Lessons from a human-in-the-loop machine learning approach for planning: identifying vacant, abandoned, and disinvested properties in Savannah, Georgia.

Abstract

Addressing issues with vacant, abandoned, and disinvested (VAD) properties is important for maintaining a healthy community housing stock. Yet, the process of identifying these properties can be difficult. Here, we create a human-in-the-loop machine learning model and apply it to a parcel-level case study in Savannah, Georgia. Our model reveals that tax delinquency, code violations, and crime history best predict VAD properties, and it identifies differences between machine vs. human generated results. The resulting model uses expert knowledge and statistical learning to streamline the process of identifying and managing VAD properties and subsequently, helps staff design comprehensive plans.

Introduction

Over time, neglected structures and land in communities degrade into unusable or unsafe infrastructure. This process adversely affects public health and welfare and can increase crime and lower property values. To protect and support growth in communities, local jurisdictions try to identify these vacant, abandoned, and disinvested (VAD) properties. However, identifying VAD properties, i.e., recording their location, is not a simple task: there is no national database nor a standardized definition with which to detect VAD properties, and as such, cities typically rely on block-by-block field surveys to count VAD properties (Mallach 2018). Some cities have attempted to use advanced techniques like spatial decision support systems (SDSS) and machine learning models to find VAD properties from large datasets, but in discussion, emphasize the technology itself rather than the combination of human, technology, and data (Hill et al. 2003; Appel et al. 2014; Hillenbrand 2016; Reyes et al. 2016).

In this work, we build a human-in-the-loop machine learning model (HILML) for the City of Savannah, Georgia, and work with housing officials from Savannah's Housing and Neighborhood Services Department (HNSD) and the Chatham County / City of Savannah Land Bank Authority (LBA) to identify VAD properties for government acquisition and eventual redevelopment as affordable housing. In the model, each property is associated with civic variables (such as code violations, tax, and crime etc.) collected using census and municipal data records stored in GIS and databases. The goal is to predict whether a property is VAD, through a 'learned' combination (calibrated by human insights) that indicates which variables best predict VAD status, thus integrating human expertise and 'common sense' thinking into a computational process.

The research questions at hand are centered on model prediction power and real-world utility. First: What combination of a-priori variables best predict whether a property should be classified as VAD? Second: How does a human-in-the-loop machine learning approach improve (or not improve) the city's current methods to identify VAD properties?

The resulting model correctly predicted labels (i.e., VAD or Not VAD) for 5,327 parcels with about 90% confidence on structureless parcels and 87% confidence on parcels with structures. The number of years and lost value of unpaid taxes associated with the property were the most important variables for predicting VAD status, followed by the number of crimes and number of code violations. We found a difference in what the ML model and the human experts would classify as VAD: machine-generated predictions covered a wider range of neighborhoods beyond the scope of human experts' field survey (i.e., driving to a few target neighborhoods with many highly ranked VAD properties and rate properties based on visual cues that align with legal blight criteria). The model also reports more VAD properties that are tax delinquent, prone to crime, and in neighborhoods with various income levels, while human experts tend to identify properties with visual cues (i.e., dilapidated roofs, boarded windows) and low property value (i.e., in low-income neighborhoods).

This research has three specific features that distinguish it from existing ML-based parcel-finder models in the literature: 1) a human-in-the-loop approach to train the model with "active learning" techniques that incorporate local experts' tacit knowledge, 2) interpretable ML methods to find discrepancy in expert-labeled data and build consensus, and 3) empirical analyses of human and machine biases in identifying VAD properties to prompt further actions. As such, a contribution of this work is a method for helping mitigate and improve the condition of municipal properties and preempt deterioration that adversely affects neighborhoods and residents. The method provides insights on the property level rather than the neighborhood level where traditional analyses based on Census data often lack granularity. Another contribution of this case study is to exemplify the development of a civic technology that balance the pursuit of efficiency as well as collaboration, context, and bias awareness. While our model's results may not be generalizable to all communities, the process of academic-practitioner collaboration and training a model using tacit knowledge and decentralized databases can serve as a guide for planning practices.

The manuscript proceeds as follows. We first review literature on property vitality and intervention. We then describe our case study, dataset, and methods. Then, we report on results, followed by a discussion on *'success stories' and 'lessons learned'.* Although this case study is centered on one municipality, we describe how our results can be extended and implemented in other locales.

Background / Literature Review

Characteristics of VAD Properties

Vacant, abandoned, and disinvested (VAD) properties are properties (parcels) that exhibit physical deterioration and neglect, which often negatively impacts the surrounding area. To our knowledge, there is no known universal definition of VAD (or blighted) properties, likely because such properties are often described to suit a specific purpose: raising awareness for substandard low-income housing, supporting massive demolition for urban renewal, revitalizing downtown through eminent domain, or revealing the impacts of mortgage crises (Schilling and Pinzon

2016). In some of these usage contexts, VAD properties are associated with “blight” which has been criticized as a stigmatizing term for properties and their neighborhoods (Mallach 2018), so we use the term VAD. VAD properties often have more than one of the following features: vacancy, property code violations, tax delinquency (Hillier et al. 2003), and are the site of public nuisances (e.g., drug crime or fire damage) (Schilling et al. 2015). One feature alone may not imply VAD status: for example, a house can be vacant but still well-maintained or delinquent in tax but occupied by low-income families. Therefore, the definition of VAD must be examined in a local context. VAD properties may also have lower property values compared with those in the same block, as found in research comparing sales price of foreclosure properties and the neighboring occupied units (Sumell 2009; Whitaker, Stephan, and Fitzpatrick 2013). Yet in practice, property value can be slow to update and vary significantly based on property attributes and thus not a stable indicator for VADs.

Causes and Impacts of VAD properties

A property can become disinvested as a result of macroeconomic and demographic shifts, a housing market failure, and public and legal policies. In the U.S., historically, an increasing number of properties have faced degradation since 1960s due to the demographic shifts in the inner city and the cost to maintain housing (e.g., mortgages) exceeding its value. This trend peaked after the housing bubble bust in 2006 and 2007 (Mallach 2018). Many legacy cities (i.e., post-industrial shrinking cities) with sustained job and population loss over the past decade have struggled to uplift neighborhoods with concentrated vacancy (Mallach, Alan, and Brachman 2013). At the local level, VAD properties with cloudy titles (i.e., unclear ownership status) or delinquent taxes and assessments exceeding the value of the property can be characterized as ‘dead to the housing market’, because they have high transaction costs and minimum values.

VAD properties can disproportionally impact marginalized communities’ local housing markets by decreasing neighboring property values and impacting neighbors’ quality of life (Mallach 2018; Whitaker, Stephan, and Fitzpatrick 2013). Neighborhoods with many VAD properties are associated with poor school quality (Sun et al. 2019), high crime rates (Branas et al. 2012), higher male unemployment rate (Appel et al. 2014), and slower growth in property sales price (Gilreath 2013). These neighborhoods are also more likely to be home to low-income and African American households (Sun et al. 2019; Silverman et al. 2013) and suffer from declining home ownership and pessimistic perceptions of neighborhood trajectories (Mallach 2021). Geographically, VAD properties also tend to cluster (Weaver, Russell, and Bagchi-Sen 2013; Hillier et al. 2003; Reyes et al. 2016), reinforcing the concentration of income, race, and housing market inequality in marginalized communities and stunting economic mobility. VAD properties also impose fiscal burdens on the city including millions of USD in lost tax revenue and unrecoverable costs of managing overgrown grass, litter, and illegal dumping; securing open structures; and demolishing properties (Sumell 2009; Mallach 2018; Immergluck 2016).

From Spatial Decision Support Systems to Human-in-the-loop Machine Learning

Early examples of identifying and managing VAD properties through technology have used spatial decision support systems (SDSS). SDSSs combine spatial and aspatial data with analytical models and geovisualization to facilitate decision-making in a spatial context (Hopkins et al 1985;

Armstrong et al 1986). Such systems typically have three components: spatial database infrastructure, a library of models, and an interface to visualize spatial outcomes and make decisions (e.g. where to distribute funds across the city) (Keenan, Peter Bernard, and Jankowski 2019).

SDSSs and related analytical models have helped cities address VAD properties. The Philadelphia Neighborhood Information System is an early example of using statistical models (in this case, logistic regression) in SDSS for blight remediation purpose (Hillier et al. 2003). The system was designed to integrate housing information and web mapping, and to leverage predictive analytics to identify likely abandoned properties (Hillier et al. 2003). The City of New Orleans subsequently developed a decision support scorecard system using logistic regression to recommend sale (on the market or privately) or demolition to city officials based on local experts' scoring of multiple criteria related to the property's condition (Hillenbrand 2016). A similar study in the City of Youngstown, Ohio overlayed multiple factors in the ArcGIS Geographic Information Systems (GIS) environment to prioritize demolition based on the highest sum values of property scores (Morckel 2016). Others have used various machine learning models (random forest, decision tree, gradient boosting, etc.) to predict the VAD properties and discover how each factor contributes to the prediction (e.g., City of Syracuse in Appel et al. 2014; City of Cincinnati in Reyes et al. 2016). Yet, most previous case studies have not involved local experts when adapting ML models nor when creating new training data independently from historical records (Hill et al. 2003; Reyes et al. 2016; Appel et al. 2014) (Hillbrand 2016 is an exception). In addition, few studies discuss the potential ethical concerns in the input data (Reyes et al. 2016 is an exception).

Human-in-the-loop machine learning (HILML) emphasizes combining human insights with the statistical data-driven model to increase model efficiency and accuracy, as well as visibility, explainability, trustworthiness, and transparency (Holzinger 2016; Zhou and Chen 2018). The traditional supervised machine learning process involves training a model on labeled datasets and running it through multiple probability and statistical tests to predict unlabeled data. In contrast, HILML highlights active human involvement in all aspects of the machine learning process, including strategies to create training data through human labeling (which is a preferred method of labeling), intelligent selection of samples and features, and explainable mechanisms (Monarch 2021), and thus best suited for creating unbiased predictions on complex and rare events, and for addressing societal problems. Few studies have operationalized HILML to understand planning issues (see Zheng, Zhibin and Sieber 2021 for topic modeling on smart city grant proposal text and Anwar 2022 for land cover mapping), but we see value in incorporating ground truthing and tacit expert knowledge within the data-driven model to help the model adopt site-specific knowledge and to allow experts to shape their tools.

Civic Technology and Data for Good

The human-in-the-loop approach speaks to the rising civic technology movement that outlines the power of technology to govern and serve as a voice for communities (Boehner, Kirsten, and DiSalvo 2016; Le Dantec 2016). Yet concerns arise as a machine learning model 'trained' on historically biased data may yield predictions that reinforce these biases and disproportionally impact minoritized groups (Eubanks 2018; D'ignazio, Catherine, and Klein, 2020). In response, feminist and action researchers have called for the use of data and technology that respect local contexts, challenge power, and generate public discourses, through activities such as collaborating with (local) expert teams, articulating issues, building and ground-truthing data, and interrogating the contexts of data production (Le Dantec 2016; Loukissas 2019; D'ignazio, Catherine, and Klein, 2020; Williams 2020). These principles correspond with the collaborative and communicative planning traditions that recognize power relations, consensus-building, and diverse interests in planning practices (Forester 1982; Innes 1995). Our human-in-the-loop machine learning approach attempts to adopt these values, which we detail in the method section.

Case Study

The City of Savannah is a historic city (city pop 150,000), popular tourist destination, and home to the Savannah Port, which is one of the busiest seaports in the United States (International Trade Administration, n.d). Savannah is classified by the Lincoln Institute of Land Policy (n.d.) as a legacy city (i.e., post-industrial shrinking city) whose population peaked in the 1960s. The city has a growing number of residential properties that are likely to qualify as VAD, as measured through code violations and tax delinquency. In 2019, there were 5,372 candidate properties, within which 1,319 properties with code violations indicate severe VAD conditions and 1,404 properties with at least three years of tax delinquent history. These properties tend to be in minority-concentrated neighborhoods with sizable Black population. In Savannah, each VAD property costs the city $1,300 USD annually for maintenance costs and loss of property tax revenue. By 2019, the city had 4.7 million USD of uncollected taxes from the 5,372 potentially VAD residential properties.

In 2019, Savannah allocated $10 million USD to acquire and redevelop 1,000 VAD properties over the subsequent 10 years to repair and redevelop VAD properties into affordable housing in neighborhoods that have been long neglected or exploited by profit-driven investors (Housing and Neighborhood Service. n.d.).

Traditionally, Savannah's HNSD and LBA staff identified potential VAD candidates for redevelopment through tacit knowledge or during tax sales or foreclosure events. As such, only dozens of properties are acquired every year. Since 2018, the HNSD has used a data-driven approach, by acquiring spreadsheet data from civic departments (e.g., police, code compliance, county/city revenue office), examining the parcel information visually on an online platform, averaging the score of each parcel's VAD features, and conducting a field survey to augment the data. Yet, the process of acquiring, cleaning, mapping, and analyzing data across the entire city was labor-intensive and time-consuming, and the data was limited to a snapshot of conditions (personal communication redacted for peer review). The scoring system is also prone to error because some features (e.g., crime) are only significant if they are combined with others (e.g., tax delinquency). The city wanted to engage researchers to design a VAD-identification system that is not only transparent, scalable, and sustainable but also considers the local context and interactions between variables.

Data and Variables

Independent Variables in the Machine Learning Model – VAD characteristics

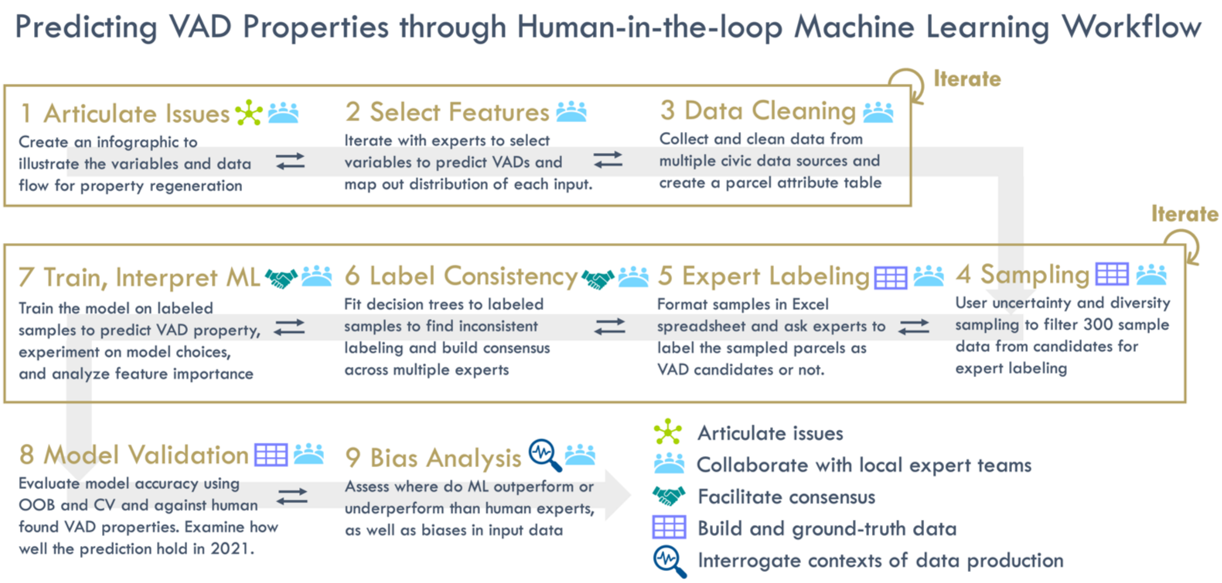
To develop the HILML model, we used data sources under seven themes: crime, code compliance, fire incidents, tax delinquency, vacancy status, building attributes, and market indicators (see Table 1). All data are from 2010-2019, except for code violations (from 2012) and are used to characterize a parcel-level dataset within the City of Savannah. We acquired most data through HNSD (see Table 1). Median neighborhood property value, five-year growth rate, and ratio variables are derived from existing variables in the dataset. We acquired parcel shapefiles, flood zone boundaries, and neighborhood boundaries from Savannah Area GIS Open Data (n.d.). Vacancy probability is compiled from a mix of USPS-generated records, field survey, and specific code violations (see S.I. Section B). Data with temporal records (e.g., crime, tax, fire, and code violations) are weighted: an incident is weighted higher if it is more recent and the type particularly contributes to VAD (e.g., drug crime or code violation that indicates unsafe structure, see S.I. Section B for details).

Table 1: Independent variables (or features) per parcel fed into the machine learning model for land or structure. See S.I. Section B for variables used in labeling and extra processing details.

Variable	Year	Description	Source
Weighted Crime	2010-2019	The number of crime incidents weighted by recency and type.	Police Department
Weighted Drug Crime	2010-2019	The number of drug crime incidents weighted by recency.	Police Department
Weighted Active Code Violation	2012-2019	The number of active code violations weighted by recency and type.	Code Compliance
Weighted Fire Incidents	2010-2019	The number of fire incidents weighted by recency.	Fire Department
Delinquent Tax	2010-2019	Total amount of delinquent city and county tax and unpaid special assessment. The unit is dollar.	County Tax Office City Tax Office City Revenue Office
Total Delinquent Years	2010-2019	The number of years that the property has tax delinquency or unpaid special assessment.	County Tax Office City Tax Office City Revenue Office
Unpaid Special Assessment Tax Pct	2010-2019	The percentage of unpaid special assessment in total delinquent tax.	Derived
Vacancy Probability	2019	The probability that the property is vacant. See S.I. section B for more details.	USPS Service Records; Field Survey
Land Size	2019	The size of land. The unit is acre.	Assessor
Qualified Sales Count	2010-2019	The count of qualified sales, which often measures title transfers for properties under market value.	Assessor
Unqualified Sales Count	2010-2019	The count of unqualified sales. The sales can occur due to heir inheritance or foreclosure and often sold below market values.	Assessor
Year Last Sold	2010-2019	The most recent year that the property was sold	Assessor
Property Value	2019	Property values estimated by computer assisted mass appraisal (CAMA). The unit is 1,000 dollars.	Savannah GIS Open Data Parcel Shapefile
Five-year Growth	2014, 2019	Five-year growth in property value, calculated as (CAMA_2019 – CAMA_2014)/CAMA_2014*100. The unit is percentage.	Derived
Median Neighborhood Property Value	2019	The median neighborhood value of land or structure in the neighborhood. The unit is 1,000 dollars.	Derived

Human-in-the-loop Machine Learning (HILML)

In the following subsections, we outline how the values of civic technology for good, including *articulating issues*, *collaborating with local expert teams*, *building and ground-truthing data*, *facilitating consensus*, *and interrogating the contexts of data production,* are operationalized in the human-in-the-loop machine learning workflow (see Figure 1).


Predicting VAD Properties through Human-in-the-loop Machine Learning Workflow

1 Articulate Issues: Create an infographic to illustrate the variables and data flow for property regeneration

2 Select Features: Iterate with experts to select variables to predict VADs and map out distribution of each input.

3 Data Cleaning: Collect and clean data from multiple civic data sources and create a parcel attribute table

4 Sampling: User uncertainty and diversity sampling to filter 300 sample data from candidates for expert labeling

5 Expert Labeling: Format samples in Excel spreadsheet and ask experts to label the sampled parcels as VAD candidates or not.

6 Label Consistency: Fit decision trees to labeled samples to find inconsistent labeling and build consensus across multiple experts

7 Train, Interpret ML: Train the model on labeled samples to predict VAD property, experiment on model choices, and analyze feature importance

8 Model Validation: Evaluate model accuracy using OOB and CV and against human found VAD properties. Examine how well the prediction hold in 2021.

9 Bias Analysis: Assess where do ML outperform or underperform than human experts, as well as biases in input data

Iterate
Iterate
Articulate issues
Collaborate with local expert teams
Facilitate consensus
Build and ground-truth data
Interrogate contexts of data production

Figure 1: Human-in-the-loop machine learning workflow. Steps 1-3 and steps 4-7 are iterated multiple times with local experts. The icons represent which 'civic technology for good' values (and human interactions) are adopted in the process.

Articulating Issues

We first sent a Q&A document to the HNSD experts to learn about the process of property regeneration (i.e., the issues) from the administration perspective and worked with them at their offices in Savannah and online to codify their process of identifying VAD properties and then 'regenerating' them for affordable housing (see S.I. Section C for infographic). From the standpoint of HNSD, the biggest challenge is managing the decentralized data, that flows between institutions, and leveraging the data for effective decision-making.

Feature Selection and Sampling

We created maps of Savannah symbolized by variables listed in Table 1 and examined them in person with the experts to rule out potential anomalies (see S.I. Section D). This discussion resulted in 5,372 residential structures (single-family, 2-4 family, and townhouse) and land that may be vacant, have no flood risks, and with records in drug crime, tax issues, code violations, and fire incidents.

We deployed a mix of random sampling and active learning techniques (uncertainty and diversity sampling) to create the training samples for expert labeling that can maximize the learning of the model. Uncertainty sampling selects data points that are close to decision boundaries and thus have maximum uncertainty (Lewis, David, and Catlett 1994), while diversity sampling rebalances the ratio of data points to amplify rare events (Monarch 2021). We conducted three rounds of sampling, each with 100 samples (a total of 300 samples). The first two rounds emphasized geographic diversity (e.g., properties from various racial-majority neighborhoods) and the representations of various types of VAD properties (properties from clusters identified through the K-means algorithm) in the samples, while the third round generated random samples (see S.I. Section D).

Expert Labeling, and Label Consistency

We asked a team of four housing experts to classify a subset of properties (n=300) as VAD or not; this serves as a dependent variable for the model's training data. The experts were two males and two females ranging in age from their 30s to 65. They have experience in the field of land and housing for 25 years, 15 years, 12 years and 3 years and have been working for the City of Savannah for 8 years, 3 years (x2), and 2 years. They had no prior experience in ML nor in labeling training data.

In July 2021, we sent each expert a spreadsheet of properties and their characteristics (see Table 1). In the spreadsheet, each column is a parcel, and each row is a parcel attribute. The experts could choose labels (i.e., VAD and Not VAD) from a dropdown cell and write comments to clarify how they made the decisions. To improve interpretability of the variables for human labeling, we disaggregated the weighted variables by their components (see S.I. Section D). One expert labeled 150 samples, and the others labeled 50 samples, respectively.

After the labeling, we asked experts via e-mail to discuss labels that either countered our instincts or were deemed as highly uncertain by the ML algorithm in Python modAI package. This process helped expose nuances in the interactions of variables, differentiate whether false labels were due to low sample points or human error, and uncover tacit variables being used in the reasoning.

We found that properties with very similar conditions were labeled differently, presenting opportunities to highlight implicit assumptions in the labeling process and to streamline the method to ensure equity across the process. To find these properties, we fitted decision trees to labeled data of lands and structures respectively and visualized the branching. For example, if one (or a small set of) VAD property took multiple splits to be separated from a large group of non-VAD properties, then it (they) was a good candidate for discussion (see S.I. Section D for visualization). When a property was labeled differently by the experts, we presented the decision trees and labeled samples to the experts to agree upon a resolution.

Machine Learning, Model Validation, and Bias Analysis

We used a *random forest algorithm* to classify properties into VAD and Not VAD properties. The random forest (RF) model captures nonlinear relationships between variables by creating a "forest" of decision trees in which each tree decides on a sequence of variable and variable values to split the data so that it minimizes label differences in each branch (Liaw, Andy, and

Wiener 2002). We chose a random forest model because it has higher model accuracy than logistic regression and the VAD status depend on a combination of variables (e.g., crime with tax delinquency); random forest can capture such feature interactions (Basu et al. 2018). Since feature correlation impacts feature importance scores in the random forest model, we ran Spearman's rank-order correlation table between all feature pairs. Lastly, we tuned the hyperparameters of the random forest model (see S.I. Section E).

In total, we ran eight random forest models. We used Python (with scikit-learn) for machine learning tasks and R (with tidyverse, sf, and tmap) for data wrangling and mapping. These models are differentiated by 1) whether they had full or reduced features, 2) whether they were trained with 200 or 300 labeled samples, and 3) whether they were trained on lands or structures. To reveal what variables were most instrumental at predicting the VAD properties, we reported drop-column feature importance score for each feature at each model. The drop-column importance method is favored over the default importance score method in RF (i.e., mean decrease in impurity) because our features have different scales of measurement (Strobl et al. 2007). A high drop-column importance score means the accuracy of the model dropped significantly without that feature. We use partial dependence plots to show how well each feature predicts outcomes (see S.I. Section E).

To evaluate the models, we used *5-fold cross-validation*, which is the average accuracy of five different training and test sets, with a ratio of 80-20 split. We also use the *out-of-bag (OOB) score*, which is the average accuracy of sample data predicted by decision trees in the random forest without these samples. Both metrics report the percentage of samples that are not in the training set and were correctly predicted.

To further validate the robustness of the final model predictions against human-identified VAD properties, we compared the model predictions with 1) the geographic distribution and VAD types (see S.I. Section F for criterion of VAD types) of human-generated VAD targets (only in a few neighborhoods) in 2019 and 2) human judgements (by HNSD experts) through manually checking all records of 100 predicted properties in 2021. As such, we could better understand how well the predictions trained on 2019 data applied to 2021 data and whether extra contextual information about the properties (e.g., field visits) contributed to more holistic assessments. We also identified other sources of biases in the machine learning process by questioning the context of data production for all the features (see S.I. Section F). Specifically, we mapped out how certain neighborhoods have low percentage of code cases from 311 calls (a major source of input for the code violation dataset) and ran linear regression to reveal neighborhood characteristics that correlate with few 311 calls.

Results

Important features in trained models

We report the drop-column feature importance score of all features in eight random forest models in Table 2 and find that tax (total delinquency and delinquent years) are the most consistently important variables in predicting VAD properties across eight models, followed by crime and active code cases. For example, when using model S4 to predict vacant land, including total tax delinquency and delinquent year as a variable improves model accuracy by 13.3%. Drug crime and the percentage of special assessment tax in total tax delinquency positively contribute to model accuracy in some models, but not the others.

This outcome is aligned with the feedback we received from the experts, as they identified variables related crime, code violations, tax issues, and low property value as key to their decisions. Other variables' role in predicting VAD properties may be contextual. For example, property value is not a necessary condition for VAD property, but a VAD property with low property value is particularly attractive to the city as it indicates low acquisition cost. The experts' interpretation was important because property value did not significantly improve model accuracy in any model, potentially because the interpretation of a "low" property value is relative to housing types, neighborhoods, and land size.

Other variables (e.g., unqualified sales, growth rate etc.) that indicated VAD conditions, in theory, also have varying importance that cannot be separated from random stochasticity. Their low importance scores may be because their contribution is minimal compared to key variables that explain most of the classifications (i.e., tax, crime, and code cases). However, we note that importance scores come from a skewed representation of the samples (see S.I. section E). When considering the distribution of values in the global sample, the importance of crime and code violations may be overestimated, while the importance of tax delinquency may be underestimated.

Table 2: Drop-column importance score in all random forest models.

	All Features				Reduced Features			
	S1: 200 samples		S2: 300 samples		S3: 200 samples		S4: 300 samples	
	Land	Structure	Land	Structure	Land	Structure	Land	Structure
Feature Importance Score (%)								
Weighted Crime Count	**2.50 (R2)**	**5.07 (R1)**	**2.86 (R2)**	**3.59 (R2)**	**3.75 (R3)**	**6.70 (R1)**	**2.86 (R3)**	**4.62 (R2)**
Weighted Drug Count	-1.18*	-0.87*	0.95*	0.00*	2.57*	-1.70*	**3.81 (R2)**	-1.54
Weighted Active Code Cases	-1.18*	**1.74 (R3)**	0.00*	**2.05 (R4)**	**4.93 (R1)**	**3.37 (R3)**	**1.91 (R4)**	**1.03 (R3)**
Weighted Fire Count	-1.18	0.83	-2.86*	1.03*	NA	NA	NA	NA
Tax Delinquency & Delinquent Years	**3.53 (R1)**	**3.37 (R2)**	**8.57 (R1)**	**8.72 (R1)**	**4.85 (R2)**	**5.00 (R1)**	**13.3 (R1)**	**5.64 (R1)**
Special Assessment Tax Pct	-4.85*	0.87	0.95	**2.56 (R3)**	-3.60	1.63*	-2.86	**1.03 (R3)**
Vacancy Probability	-2.35*	0.91*	-0.95*	1.03*	NA	NA	NA	NA
Property Value	-3.60*	0.04*	-0.95*	0.51*	-2.43	-0.91*	0.00*	-1.54
Land Size	-3.53*	0.00*	-0.95*	-1.03	NA	NA	NA	NA
Qualified Sales	-3.60*	0.04*	0.95*	0.00*	NA	NA	NA	NA
Unqualified Sales	-3.60*	0.04	-0.95*	1.03*	NA	NA	NA	NA
Year Last Sold	-3.60*	0.04	-0.95*	-0.51*	NA	NA	NA	NA
Growth Rate	-3.60*	-0.87*	-0.95*	1.54*	NA	NA	NA	NA
Median Neigh PV	-3.60*	0.00*	0.00*	1.03*	NA	NA	NA	NA
Evaluation Metrics (%)								
Cross Validation Accuracy	87.94	90.65	91.43	87.69	91.62	91.45	95.24	87.69
OOB score	91.46	91.53	88.57	87.18	92.68	91.53	90.48	88.21

* Indicates that the number fluctuates above or below zero depending on the random state of the random forest algorithm and thus deemed uncertain. The unit of the number is %. *R* in parenthesis indicates the ranking of features that have stable contributes to at least 1% drop-column importance. Numbers bolded are important for each model. Our final model is set 4 (S4).

Model Accuracy and Validation

Overall, the model accuracy (with both cross validation and OOB score) varies from 87-95% for land and 87-92% for structures (see Table 2). Reducing the features to a few key variables slightly improves the model accuracy, indicating that the variability in the additional variables may confuse the model more than providing useful information. Given these comparisons, we chose the model trained with reduced features and 300 samples as our final model to give predictions (in Fig. 3).

To further validate our model predictions, we compared machine-predicted VAD properties with a list of VAD targets labeled by HNSD staff in 2019 through a field survey. We found that model prediction agrees with 72% of the human found targets. The ground-truth accuracy may be even higher. Figure 3 shows that machine predicted VAD properties are more spatially diverse and larger than human labeled targets, which is not surprising because a field survey is often confined to selected neighborhoods. Our model identified 1,234 VAD properties among 5,327 candidates, while the human-generated list contains 693 properties. Within the 693 human found properties, 286 (41%) did not meet the basic requirements to be machine learning candidates, because they either have zero records in code violation, tax delinquency, drug records, fire incidents, or have flood risks. These properties are excluded from the comparisons. The experts may have labeled them simply due to vacancy (although vacant land alone does not qualify as a VAD property) or our data do not reflect the current VAD conditions of the properties because they may be outdated.

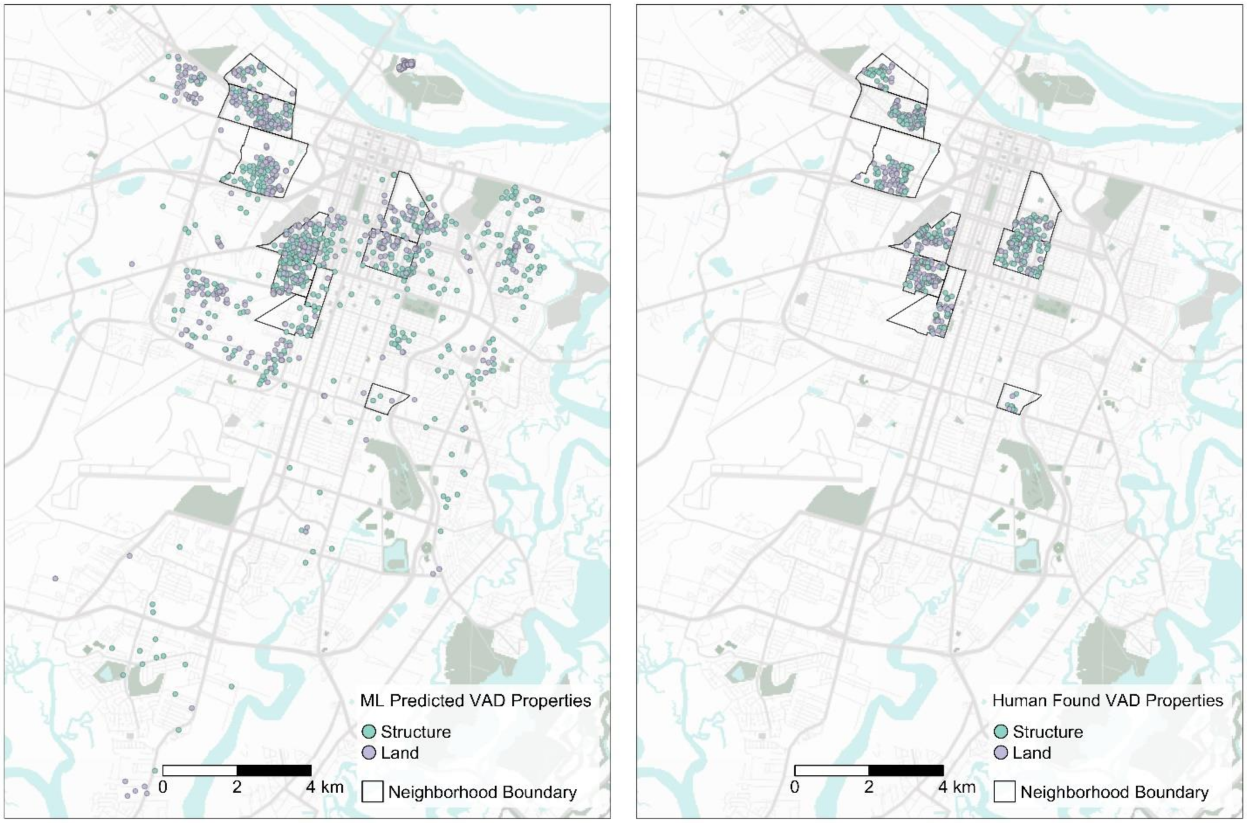


Figure 3: Compare model predicted VAD properties (left) with human found VAD properties (right).

By comparing the percentage of VAD types (i.e., VAD status driven by various factors) between machine vs. human, we found that machine predicted VAD land parcels tend to have more crime (8.1% vs. 6%), tax delinquency (100% vs. 67.5%), and code violation (29.7% vs. 24.4%) types, and less low property values (91% vs. 97.5%) type. For land with structures, machine predicted VAD properties tend to have more crime (44.6% vs. 27%) and tax delinquency (98% vs. 88%) types, and fewer code violation (33.8% vs. 43.3%) and low property values (68% vs. 72%) types. The VAD types are derived by looking at whether crime, tax, code, and property value of a property passes certain thresholds to be significant at determining the VAD status (see S.I. Section F).

Lastly, we found that only 66% of labels in model predictions using the 2019 input data stayed the same with 2021 input data. For the 20% properties that changed from Not VAD (2019) to VAD (2021), some may have new conditions that contributed to VAD status. For the 13% that changed from VAD (2019) to Not VAD (2021), some may already receive interventions (a few experienced tax sales) or simply have been mislabeled in 2019. Thus, while streamlining data is difficult, continuing investments to run the model with updated data is crucial to keep the model accuracy rate.

Bias Analysis

Our model builds on data collected from decentralized and heterogeneous sources. We found that neighborhoods with low home-ownership rate, low-income, and high concentration of African Americans tend to have less code cases coming from 311 calls (see maps in S.I. Section F), (as in Kontokosta, Constantine, Hong, 2021). Thus, these neighborhoods may either have less accurate code violation data or experience 'over-policing' from code compliance officers. We listed biases identified for each variable in S.I. Section F.

Discussion and Conclusion

In this paper, we presented a human-in-the-loop approach to develop a random forest model that predicts vacant, abandoned, and disinvested (VAD) properties in the City of Savannah, Georgia. Different from traditional machine learning approaches that focus solely on optimization, we involved local experts from the Housing and Neighborhood Services Department and Chatham County / City of Savannah Land Bank Authority in building, training, and evaluating the model. We articulated issues of property regeneration and identified appropriate places for machine learning intervention before jumping into the application; we built our own training data through active learning techniques, ground-truthed our predictions with human-found VAD targets, and examined how well the prediction held over time. We used interpretable machine learning methods (e.g., decision trees and feature importance) to facilitate learning and consensus among local experts. Finally, we identified neighborhoods that were vulnerable to the context that code violation data were collected.

We found that tax delinquency (and delinquent years), crime history, and code violations are the three most important features at identifying VAD properties. The predictions from the model agreed with the expert labels about 72% of the time. And the model could predict about 66% of labels after two years. To reach a higher accuracy, the data should be annotated using information from field visits. The comparison also revealed that machine-generated predictions are more spatially diverse and less focused on code violations and low property values than human-identified candidates. We recommend that the city prioritize building database connections to tax, crime, and code records, as these are the main determinants for the VAD status.

Our process suggested that developing a machine learning model using a human-in-the-loop approach was time-consuming and challenging yet rewarding. Involving housing experts helps us contextualize the selection of variables, build training data, adapt the model, and validate the model outcomes. In return, some experts reported that the labeling process was enlightening, as it helped organize their thoughts in making decisions and reflect upon why they choose certain outcomes for the parcels. After the process, they felt more confident and knowledgeable about deploying the model in future operations. The human factor in ML also brings new challenges: experts have discrepancies at labeling VAD properties. Our HILML process highlighted these discrepancies and thus facilitated more agreements of the decision criteria. Thus, the HILML approach streamlines the identification process and makes it more transparent.

However, our study also has some limitations. First, our model focuses on challenges at the institutional scales rather than challenges of the property owners. As such, we did not incorporate voices and inputs from communities and property owners, reinforcing a transactive mode of governance (where citizens are consumers of public services) rather than a relation mode (where citizens are co-creators). Second, updating the model requires data infrastructure that are dependent on third-party services, which is a common struggle in many smart city initiatives (Kitchin 2014). Lastly, while the model suggests candidates based on VAD characteristics, the acquisition of VAD candidates in real life is still political, as many minority communities are concerned with the impacts of governmental interventions. Altogether, we are unsure whether the investments to generate such model outweigh the burdens of maintenance and education needed for the model to sustain. Future work should examine these issues and recreate this approach for other cities and case studies.

In conclusion, this research presents a case study that illustrates a human-in-the-loop machine learning approach for classifying land and residential structures as potentially in need of attention. By using a collaborative technique that engages experts while using large datasets and statistical analysis, we were able to generate new insights into how to potentially automate or improve the local government’s ability to identifying vacant, abandoned, and disinvested properties. The result is a more reliable method for managing assets in a municipal setting. This HILML approach may be used in other planning efforts, such as incorporating local knowledge to detect missing buildings from satellite imagery, supporting community needs by training models on crowd-sourced data, or improving machine-generated urban design scenarios with human feedback. We suggest that more researchers and practitioners use machine learning in their planning efforts, but also incorporate human input and expertise as they develop and test these models.

References

Anwar, Sajjad. 2022. “Open sourcing PEARL – human-in-the-loop AI for land cover mapping”. *developmentSEED*, Accessed March 15, 2022. https://developmentseed.org/blog/2022-03-15-open-sourcing-pearl

Appel, Sheila U., Derek Botti, James Jamison, Leslie Plant, Jing Y. Shyr, and Lav R. Varshney. "Predictive analytics can facilitate proactive property vacancy policies for cities." *Technological Forecasting and Social Change* 89 (2014): 161-173. https://doi.org/10.1016/j.techfore.2013.08.028

Armstrong, M. P., Densham, P. J., & Rushton, G. 1986. “Architecture for a microcomputer based spatial decision support system.” In *Proceedings of the Second International Symposium on Spatial Data Handling,* pp.120-131. Seattle, 1986.

Basu, Sumanta, Karl Kumbier, James B. Brown, and Bin Yu. "Iterative random forests to discover predictive and stable high-order interactions." *Proceedings of the National Academy of Sciences* 115, no. 8 (2018): 1943-1948. https://doi.org/10.1073/pnas.1711236115

Boehner, Kirsten, and Carl DiSalvo. "Data, design and civics: An exploratory study of civic tech." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2970-2981. San Jose, 2016. https://doi.org/10.1145/2858036.2858326

Branas, Charles C., David Rubin, and Wensheng Guo. "Vacant properties and violence in neighborhoods." *ISRN Public Health*, no. 2012 (2012): 1-23. https://doi.org/10.5402/2012/246142

D'ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Cambridge, MA: MIT press, 2020.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.

Forester, John. "Planning in the Face of Power." *Journal of the American Planning Association* 48, no. 1 (1982): 67-80.

Gilreath, Morgan B. "A model for quantitatively defining urban blight by using assessment data." *Fair & Equitable*, 2013. Accessed July 3, 2022. https://www.iaao.org/media/Topics/FE_Aug13_Gilreath.pdf

Hillenbrand, Katherine. "New Orleans brings data-driven tools to blight remediation. Harvard Data Smart City Solutions". Harvard University. Accessed February 16, 2022. https://datasmart.ash.harvard.edu/news/article/new-orleans-brings-data-driven-tools-to-blight-remediation-915

Hillier, Amy E., Dennis P. Culhane, Tony E. Smith, and C. Dana Tomlin. "Predicting housing abandonment with the Philadelphia neighborhood information system." *Journal of Urban Affairs* 25, no. 1 (2003): 91-106. https://doi.org/10.1111/1467-9906.00007

Holzinger, Andreas. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?." *Brain Informatics* 3, no. 2 (2016): 119-131. https://doi.org/10.1007/s40708-016-0042-6

Hopkins, Lewis D., and Marc P. Armstrong. "Analytic and cartographic data storage: a two-tiered approach to spatial decision support systems." In *Proceedings, Seventh International Symposium on Computer-Assisted Cartography. Washington, DC: American Congress on Surveying and Mapping*. Washington DC, 1985.

Housing and Neighborhood Service. n.d. City of Savannah (website). Accessed February 16, 2022. https://www.savannahga.gov/484/Housing-and-Neighborhood-Services.

Innes, Judith E. "Planning theory's emerging paradigm: Communicative action and interactive practice." *Journal of Planning Education and Research* 14, no. 3 (1995): 183-189. https://doi.org/10.1177/0739456X9501400307

Keenan, Peter Bernard, and Piotr Jankowski. "Spatial decision support systems: Three decades on." *Decision Support Systems* 116 (2019): 64-76. https://doi.org/10.1016/j.dss.2018.10.010

Kitchin, Rob. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. New York, NY: Sage, 2014.

Kontokosta, Constantine E., and Boyeong Hong. "Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions." *Sustainable Cities and Society* 64 (2021): 102503.

Immergluck, Dan. "Examining changes in long-term neighborhood housing vacancy during the 2011 to 2014 US national recovery." *Journal of Urban Affairs* 38, no. 5 (2016): 607-622. https://doi.org/10.1111/juaf.12267

International Trade Administration. n.d. “Maritime Services Trade Data”. U.S. Department of Commerce. Accessed February 16, 2022. https://www.trade.gov/maritime-services-trade-data

Le Dantec, Christopher A. *Designing Publics*. Cambridge, MA: MIT Press, 2016.

Lewis, David D., and Jason Catlett. "Heterogeneous uncertainty sampling for supervised learning." In *Machine Learning Proceedings 1994*, pp. 148-156. New Brunswick, 1994. https://doi.org/10.1016/B978-1-55860-335-6.50026-X

Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2, no. 3 (2002): 18-22.

Lincoln Institute of Land Policy. n.d. “Legacy Cities”. Accessed February 16, 2020. https://www.lincolninst.edu/research-data/data-toolkits/legacy-cities

Loukissas, Yanni Alexander. *All Data are Local: Thinking Critically in a Data-driven Society*. Cambridge, MA: MIT Press, 2019.

Mallach, Alan, and Lavea Brachman. *Regenerating America's legacy cities*. Cambridge, MA: Lincoln Institute of Land Policy, 2013.

Mallach, Alan. *The Empty House Next Door: Understanding and Reducing Vacancy and Hypervacancy in the United States*. Cambridge, MA: Lincoln Institute of Land Policy, 2018.

Mallach, Alan. *Making the Comeback: Reversing the Downward Trajectories of African American Middle-Income Neighborhoods in Legacy Cities*. Cambridge, MA: Lincoln Institute of Land Policy, 2021.

Morckel, Victoria C. "Using suitability analysis to prioritize demolitions in a legacy city." *Urban Geography* 38, no. 1 (2016): 90-111. https://doi.org/10.1080/02723638.2016.1147756

Monarch, Robert Munro. *Human-In-The-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Simon and Schuster, 2021.

Reyes, Eduardo Blancas, Jennifer Helsby, Katharina Rasch, Paul van der Boor, Rayid Ghani, Lauren Haynes, and Edward P. Cunningham. "Early detection of properties at risk of blight using spatiotemporal data." *Data for Policy 2016: Frontiers of Data Science for Government: Ideas, Practices and Projections.* Cambridge, United Kingdom, 2016. https://doi.org/10.5281/zenodo.556510

Savannah Area GIS Open Data. n.d. Savannah Area Geographic Information System (website). Accessed February 16, 2022. https://data-sagis.opendata.arcgis.com/

Savannah, Georgia, Municipal Code § 2-18 (2016).

Silverman, Robert Mark, Li Yin, and Kelly L. Patterson. "Dawn of the dead city: An exploratory analysis of vacant addresses in Buffalo, NY 2008–2010." *Journal of Urban Affairs* 35, no. 2 (2013): 131-152. https://doi.org/10.1111/j.1467-9906.2012.00627.x

Sumell, Albert. "The determinants of foreclosed property values: Evidence from inner-city Cleveland." *Journal of Housing Research* 18, no. 1 (2009): 45-61. https://doi.org/10.1080/10835547.2009.12091996

Sun, Wei, Ying Huang, Ronald W. Spahr, Mark A. Sunderman, and Minxing Sun. "Neighborhood blight indices, impacts on property values and blight resolution alternatives." *Journal of Real Estate Research* 41, no. 4 (2019): 555-604. https://doi.org/10.22300/0896-5803.41.4.555

Schilling, Joseph, Katie Well, Jimena Pinzon, John Kromer, and Ed Rendell. *Charting the Multiple Meanings of Blight: A National Literature Review on Addressing the Community Impacts of Blighted Properties*. Alexandria, VA: Vacant Property Research Network, 2015.

Schilling, Joseph, and Jimena Pinzon. *The Basics of Blight: Recent Research on its Drivers, Impacts, and Interventions*. Alexandria, VA: Vacant Property Research Network, 2016.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC Bioinformatics* 8, no. 1 (2007): 1-21. https://doi.org/10.1186/1471-2105-8-25

Weaver, Russell C., and Sharmistha Bagchi-Sen. "Spatial analysis of urban decline: The geography of blight." *Applied Geography* 40 (2013): 61-70. https://doi.org/10.1016/j.apgeog.2013.01.011Cullen

Whitaker, Stephan, and Thomas J. Fitzpatrick IV. "Deconstructing distressed-property spillovers: The effects of vacant, tax-delinquent, and foreclosed properties in housing submarkets." *Journal of Housing Economics* 22, no. 2 (2013): 79-91. https://doi.org/10.1016/j.jhe.2013.04.001

Williams, Sarah. *Data Action: Using Data for Public Good*. Cambridge, MA: MIT Press, 2020.

Zhou, Jianlong, and Fang Chen, eds. *Human and Machine Learning: Visible, Explainable, Trustworthy, and Transparent*. Cham, Switzerland: Springer Cham, 2018.

Zheng, Zhibin, and Renee Sieber. "Putting humans back in the loop of machine learning in Canadian smart cities." *Transactions in GIS* 26, no. 1 (2022): 8-24.