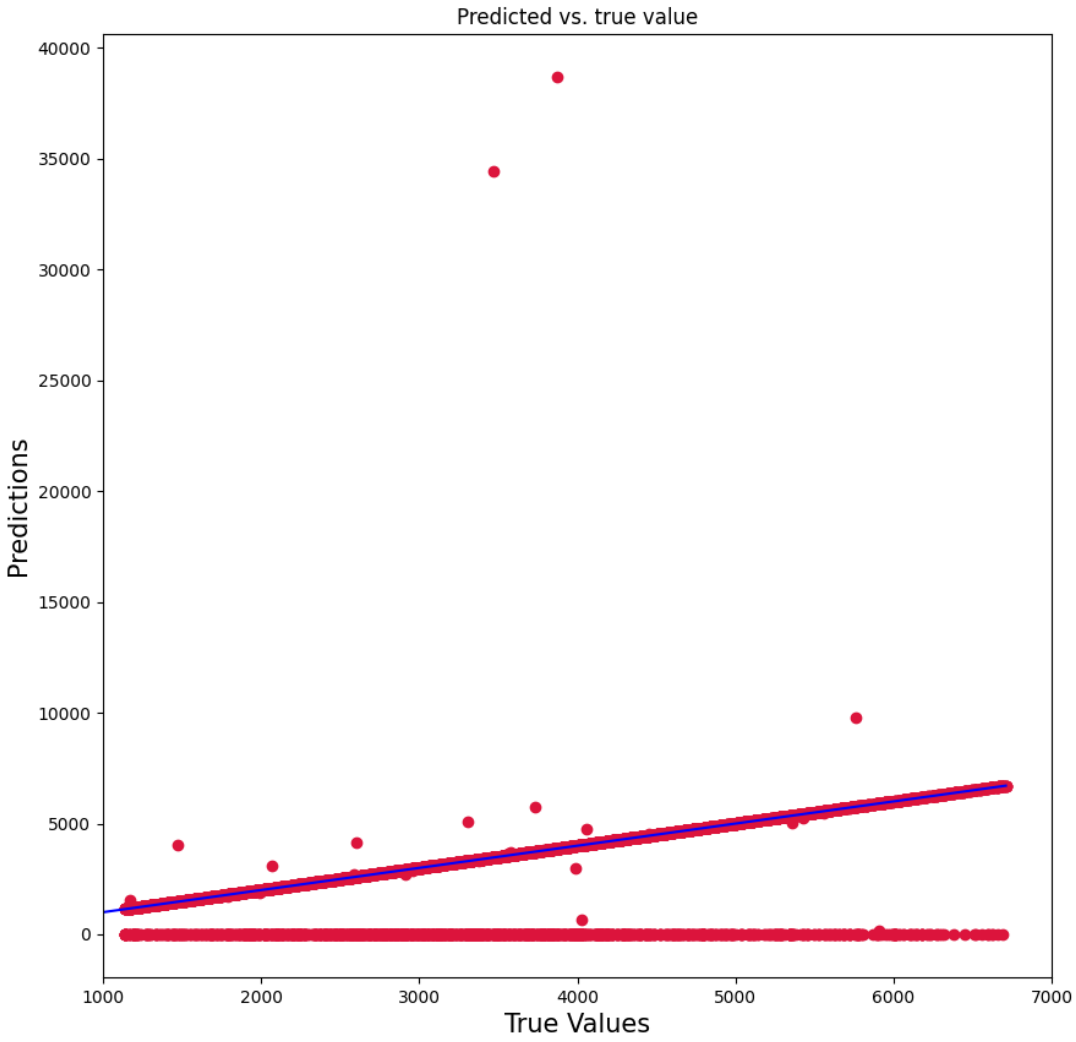



Figure 4.2: OCR Model Predictions

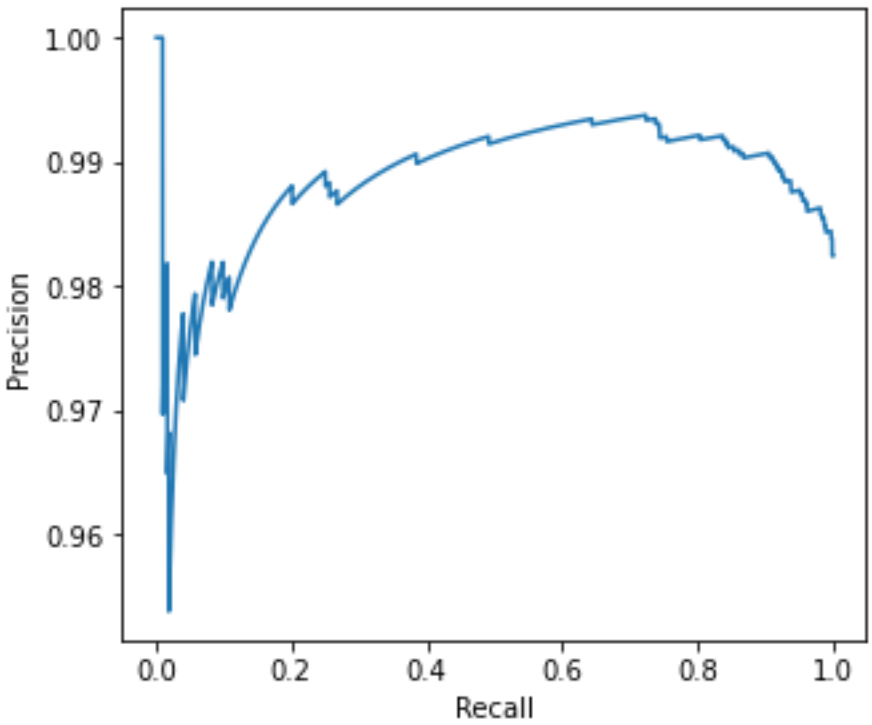



Figure 4.3: OCR Model Precision vs Recall

Table 4.2: Prediction metrics of OCR models for different confidence thresholds

| Metrics | Top 90% | Top 95% | Top 99% | All 100% |
|---|---|---|---|---|
| $R^2$ (higher is better) | 0.7663 | 0.6958 | 0.6350 | 0.6264 |
| MAPE (lower is better) | 5.417% | 10.36% | 13.86% | 14.72% |
| RMSPE (lower is better) | 26.21% | 34.26% | 38.97% | 40.04% |
| MPE (lower is better) | 0% | 0% | 0% | 0% |
| Within 5% of True Value (higher is better) | 94.68% | 89.73% | 84.19% | 85.37% |
| Within 10% of True Value (higher is better) | 94.71% | 89.76% | 86.25% | 85.39% |
| Within 20% of True Value (higher is better) | 94.72% | 89.77% | 86.26% | 85.40% |

We find the best performing OCR model to be TrOCR fine tuned on a mixture of our Hamilton county dataset combined with additional handwritten digit data from CAR-B. Fine tuning on the DIDA dataset was found to be detrimental since the digit strings are primarily year values recorded in church documents caused the fine tuned TrOCR model to incorrectly predict values between 1800-1940 more often. The results from the best performing TrOCR model trained on 7375 entries randomly sampled from our Hamilton county dataset and 3000 entries from CAR-B, see Figure 4.2. We find that this model is relatively accurate with low MAPE and MPE values. Upon further analysis, we find that while errors are rare, as evident by the fact that 85.36% of all predictions falling within 5% of their true values, the magnitude of the errors are large. This is often due to the insertion or deletion of digits which creates extremely large errors and results in large RMSPE and is reflected by the outliers in Figure 4.2. We note that another common source of error for this model is predicting a value of ”0” which occurs when the OCR model fails to detect recognizable digits. Fortunately, these errors are usually accompanied by a low confidence score which allows these low confidence predictions to be filtered. The Precision-Recall curve in Figure 4.3 illustrates the impact of varying the confidence threshold for OCR predictions. By choosing an appropriate threshold, we can achieve an exact match accuracy of up to 99.4%. To evaluate the impact of this filtering on the outputs of the model, we report the accuracy metrics for retaining top 90%, 95% and 99% of the most confident predictions, see Table 4.2. We see significant improvements in the model performance if we retain only the top 90% of the most confident predictions, achieving and MAPE of 5.417% and able to make a prediction within 5% of the true value for 94.68% of the test cases.

## 4.4 ML Models

The chosen random forest regressor model predicts the target value with an MAPE of 17.48%, which is a substantial improvement over the baseline method. As seen in Figure 4.5, the ML models seem to perform worse on higher-value properties, with larger over-predictions and

under-predictions. The two most important features were the attic size and basement square footage, with the building's height (stories), garage capacity, and floor 1 square footage also appearing in the top 5 features. See Appendix G for a plot of all the feature importances.

A relevant question for our proposed approach is the number of samples that need to be manually digitized for the machine learning model to predict the target value accurately. Figure 4.4 shows the improvement in MAPE as the size of the training set increases. As the number of labeled samples in the training set increases from 3,000 to 8,000, the MAPE drops from roughly 18.6% to 17.5%. We did not collect additional samples, but based on the trend it appears that additional data would improve performance.

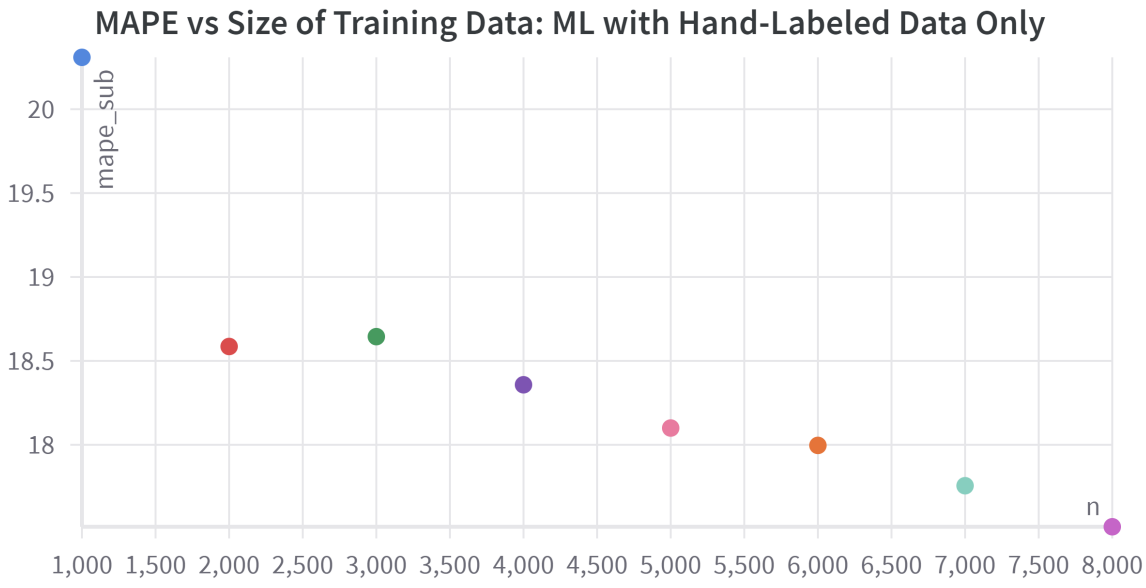



Figure 4.4: MAPE as size of hand-labeled training data increases

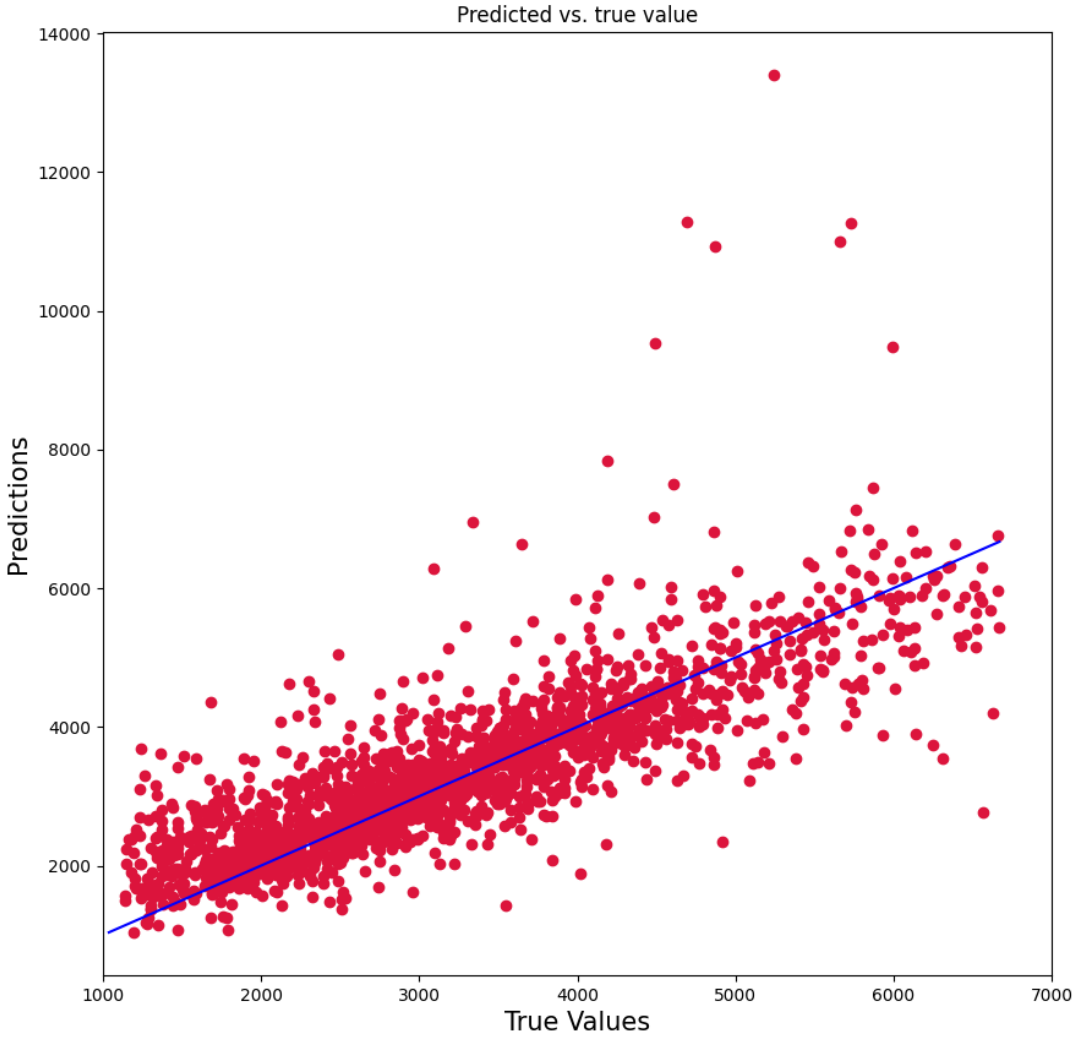



Figure 4.5: OCR Model Predictions

## 4.5 Augmented ML Models

Augmenting the number of training samples by using the predictions from the OCR models on unlabeled Hamilton County samples as training samples drastically increase the number of

training data for the ML models. As discussed in Section 4.3, in order to improve the accuracy of the OCR labels used for training the ML models, a threshold on the OCR model prediction confidence should be used. This presents a trade off between the quantity and quality of the training samples using this augmentation method. Choosing a high confidence threshold of the OCR predictions means the training samples are low but also contain fewer incorrect labels and vice versa. To examine this effect, the performance of the augmented ML models using different OCR prediction confidence thresholds retaining the top 99%, 90% 75% and 50% of the most confident prediction, is shown in Figure 4.6. Here we find that while we see a slight improvement in some accuracy measure such as MAPE from 17.48% to 16.12% and RMSPE from 27.73% to 24.01%, other measures MPE see a slight decline. We also note that while the amount of predictions within 20% of the true values improved from 73.44% to 74.55%, the amount of predictions within 5% and 10% of the true values decreased. This result suggest that is is not conclusive that augmenting the ML models using OCR predictions is beneficial. From our analysis, the outliers in the OCR predictions, despite our efforts to remove them by applying a threshold for the OCR prediction confidence, are highly detrimental to ML models and offset the benefits of additional training samples.

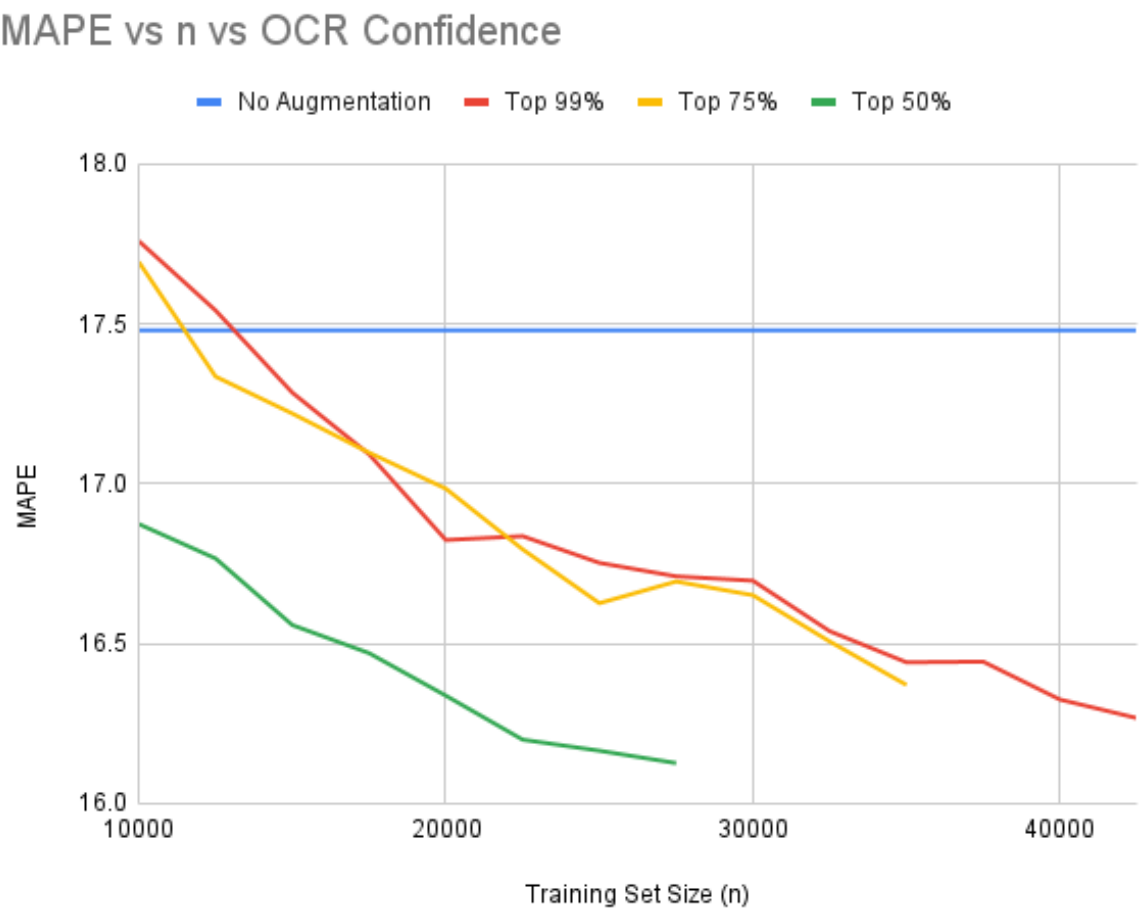



Figure 4.6: MAPE and OCR Confidence Threshold vs n

## 4.6 Generalization

The results on the Franklin County test set are significantly worse than for the Hamilton County test set, with an MAPE of 40.19%. The ML model underperforms the simple baseline method. One possible explanation for this is the limited set of features used in the model. However, using this same limited model, the Hamilton County test set performance remains stable at 17% relative to the full model discussed in the previous sections.

Within the Franklin test set, we observe that the median target value is $2,300, which is lower

than the Hamilton median of $3,085. Figure 4.7 shows that the model seems to be consistently over-predicting the values, since we do not perform an adjustment for price differences between the counties. With additional hand-labeled samples, a domain adaptation approach could be considered wherein a small number of samples from Franklin are also included in the training data to improve generalization performance.

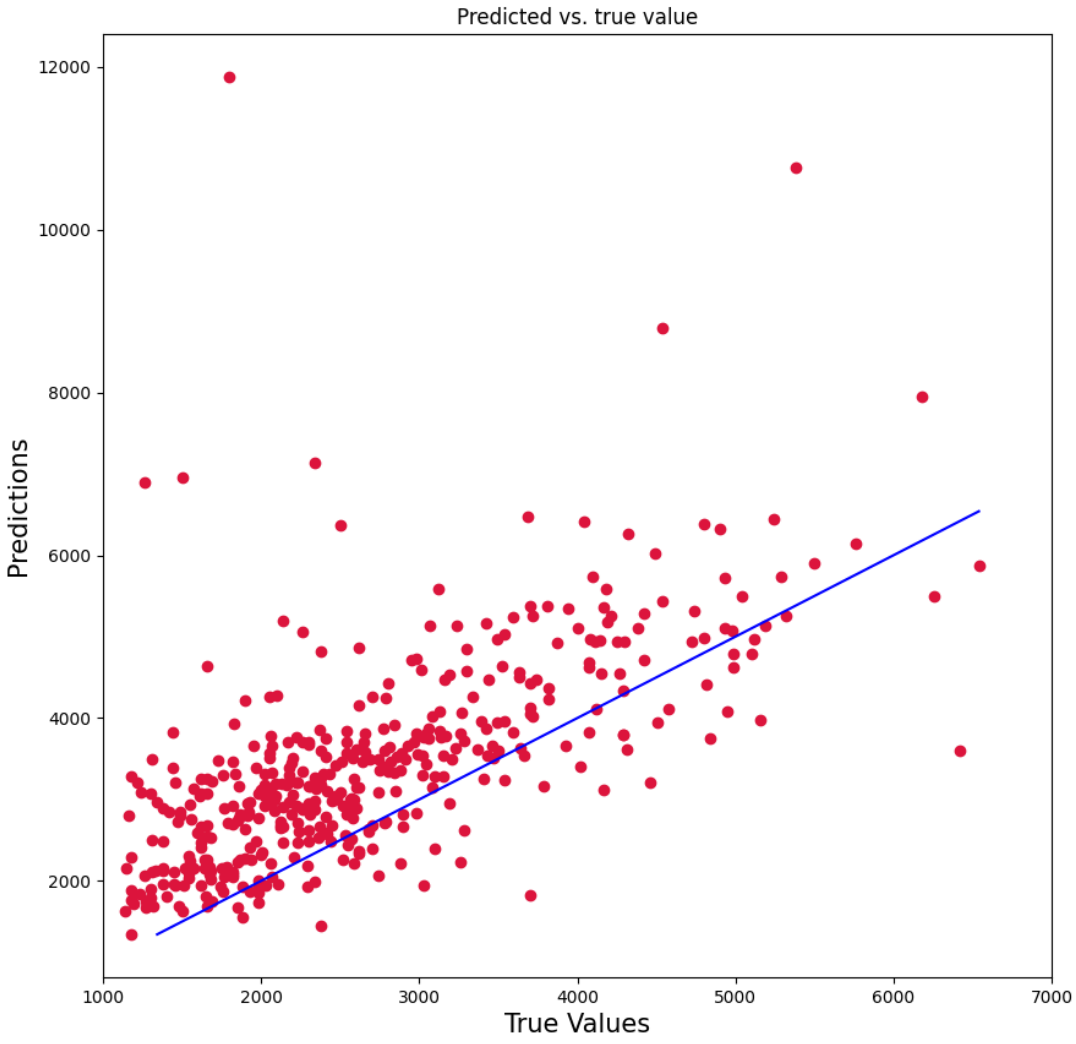



Figure 4.7: Generalization Predictions for Franklin County

# Chapter 5

# Discussion

## 5.1 Cost Accuracy Trade-off

One of the main benefits of the proposed OCR and ML techniques for extracting values from historical records is the ability to scale to large numbers of documents with minimal cost. As a baseline we consider a hypothetical collection of 353,973 each with a single value to extract, which matches the number of properties as in Hamilton County. For details on the following estimates calculations, see Appendix H.

To estimate the cost savings of the OCR methods, we assume scanned documents are available and compare the estimated costs of manual data entry of a single target value against adapting the OCR methods. Manually extracting a single value from scanned documents at the rate we used for the manual labeling process on 353,973 documents will cost an estimated $24,789.22. In contrast, the cost of employing a data scientist to adapt the OCR methods described in this work to a different document will only cost $4,698.12 which is a cost saving of 81%. The drawback for this cost reduction is the reduction in accuracy with an MAPE of 14.72%.

In the scenario where scanned documents are not available, we compared the cost of fully manual scanning and data entry process against the the proposed ML methods. Using online estimates of document scanning services, this will incur an average cost of $35,570.42 for 353,973

documents. Combined with the data entry costs listed previously, this gives a total cost of $60,359.64 for the manual process. We estimate developing the ML model will be comparable to developing the OCR methods and also cost $4,698.12. Combined with an additional cost of $2118.37 to generate 12,423 training samples, which is the number of training samples we used in our experiments, using the manual process gives a total cost of $6816.49 which is a cost saving of 89%. Using this method further reduces the accuracy to an MAPE of 17.48%.

## 5.2 Future Work

Here we present several additional paths to explore for improving the results we presented here and address open questions.

### 5.2.1 OCR of Entire Historical Document

Our OCR methods currently only attempt to extract the initial building value as a proxy of the construction cost, this approach is ignoring a large amount of data present on the ownership documents including the value of the land or other components, changes in the valuation across the years and comments or details about the valuation changes. These data can be extracted by adapting our current segmentation technique to the entire document or use a more sophisticated deep learning based segmentation model to automatically recognize the positions and relationships in the tabular document. With additional time and computational resources, we can evaluate the feasibility of this approach.

### 5.2.2 Additional Contemporary Data and training samples for ML models

One way to improve our ML models is to improve the richness and quantity of the input data. From out analysis, we suspect that our ML models were limited due to missing input signals and the number of training samples we were able to generate using a manual process. For example,

in our case study of estimating home values, additional contemporary data include home improvement permits and contemporary photos of the building. These data would help in bridging the gap between the contemporary information on the building and the changes in building value due to renovations and modifications as well as visual indications of how well the buildings have been maintained through the years. Extending this concept beyond the case study, we expect the performance of ML models to be sensitive to the quantity and quality of input samples so gathering additional training samples and exploring additional sources of input features to be critical to model performance.

### 5.2.3 Deep Learning Models

The ML models we experimented with in this work are relatively simple models and do not take advantage of the latest development in deep learning. We chose to use these simpler models due to the simplicity of our input features and the small number of training samples. Using our case study as an example, a richer set of input features using additional sources of contemporary data along with more samples collected beyond Hamilton and Franklin counties, a more complex deep learning model may prove more successful at predicting the historical value of interest.

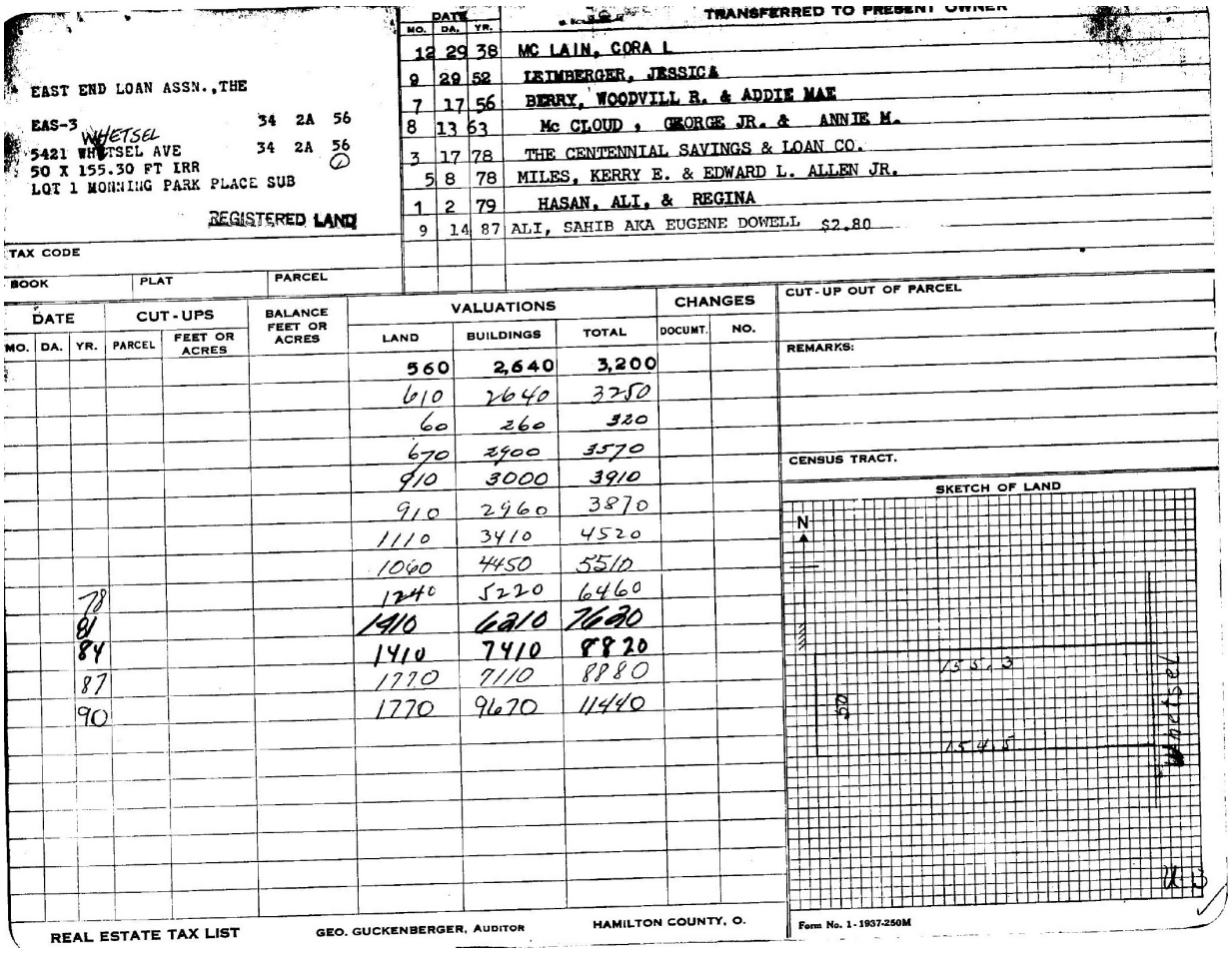
# Chapter 6

# Conclusion

Through this work we are able to show that machine learning and computer vision methods are viable approaches for digitizing data from tabular historical documents with prediction accuracy of 14.72% MAPE and 17.48% MAPE, respectively. We also demonstrate that these methods are cost effective compared to existing manual methods, saving up to 81% with the OCR methods and 89% with ML methods. Though we show the feasibility of augmenting ML model training samples with OCR generated labels, additional work needs to be done to conclusively demonstrate its effectiveness. With potential improvements from expanding the complexity of the ML model, increasing the richness of the ML model inputs and applying our OCR methods on the full historical document, we expect our proposed methods to perform even better, given sufficient time and resources to explore these approaches. We hope our work highlights the benefits of using machine learning and computer vision in unlocking the wealth of data potentially available in historical documents and serves as a guide for how these tools can be practically applied.

# Appendix A

# Sample Hamilton County Ownership Card



EAST END LOAN ASSN.,THE

EAS-3 WHETSEL 34 2A 56
5421 WHETSEL AVE 34 2A 56
50 X 155.30 FT IRR
LOT 1 MORNING PARK PLACE SUB

REGISTERED LAND

TAX CODE

BOOK | PLAT | PARCEL

| DATE MO. | DA. | YR. | TRANSFERRED TO PRESENT OWNER |
|---|---|---|---|
| 12 | 29 | 38 | MC LAIN, CORA L |
| 9 | 29 | 52 | LEIMBERGER, JESSICA |
| 7 | 17 | 56 | BERRY, WOODVILL R. & ADDIE MAE |
| 8 | 13 | 63 | Mc CLOUD , GEORGE JR. & ANNIE M. |
| 3 | 17 | 78 | THE CENTENNIAL SAVINGS & LOAN CO. |
| 5 | 8 | 78 | MILES, KERRY E. & EDWARD L. ALLEN JR. |
| 1 | 2 | 79 | HASAN, ALI, & REGINA |
| 9 | 14 | 87 | ALI, SAHIB AKA EUGENE DOWELL $2.80 |

| DATE MO. | DA. | YR. | CUT-UPS PARCEL | FEET OR ACRES | BALANCE FEET OR ACRES | VALUATIONS LAND | BUILDINGS | TOTAL | CHANGES DOCUMT. | NO. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 560 | 2,640 | 3,200 | | |
| | | | | | | 610 | 2640 | 3250 | | |
| | | | | | | 60 | 260 | 320 | | |
| | | | | | | 670 | 2900 | 3570 | | |
| | | | | | | 910 | 3000 | 3910 | | |
| | | | | | | 910 | 2960 | 3870 | | |
| | | | | | | 1110 | 3410 | 4520 | | |
| | | | | | | 1060 | 4450 | 5510 | | |
| | | 78 | | | | 1240 | 5220 | 6460 | | |
| | | 81 | | | | 1410 | 6210 | 7620 | | |
| | | 84 | | | | 1410 | 7410 | 8820 | | |
| | | 87 | | | | 1770 | 7110 | 8880 | | |
| | | 90 | | | | 1770 | 9670 | 11440 | | |

CUT-UP OUT OF PARCEL

REMARKS:

CENSUS TRACT.

SKETCH OF LAND

N

155.3

50

154.5

REAL ESTATE TAX LIST GEO. GUCKENBERGER, AUDITOR HAMILTON COUNTY, O.

Form No. 1-1937-250M

Figure A.1: Sample Ownership Card

# Appendix B

# Manual Labeling

To ensure we have a reliable set of baseline labels for our models, we used Upwork to find contractor(s) to manually label a subset of our samples at a rate of up to 15 US$ an hour. We provided a total of 12,423 sample for which the first value in the ”BUILDING” column was recorded. Additional information such as the year this value was estimated as well as whether the value was handwritten were also recorded to distinguish whether the building values were the original value estimates from 1933. Finally we sample 1000 of the generated data and verify the correctness ourselves to ensure the accuracy of the labels before using it as ground truth for our models which showed that all manual labels we received were accurate.

# Appendix C

# Cleaning and processing of structured data from Hamilton County

## Step 1: Load data

All raw data files were downloaded from source and placed into a Google Drive folder.

The data files for sources 4.1 and 4.2 were downloaded directly as single CSV or Excel files from the Hamilton County Auditor's site downloads page, linked here. 'Tax Year Information Export' contains the tax assessment information, while both 'Historic Sales' and 'Building Information Export' contain building information.

The data for source 4.3 was downloaded as multiple CSV files from this link, with a separate file for each year of property transfer records from 1998 to 2022. We wrote a Python script to append this information into a single file, handling a change in format between 2006 and 2007. The format change meant adding new columns to the old data (filled with null values), and standardizing column names.

Finally, we wrote a script to pull all the data from the Google Drive into a PostgreSQL database. All further processing happens in the database using SQL scripts.

## Step 2: Fixing basic formatting issues

The first round of cleaning focused on fixing basic formatting and consistency issues. These include:

- Making the parcel identifier (parcelid) consistent across tables. For example, the parcelid had to be manually constructed in the older property transfer files by concatenating book, plat, parcel, and multi-owner (the fields that make up the parcelid) after removing special characters. In other files, parcelids had to be converted to upper case.
- Standardizing NULL values. For example: in property class, null values were captured as two blankspace characters, while in property value the text 'New' was used.
- Optimizing the tables for query performance. We added indices on parcelid and converted string formats to numeric or datetime where possible.

We used this script to implement the cleaning, moving tables from a raw schema to a 'cleaned' schema in the database.

## Step 3: Data quality issues and fixes

Once the basic cleaning was done we performed a more comprehensive data exploration. This raised further issues and inconsistencies which required discussion and decisions on how to handle such cases. These are summarized in Table C.1:

Table C.1: List of data quality issues and resolutions

| **Issue** | **Decision** |
|---|---|
| Property class is captured in multiple tables, with inconsistent values for the same parcel | Use the tax assessment value, because it is the most updated source |
| Some parcels do not merge across tables. E.g., building info has 289 parcelids that don't merge to tax assessment | Drop rows in other tables that don't merge to tax assessment, as it is the most updated source. |
| Parcelids have duplicates because of multiple buildings on a parcel | For now, only analyse parcels with one building. Going forward, reshape data to wide format at parcel level, retaining info about multiple buildings. |
| Some buildings have 0 total square footage | For cases where other square footage fields are nonzero (e.g. floor 1, attic), impute value by summing these up. For buildings where all square footage columns are 0, drop rows because these buildings are torn down. |

# Step 4: Generating features

TODO: talk about types of transformations (collapsing categories, making proportions, etc.)

We used this script to implement the additional cleaning, moving tables from the 'cleaned' schema to 'processed'. The 'processed' schema is the final cleaned data fed as inputs to the modeling pipeline.

## C.1 Standardizing features across Hamilton and Franklin County

TODO: features that were standardized between both counties

# Appendix D

# Segmentation

This task involves recognizing the column header “Buildings” in the image and extracting the bounding boxes of the first cell below it. In this work, we are concerned with extracting the initial construction cost of the building for which we deem the first entry under the ”Buildings” column to be a good proxy.

For the task of locating each cell segment, we begin with TesseractOCR as a baseline to label the bounding boxes for sequences of letters and digits. However, this proved to be difficult since there were many false positives and negatives.

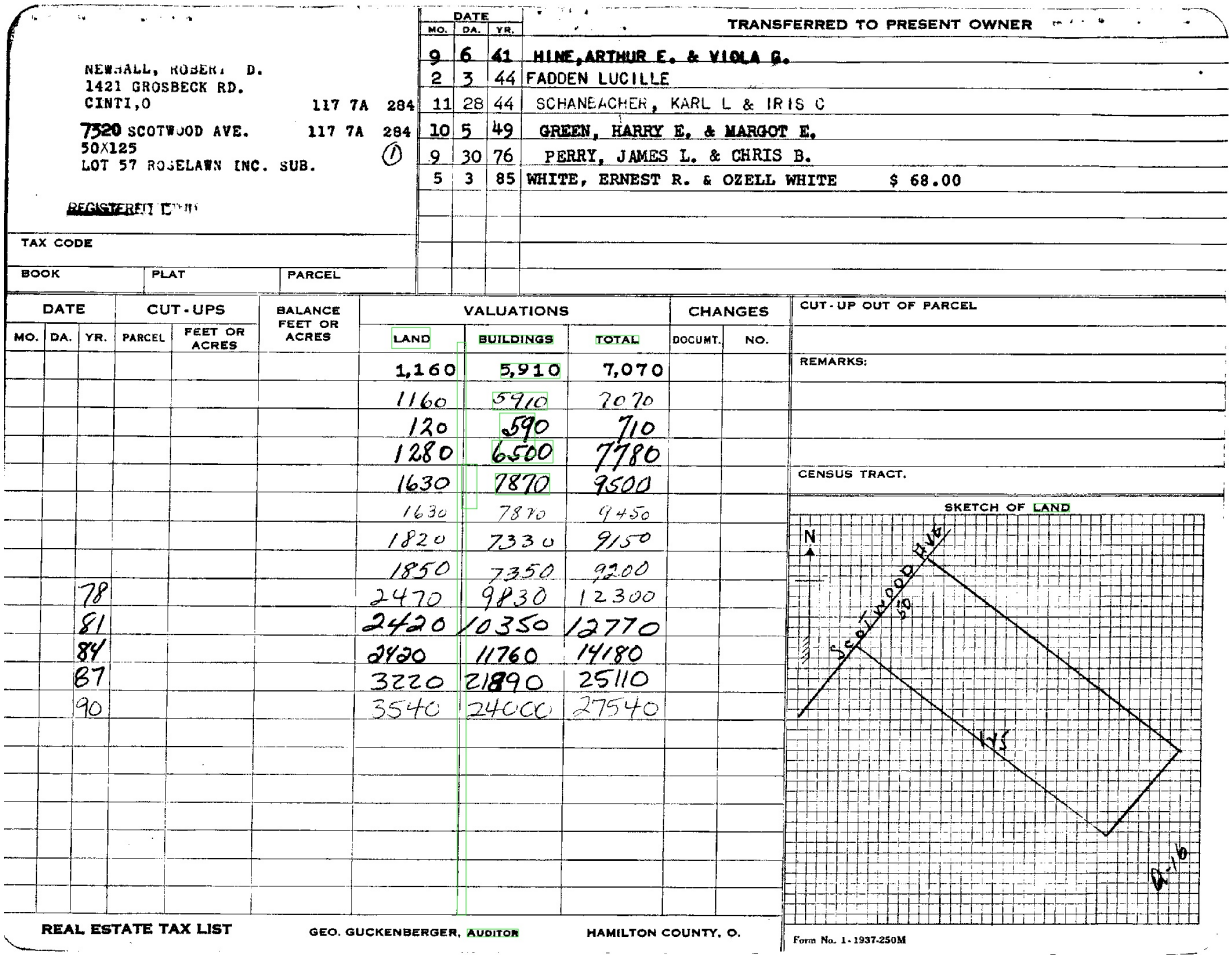



Figure D.1: Sample TesseractOCR Output

Here we can see several issues. First, there are false positives where non digit elements such grid lines being recognized as characters by TesseractOCR. Second there are false negatives where digits further down the column are not recognized. Furthermore, some sequences of characters are not fully recognized. For example only the ”59” of the ”590” sequence is recognized. Finally the recognized characters are not always correct. For example, the first three rows were recognized as ”5,910”, ”SULO” and ”Ff” of which only the first row is correct. Given that TesseractOCR is a pretrained model, we found it difficult to modify its behavior for our particular problem and proceeded with building our own solution.

For the first step, we retain the use of TesseractOCR for locating the ”Buildings” column

header and creating a cropped image around the column header. For example of the cropped document containing the detected column header, see Figure D.2.

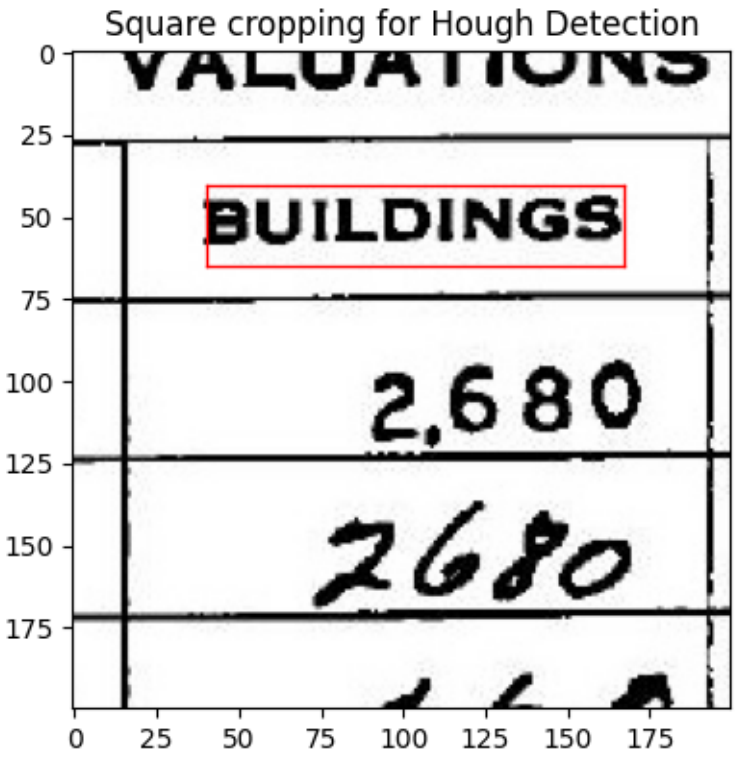



Figure D.2: Sample cropped document

To extract the cells below the header, we then use Hough Transform [7] to detect the main line segments in the cropped image. An example of the document with detected lines overlaid on top is shown in Figure D.3.

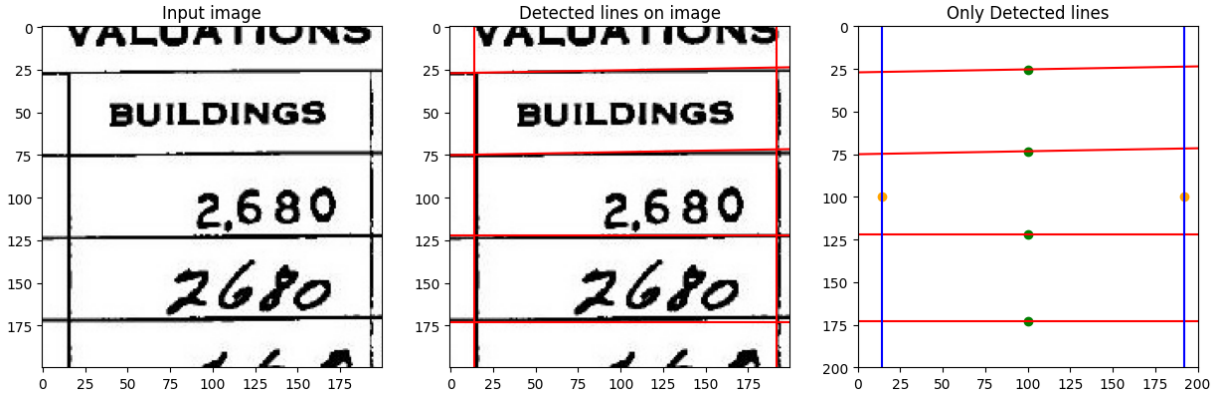



Figure D.3: Sample line detection using Hough Transform

Finally, we use the detected lines and compute the intersections to determine the corners containing the cell we are interested in, which is then used to create a final image of the cell stretched to be a regular rectangle, see Figure D.4.

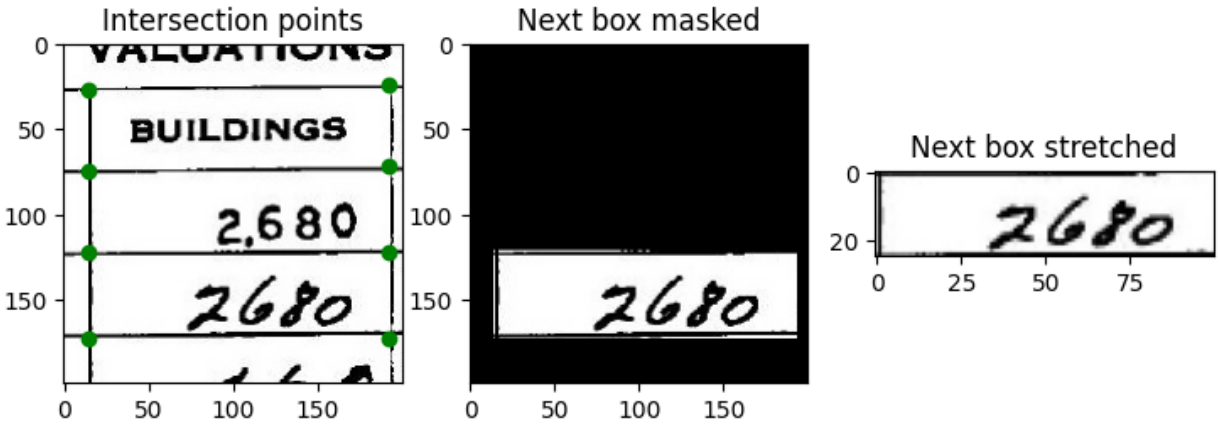



Figure D.4: Extracting a sample cell as a rectangular image

The final output is then ready to be used as an in put to OCR models.

# Appendix E

# OCR models

For the OCR task, we aim to retrieve a numeric value from the segments collected by the process described in the previous section. We experiment with both TesseractOCR and TrOCR to detect numbers and found the results of TrOCR to be significantly better than those obtained with TesseractOCR.

## E.1 TesseractOCR

Our initial experiments with TesseractOCR involved using it for both segmentation and OCR since it outputs the bounding boxes, characters detected as well as its confidence of the predictions. This is promising since it provides all of the required information for constructing a structured output for tabular data. However, we quickly found that TesseractOCR is trained to be a general OCR tool that also recognizes letters and punctuation in addition to the digits that we are interested in and often confuses between them. Furthermore, TesseractOCR performs especially poorly on handwritten digits. As a result, we found that we needed to do significant amount of post-processing to retrieve any meaningful results. Even with all of the processing we were still only able to accurately retrieve the target value in 52.5% of our test cases, see Figure E.1 for the example predictions. Given these poor results we abandoned further work using this

tool for the OCR task.

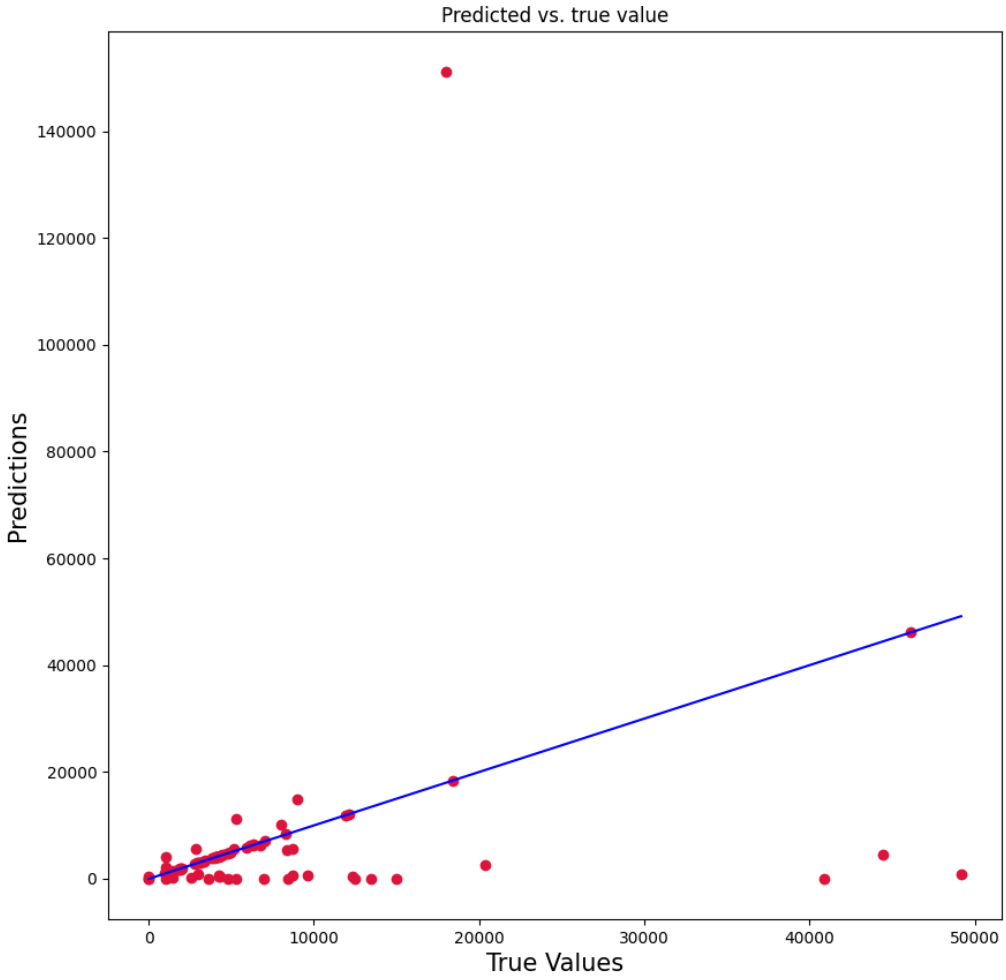



Figure E.1: TesseractOCR predictions

## E.2 TrOCR

Our experiments with the TrOCR model is more successful. While the pre-trained TrOCR model suffers from similar errors as TesseractOCR such as recognizing letters and punctuation in addition to the digits we are interested in, we found that even with minimal fine-tuning on 500

training samples, we can achieve up to 95% exact match in our test set, a drastic improvement over TesseractOCR. Analysing the errors suggested that TrOCR was performing poorly on handwritten digits due to the lack of training samples containing handwriting. To address this deficiency, we combined our training set with the CAR-B dataset [6] of handwritten digit strings from checks to our training samples and surpassed the performance of TrOCR trained on only our dataset or only on CAR-B. A table of the performance of our TrOCR fine tuning experiments is found in Table E.1.

Table E.1: TrOCR Fine-tuning experiments

| Fine-tuning Experiments | Exact match accuracy |
|---|---|
| Our Dataset n=500 (3 iters) | 95% |
| CAR-B n=3k (3 iters) | 4.90% |
| Our Dataset n=5k (3 iters) | 97.17% |
| Our Dataset n=7k combined with CAR-B n=3k (3 iters) | **98.69%** |
| CAR-B n=3k (3 iters) then Our Dataset n=7k (3 iters) | 95.51% |

Further ablation studies on hyperparameters for TrOCR fine-tuning iterations did not yield significant improvements and we selected our best performing experiment as the model used to report our results.

# Appendix F

# Model class selection

Results of a preliminary search for promising model classes to conduct hyperparameter searches on.

Table F.1: Performance of regression model classes (no tuning)

| Model Class | RMSE |
|---|---|
| Poisson Regressor | 1068.42 |
| Random Forest Regressor | 1103.62 |
| Huber Regressor | 1117.24 |
| Gamma Regressor | 1144.69 |
| XGB Regressor | 1226.48 |
| LassoLarsCV | 1229.35 |
| Gradient Boosting Regressor | 1243.21 |
| Lasso | 1255.50 |
| Light GBM Regressor | 1271.28 |
| ElasticNet | 1303.20 |
| Ridge | 1432.39 |
| Linear Regression | 1444.08 |
| Decision Tree Regressor | 1681.67 |
| AdaBoost Regressor | 1681.67 |

# Appendix G

# Feature Importance

TODO: add labels to this plot

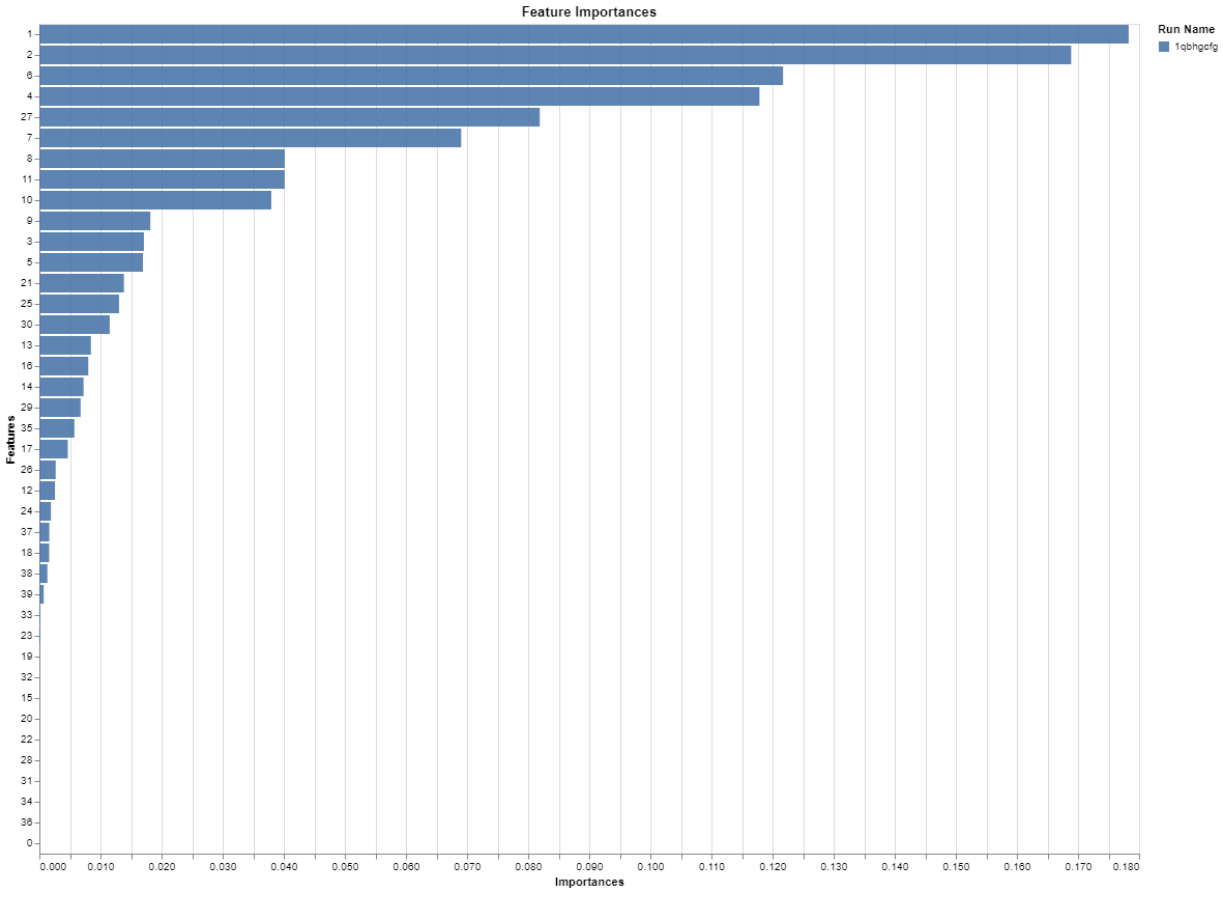



Figure G.1: Feature Importances: ML Model (without OCR augmentation)

# Appendix H

# Cost estimation

We find that most digital record services that offer data entry of specific values in the document to involve two steps, like at Iron Mountain [18]. Typically, the document is first scanned, then OCR or manual entry is performed on the scanned document. This workflow is also used in previous research into historical document digitization [27]. As such we estimate cost of the two steps individually as part of calculations.

For the estimation of scanning 353,973 pages of ownership documents we use the online estimators from two separate services. SecureScan [23] gives a quote of $45,477.80 and ILM Corp [4] gives a quote of $25,663.04, giving an average estimated cost of $35,570.42 or $0.10049 per document.

For the estimation of hiring contractors to extract the initial construction costs from scanned documents we use the same rate as our manual labeling contract on Upwork. In our case, we charged a rate of $15/hr and was able to label 12,423 samples in 58 hours. Extrapolating from this rate to 353,973 gives an estimated cost of $24,789.22 or $0.07003 per document.

We then estimate the cost of developing the OCR and ML models. Considering the time to develop the two proposed models were comparable and required one 14-week semester of work at an estimated 12 hours per week, it took about 84 hours to develop each individual model. Using an estimate of an average Data Scientist salary of $55.93 from Indeed.com [14], we estimate the

cost of developing each model at $4698.12. For developing the ML model, additional costs need to be included for generating the training labels, potentially from physical documents that are not scanned. For our model, we collected 12,423 training samples. Using the scanning and data entry costs per document listed above this would add an additional $2118.37 to the development of the ML model.

# Bibliography

[1] Hamilton County Auditor. Hamilton county auditor: Real estate tax valuation. https://www.hamiltoncountyauditor.org/revalue.asp, 2023. Accessed: 2023-04-23. 2.1.2

[2] Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdez, and Carlos Afonso. Identifying real estate opportunities using machine learning. *Applied sciences*, 8(11):2321, 2018. 1.1

[3] Callum Booth, Robert Shoemaker, and Robert Gaizauskas. A language modelling approach to quality assessment of OCR'ed historical text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5859–5864, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.630. 1.1

[4] ILM Corp. ILM Corp cost of document scanning. https://www.ilmcorp.com/tools-and-resources/cost-of-document-scanning/, 2023. Accessed: 2023-04-10. H

[5] Sergio Correia and Stephan Luck. Digitizing historical balance sheet data: A practitioner's guide. *Explorations in Economic History*, 87:101475, 2023. ISSN 0014-4983. doi: https://doi.org/10.1016/j.eeh.2022.101475. URL https://www.sciencedirect.com/science/article/pii/S0014498322000535. Methodological Advances in the Extraction and Analysis of Historical Data. 1.1

[6] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M. Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S. Oliveira. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsrc 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 779–784, 2014. doi: 10.1109/ICFHR.2014.136. 3.1.2, E.2

[7] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, jan 1972. ISSN 0001-0782. doi: 10.1145/361237.361242. URL https://doi.org/10.1145/361237.361242. D

[8] James J. Feigenbaum. Automated census record linking: A machine learning approach. Accessed: 2023-04-23, 2016. 1.1

[9] Pascal Fischer, Alen Smajic, Giuseppe Abrami, and Alexander Mehler. Multi-type-td-tsr–extracting tables from document images using a multi-stage pipeline for table detection and table structure recognition: From ocr to structured table representations. In *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44*, pages 95–108. Springer, 2021. 1.1

[10] Michel Fouquin and Jules Hugot. Two centuries of bilateral trade and gravity data: 1827-2014. Accessed: 2023-04-23, 2016. 1

[11] Winky KO Ho, Bo-Sin Tang, and Siu Wai Wong. Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1):48–70, 2021. 1.1

[12] Junia Howell. Reimagining equity: Towards an equitable economic model for social housing. Vienna International Summer School on Social Housing Production, 2022. URL https://iba-researchlab.at/summer-school-2022/. 1

[13] Junia Howell and Elizabeth Korver-Glenn. The Increasing Effect of Neighborhood Racial Composition on Housing Values, 1980–2015. *Social Problems*, 68(4):1051–1071, 09 2020. ISSN 0037-7791. doi: 10.1093/socpro/spaa033. URL https://doi.org/10.1093/

socpro/spaa033. 1

[14] Indeed.com. Indeed data scientist salary in united states. https://www.indeed.com/career/data-scientist/salaries, 2023. Accessed: 2023-04-13. H

[15] Huseyin Kusetogullari, Amir Yavariabdi, Johan Hall, and Niklas Lavesson. Digitnet: A deep handwritten digit detection and recognition method using a new historical handwritten digit dataset. *Big Data Research*, 2020. 3.1.2

[16] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2021. URL https://arxiv.org/abs/2109.10282. 1.1

[17] Jiří Martínek, Ladislav Lenc, and Pavel Král. Training strategies for ocr systems for historical documents. In John MacIntyre, Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 362–373, Cham, 2019. Springer International Publishing. ISBN 978-3-030-19823-7. 1.1

[18] Iron Mountain. Iron Mountain document scanning & digital storage services. https://www.ironmountain.com/services/document-scanning-and-digital-storage#howitworks, 2023. Accessed: 2023-04-13. H

[19] Jonas Mueller-Gastell, Marcelo Sena, and Chiin-Zhe Tan. A multi-digit ocr system for historical records (computer vision). Accessed: 2023-04-13, 2020. 1.1

[20] Smita Pallavi, Raj Ratn Pranesh, and Sumit Kumar. A conglomerate of multiple OCR table detection and extraction. *CoRR*, abs/2010.08591, 2020. URL https://arxiv.org/abs/2010.08591. 1.1

[21] Joseph Price, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. Combining family history and machine learning to link historical records. Technical report, National Bureau of Economic Research, 2019. 1.1

[22] Ashish Ranjan, Varun Nagesh Jolly Behera, and Motahar Reza. *OCR Using Computer Vision and Machine Learning*, pages 83–105. Springer International Publishing, Cham, 2021. ISBN 978-3-030-50641-4. doi: 10.1007/978-3-030-50641-4_6. URL https://doi.org/10.1007/978-3-030-50641-4_6. 1.1

[23] Secure Scan. Secure Scan document scanning price calculator. https://www.securescan.com/document-scanning-price-calculator/, 2023. Accessed: 2023-04-10. H

[24] R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991. 1.1

[25] Dieudonné Tchuente and Serge Nyawa. Real estate price estimation in french cities using geocoding and machine learning. *Annals of Operations Research*, pages 1–38, 2022. 1.1

[26] Bogdan Trawiński, Zbigniew Telec, Jacek Krasnoborski, Mateusz Piwowarczyk, Michał Talaga, Tedeusz Lasota, and Edward Sawiłow. Comparison of expert algorithms with machine learning models for real estate appraisal. In *2017 IEEE international conference on innovations in intelligent systems and applications (INISTA)*, pages 51–54. IEEE, 2017. 1.1

[27] Amir Yavariabdi, Huseyin Kusetogullari, Turgay Celik, Shivani Thummanapally, Sakib Rijwan, and Johan Hall. Cardis: A swedish historical handwritten character and word dataset. *IEEE Access*, 10:55338–55349, 2022. doi: 10.1109/ACCESS.2022.3175197. 1.1, H

[28] Yun Zhao, Girija Chetty, and Dat Tran. Deep learning with xgboost for real estate appraisal. In *2019 IEEE symposium series on computational intelligence (SSCI)*, pages 1396–1401. IEEE, 2019. 1.1