Landscape and Urban Planning 216 (2021) 104239

Contents lists available at ScienceDirect

Landscape and Urban Planning

journal homepage: www.elsevier.com/locate/landurbplan

Leveraging machine learning to understand urban change with net construction

Nathan Ron-Ferguson [a,*], Jae Teuk Chin [b,*], Youngsang Kwon [c]

[a] Department of Earth Science, University of Memphis, 109 Johnson Hall, Memphis, TN 38152 United States
[b] Department of City and Regional Planning, School of Urban Affairs and Public Policy, University of Memphis, 208 McCord Hall, Memphis, TN 38152 United States
[c] Department of Earth Science, University of Memphis, 109 Johnson Hall, Memphis, TN 38152 United States

HIGHLIGHTS

- Machine learning offers potential to capture the inherent complexity of urban change.
- Net construction quantifies construction and demolition as a single combined metric.
- Random forest regression using net construction outperformed independent models.
- Net construction reveals land use mix as the most important feature.

ARTICLE INFO

Keywords:
Urban change
Net construction
Machine learning
Random forest
Big data
Building permit data

ABSTRACT

A key indicator of urban change is construction, demolition, and renovation. Although these development activities are often interrelated, they are typically studied independent of one another. Analytic methods relying on a strict set of modeling assumptions limit our ability to understand this change holistically. Machine learning has demonstrated the potential when combined with big data to discover patterns and relationships between seemingly unrelated variables. This research explores urban change through *net construction*, a composite value that treats demolition as a deductive process that is subtracted from construction activity which provides for a more holistic and nuanced understanding of development activity. Once validated through a visual analysis of its reliability as a measure of urban change, we then used a series of random forest regression models to evaluate the predictive accuracy of *net construction* compared with independent models of construction and demolition. Applying the approaches to an urban county in the United States, we compiled 122 independent variables to provide a comprehensive view of individual neighborhoods from multi-disciplinary data sources such as socio-economic, built environment characteristics, and landscape metrics. We then analyze the feature importance scores derived from the random forest models in an effort to assess the similarities and differences between the variables that have the greatest influence on model accuracy. The net construction model produced more accurate results than models that used construction and demolition activity independently. While many of the most important features aligned with those from the independent models, land use mix drawn from landscape metrics appeared as the most important, representing a departure from previous studies. This study provides a scalable method for modeling urban change using machine learning techniques and reveals the importance of applying data-driven algorithms that can help communities become more informed about their pressing issues.

1. Introduction

A community's urban form and physical characteristics are shaped by the interaction of transportation networks and adjacent land uses. This association between contributing factors to the built environment and the impact it can have on socio-economic conditions has been analyzed to comprehend and measure a variety of development patterns. A deeper understanding of development patterns is crucial to understanding how cities change and how that change may affect the people residing in those communities. Varying measures have been used

* Corresponding authors.
E-mail addresses: jchin3@memphis.edu (J.T. Chin), ykwon@memphis.edu (Y. Kwon).

https://doi.org/10.1016/j.landurbplan.2021.104239
Received 10 November 2020; Received in revised form 23 August 2021; Accepted 1 September 2021

to gain insight into how cities are changing (Ewing and Hamidi, 2015) but are typically independent of indicative metrics which often fails to present a comprehensive view of how a city is evolving, in particular at the parcel or neighborhood scale (Coulton, 2012).

Construction, demolition, and renovation are key processes of development that promote a changing urban configuration. These three activities are a common property development unit and provide insights on how the built environment functions and relates to urban form at the parcel level (Hollander et al., 2019). This study seeks to understand what role construction, demolition, and renovation activity can play in analyzing and predicting urban change. Additionally, to analyze these processes holistically, we developed a *net construction* value by treating demolition as a deductible activity that, once removed from a neighborhood's overall construction rate, provides a more nuanced understanding of how it is changing. To achieve this goal, we address three research questions to be empirically modeled and tested. First, can the inclusion of construction, demolition, and renovation permit data as a standardized single composite value provide a better and more holistic depiction of community change than each activity independently? Secondly, what are the underlying characteristics of communities that experience different types and rates of change in development patterns? Finally, how do these characteristics differ in physical and social terms from those previously described in the literature, and what might this imply to intervene and manage inefficient urban form?

One of the most significant challenges in modeling urban change and development patterns is how to account for the complexity and non-linear relationships between the various factors influencing urban form. Conventional methods such as linear regression are limited in their ability to handle colinear or non-linear relationships thus it is necessary to leverage methods whose inputs are not constrained by distributions or data types. In recent years, the emergence of machine learning (ML) has helped various disciplines, from computer science to statistics to engineering, discover new patterns as a result of improved performance for computing complexity. However, the empirical applications of ML in urban studies are still at an early stage that has mainly been focused on urban modeling using remotely sensed inputs to classify land use and land cover (Abrantes et al., 2019; Ghosh et al., 2014). By contrast, this study introduces and utilizes physical and social attributes of individual neighborhoods collected from various sources, including national census and local data from municipalities. To account for the complexity of integrating numerous and extensive data sets with varying structures, we present a data-driven methodology using random forest (RF). This scalable approach applies to an urban county in the United States to demonstrate how it performs when modeling fluctuating development patterns across neighborhoods. This will make an analytical and empirical contribution to the literature in ML and urban studies.

To the best of our knowledge, this paper is one of very few machine learning studies investigating urban change at the neighborhood level with built environment characteristics. ML methods use data-driven model selection by creating and tuning alternative models with the data itself, which differs from developing one model specified by the researcher. The computational power of ML has been enhanced to deal with complicated features of the changing urban environment and may offer a new way of city management to implement data interventions (Glaeser et al., 2018). This paper is structured into the following sections. The second section outlines the modeling approach for urban complexity and reviews machine learning applications in urban studies. The third section introduces data and research methods, and the fourth section discusses the results of our analysis. The fifth highlights the key findings, and the final section concludes with the implications of the research.

2. Modeling complexity and machine learning in urban studies

The literature of modeling urban change has evolved in response to developing new analytical methods and the advancement of data technologies. Unfortunately, this evolution often occurs in a disciplinary vacuum due to a lack of common definitions and measurements regarding what constitutes urban change because of its multidisciplinary nature (Clifton et al., 2008) and resulting complexity (Boeing, 2018). Research into the underlying factors affecting change is further complicated by a need to utilize methods that generate interpretable coefficients to promote actionable policy decisions (Waddell and Besharati-Zadeh, 2020).

One of the most common tools in studying urban development and resulting urban form has been the suite of regression approaches. Knaap et al. (2007) and Lowry and Lowry (2014) utilized linear regression to explore the relationship between development patterns and urban configuration. Although many of the variables in their research exhibited a linear relationship, reliance upon simple regression models limited the types of data used and failed to consider the complex nature of urban environments. When a categorical dependent variable was used, Yin and Silverman (2015) utilized logistic regression to evaluate the growth priorities of Buffalo, New York using construction and demolition permits with property acquisition data. Their study is significant as it allows for nonlinear relationships among the variables and its use of permits to detect where changes occurred. Charles (2011) also used multilevel logistic regression to analyze suburban gentrification and stressed the importance of demolition, in case replaced with larger rebuilt housing. While these efforts incorporate nonlinear effects, their emphasis on abandonment and demolition fails to take a more intricate look at development activity in their models.

Expanding beyond the regression approaches, Talen et al. (2018) incorporated some of the complexity by developing a global typology of 27 built landscape types to associate morphological patterns with social disparities across neighborhoods. They found connections between previously unexplored relationships such as segregated land uses and racial segregation. However, they relied heavily on a subjective manual labeling process that limited the ability to replicate their methods and the number of variables used. The work of Boessen et al. (2018) supports the notion that the built environment plays a significant role in social ties in an empirical study. While they confirmed an association between morphology and social outcomes, all but the distribution of social networks lay beyond the purview of observable data, which implies uncertainty in results when modeling the effects of urban form on socio-economic standards. This empirical limitation of modeling has generated efforts to review the built environment effect across disciplines systematically. Mazumdar et al. (2018) and Carmona (2019) reached similar conclusions analyzing different studies and methods. Mazumdar et al. examined 23 studies of relationships between morphology and social capital to report that design and the presence of destinations were significantly related to social capital outcomes. Carmona took morphology to be embodied policy and argued that place value may be described in terms of the extent to which qualities of the built environment impact, either positively or negatively, on different social policy goals.

This modeling complexity makes data-driven approaches difficult and thus restricts comprehensive understanding of community change. The ML methods, once combined with big data, offer alternative methodological pathways for urban researchers to reveal new dimensions of urban phenomena. Although the field of urban studies has not seen the same widespread adoption as other disciplines, there has been progress in this regard and it is therefore important to explore what ML is, and how it has been applied within the context of urban issues.

In general, machine learning problems fall into three categories: supervised, unsupervised, and reinforcement learning (Bishop, 2006). Empirical research relying on big data often applies supervised learning, in which input vectors are used to train or supervise an algorithm for predicting outcomes. RF models, in particular, are a supervised learning model that can handle both regression and classification problems depending on the input data (Breiman, 2001). RF has been utilized by a number of researchers within the realm of urban studies, but its

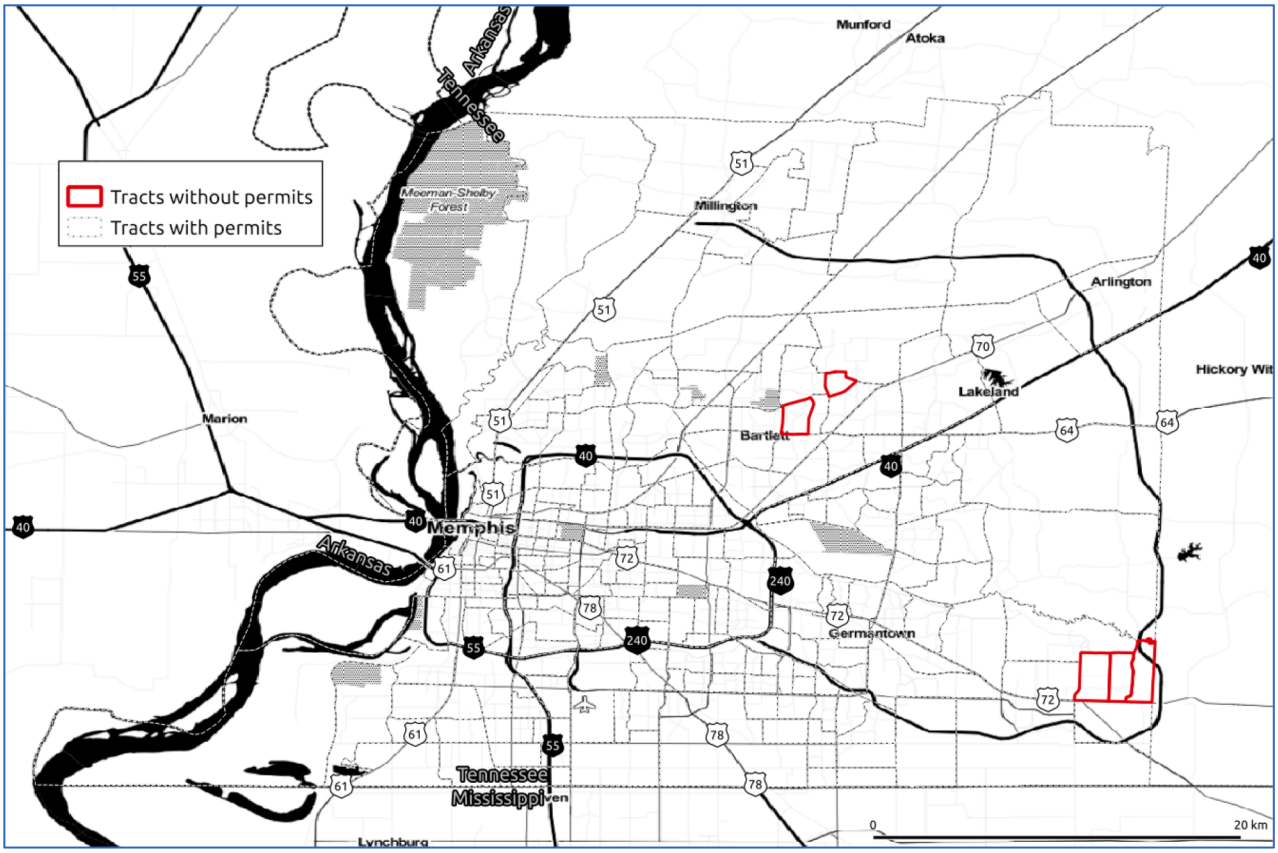

Fig. 1. Study area census tract. Census tracts were used to aggregate permit counts. Census tracts without permits are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

application has been mainly using remote sensing data at the regional scale, where it is used to predict or model land use classification, land use change, or urban growth (Ghosh et al., 2014; Rodriguez-Galiano et al., 2012; Shafizadeh-Moghadam, 2019; Walde et al., 2014). At a more localized scale, Ossola et al. (2019) utilized RF methods to handle a large volume of LiDAR data to develop a range of yard metrics such as canopy height and coverage to study morphological characteristics of residential parcels.

Apart from the use of remotely sensed data in RF, Reades et al. (2019) used RF to model gentrification in London using socio-economic variables with the presence of collinearity between input variables. They compared RF against more traditional methods such as simple and multiple regression to determine its effectiveness within the human geography domain. RF outperformed both simple and multiple regression regardless of parameter tuning, achieving higher R^2 values. Another application of RF is as a feature selection method. Yoo et al. (2012) implemented RF to identify variables for use in hedonic models to predict housing price. Like Reades et al. (2019), they used socio-economic data along with numerous housing characteristics and observed the potential for missing an optimal combination of variables when using a stepwise regression approach. Most recently, utilizing a combination of socio-economic, land assessment, and housing sales data sets in RF methods, Auerbach et al. (2020) and Hu et al. (2019) predicted future property values and housing rental prices, respectively. Sapena et al. (2020) employed RF regression to estimate the predictive power morphology in social outcomes. For instance, as much as 68% of income variance or 32% of employment rate variance across the six hundred urban areas studied may be predicted by spatial data including road networks, land cover densities, and built-up footprints. While RF regression proved to be extremely powerful when comparing regions, it was also highly dependent upon data availability and quality.

Several transportation studies have also used RF in modeling mobility behaviors and route selections. For example, by using national household travel survey data, Sabouri et al. (2020) used RF to study the relationship between ride-sourcing services like Uber and vehicle ownership. This study also corroborates previous findings that ML has superior predictive power than conventional models. In this vein, Tribby et al. (2017) noted that data-driven RF methods for variable selection produced more accurate models than the theory-driven models based on predefined walkability factors when analyzing walking route choice with GPS-derived trip data.

Recent applications of ML, artificial intelligence, and urban big data extend beyond domains like remote sensing and transportation where data format and availability made for easy incorporation. Innovations in methodological techniques include the use of computer vision to assess change in neighborhood appearance and to correlate that change with underlying characteristics (Naik et al., 2017) or the use of recurrent neural networks (RNN) to detect urban change (Papadomanolaki et al., 2019). There have also been numerous innovations regarding the data used to understand the relationships between the natural and urban environments and their inhabitants. Cao et al. (2021) incorporated 3-dimensional building data to study the effect of building morphology and air temperature and seasonality. Mora et al. (2018) utilized social network service data to determine whether the digital by-products of online communication could improve municipal service delivery.

Despite these recent innovations in urban morphological research, there is a need to more closely analyze the systemic nature of urban change, and the tendency to draw variables from a single discipline reveals a need for methods that can accommodate non-linear relationships between neighborhood characteristics and urban change. The following sections contain an overview of the methodology that we developed to incorporate a more data-centric approach to modeling

Table 1
Sample permit data after conversion. Personally identifiable information has been replaced with "–".

Column Name	Construction Permit	Renovation Permit	Demolition Permit
parcelid	D0209B C00020	089001 00056	060222 00140C
permit	B0884362	E0403646	BD000705
issued	2002-09-17	2005-10-21	2011-04-29
sub_type	res	com	bus
const_type	new	new	demo
valuation	100000	0	0
address	–	–	–
zip_code	NULL	38134	38123
fraction	Cordova	Memphis	Memphis
map_pg	160	33B	63H
lot	289	NULL	NULL
subdivision	LEE LINE FARMS	NULL	NULL
zone	RS6	NULL	EMP FP
height	NULL	NULL	NULL
sq_ft	1974	0	0
cu_ft	0	0	0
fire_o	N	NULL	N
sprinkled	N	NULL	N
health_dep	NULL	NULL	NULL
num_floors	1	0	0
relationship	OWNER	CONTRACTOR	CONTRACTOR
name	–	–	–
license_no	NULL	E3715	B3905
address_1	–	–	–
address_2	MEMPHIS TENNESSEE	MEMPHIS, TN	MEMPHIS, TN
address_3	NULL	NULL	NULL
zip	38018	38134	38108
phone	–	–	–
description	NEW DWELLING WITH ATTACHED GARAGE	INSTAL FA	126 X 66′ DEMOLISH ONE STORY STORAGE LOADING DOCK BUILDIG
buildinguse	RESIDENTIAL	NULL	UNOCCUPIED
year	2002	2005	2011

urban change.

3. Data and methodology

The data used for the study represent a complex and comprehensive view of communities. The input variables were drawn from a diverse field of disciplines to offer a glimpse into their social, economic, environmental, and transportation characteristics.

3.1. Data

This study is centered on Shelby County, Tennessee and its principal city, Memphis which was established on the eastern banks of the Mississippi River (see Fig. 1). The presence of a bluff along the Mississippi that today runs roughly between the Interstate 40 and 55 bridges on the western edge of the city made it an ideal stop for riverboat traffic and the modern central business district (CBD) still contains many of the early warehouses from that era. As it developed, the city spread eastward, with most of the development occurring between two of the primary tributaries, which parallel Interstate 40 and Interstates 55/240, respectively. The city's core falls mostly inside the Interstate 40/240 loop, and except for a few neighborhoods clustered along U.S. Highway 72 and the CBD, it has experienced decades of decline and disinvestment. As residents left the city, they moved eastward to both unincorporated regions on the eastern edge of the county line and several suburban municipalities.

We used 2012 census tract boundaries as an analytic unit to serve as a proxy for neighborhood boundaries for data aggregation and assessment since 2012 was the start year for the 5-year American Community Survey (ACS) estimate used for socio-economic variables. While there has been some debate into the validity of census tracts as neighborhood units (Clapp and Wang, 2006; Coulton et al., 2001), census tract boundaries still provide easy access with which socio-economic, demographic, and housing characteristics can be joined for statistical purposes.

Several characteristics of census tract boundaries make them ideal units for making comparisons: they are relatively permanent, represent a degree of socio-economic similarity among households that are contained within them, and on average, contain about 4,000 individuals, although this number can range from 2,500 to 8,000 (U.S. Census Bureau, 1994). Shelby County exhibits all of these characteristics with a total of 221 tracts comprised of an average population of 4,243. There is a fair amount of variation in land area across tracts as less dense regions may require a larger area to capture the target population set by the Census Bureau. This variability is demonstrated in Shelby County's tract acreage ranges from 138 to 34,090 acres.

3.1.1. Dependent variables

The dependent variables were derived from construction, renovation, and demolition permit data. Drawing a direct connection between the specific locations where construction and demolition activity occurred within a community is a departure from similar research where it is typically aggregated at larger geographic units using median year built (Knaap et al., 2007; Lowry and Lowry, 2014). This approach allowed for a more granular understanding of urban change, especially when considering new construction or renovation occurring in older neighborhoods within a central city.

Permit data were provided by the Memphis and Shelby County Office of Planning and Development (OPD), a joint city and county office that oversees the administrative functions of planning for both administrative districts. OPD permit data covers 15 years from 2002 to 2016 and, once processed, consisted of 152,325 records (rows) relating to construction and renovation activity and 9,073 records relating to demolition activity (see Table 1 for a sample of the data after conversion). While the construction data contained renovation permits, OPD's encoding method made it difficult to separate the two with complete accuracy; therefore, renovation and new construction permits were left aggregated as a single permit type. For example, in many instances, renovation permits can be identified by a valuation amount equal to $0, but the existence of numerous new construction permits missing a valuation amount made it difficult to separate the two without introducing unnecessary errors.

Permits were geocoded to convert each application's address into a longitude/latitude pair to analyze geospatial relationships. Once all permits were geocoded, the total number of permits covering the 15-year period were aggregated using census tract boundaries and any census tract missing permits was eliminated from consideration to ensure a more accurate account of where potential net neutral development activity (as defined by the relative difference between construction and demolition discussed below). There were 5 tracts without any development activity over the 15-year period which dropped the sample size from 221 tracts to a total of 216.

While the use of the full 15-year set of permits seems to run counter to a methodology that typically relies on 10-year increments in association with the decennial census, 2008 and 2009 were landmark years for most of the United States as the country grappled with the worst economic downturn since the Great Depression. This 15-year period provided a critical sample of data both before and after the recession that allowed us to establish reliable trends to gain insight into community changes and development activities.

3.1.2. Independent variables

A total of 122 variables drawn from 17 different sources (see Table A1 in Appendix A) were selected to provide a comprehensive view of individual communities from the perspective of their inhabitants, employers, transportation system, and physical environment. Many of

Table 2
Summary statistics. Summary statistics for permits before and after normalization.

Variable	Mean	Std	Min	25%	50%	75%	Max	Number Permits
Construction/Renovation[1]	681.78	1,407.31	1.00	154.50	267.00	605.75	13,695.00	152,325
Demolition	39.08	48.32	0.00	6.00	18.00	59.25	252.00	9,073
Scaled Construction	0.05	0.10	0.00	0.01	0.02	0.04	1.00	–
Scaled Demolition	0.16	0.19	0.00	0.02	0.07	0.24	1.00	–
Net Construction	−0.11	0.22	−0.98	−0.20	−0.04	0.00	0.93	–

1 represents the total number of construction and renovation permits within each census tract

the variables chosen for the study were based upon the findings of Reis et al. (2016), whose comprehensive inventory of variables relating to urban change demonstrated a need for a more robust collection of features that included both spatial metrics alongside socio-economic and demographic variables.

The most extensive collection of variables used in the study were pulled from the 2012–2016 ACS 5-year estimate performed by the U.S. Census Bureau. The 2012–2016 survey was selected because of its overlap with the latter years of available permit data as discussed above. Since permits are treated as a cumulative snapshot in time, the survey data represents the total effect of that aggregate activity. This approach aligns with one of the objectives of this study: To determine whether net construction (see section 3.2.1) is an effective mechanism for assessing urban change.

While most variables from ACS were collected without any modification, a few were aggregated for simplicity or to eliminate unnecessary redundancy. For example, instead of loading individual racial groups and evaluating them independently, all minority groups were combined for a single estimate of the non-white population (*pct_nonwh*). Similar aggregation was performed for travel time and travel mode to work, combining all age groups for each variable rather than loading each age group individually.

The next largest source of variables for the study was drawn from the U.S. Census Bureau's Longitudinal Employer-Householder Dynamics (LEHD) program, specifically the LEHD Origin-Destination Employment Statistics dataset (LODES). A total of 25 variables relating to the types of industries and the number of jobs by industry were used in the study. Unlike data from the ACS, the LODES data were originally reported at the census block level thus, we aggregated the sum for the LODES data to the census tract level.

Local cadastral and assessment data, provided by the Shelby County Assessor of Property (SCAP), were used to overview physical characteristics and the built environment in terms of the structure of the neighborhoods and the structures that inhabit it. Many of the variables derived from the Assessor such as the number of living units (*livunit*), the total number of rooms (*rmtot*), or the total square footage of living area (*sfla*) were derived using the median value of all parcels within census tracts. Others were used to calculate a percentage of the total for the tract (e.g., *pct_dev*, percent developed parcels; *pct_mf*, percent multi-family parcels; *pct_comm*, percent commercial square footage).

The SCAP data were also used to calculate a series of landscape metrics drawn from the landscape ecology literature to provide a deeper understanding of both the physical shape of land as well as the uses occurring on each parcel. Connections between urban form and environmental health have a long history in planning literature and can be traced back to early writings like Ian McHarg's *Design with Nature* (1969) or Kevin Lynch's *Theory of Good City Form* (1981). Despite the long history between urban form and the natural environment, assessing their relationship systematically has often been challenging (Alberti, 1999). One of the earliest attempts to develop a systematic approach was Alberti et al (2001) who connected urban development patterns to ecological conditions. Alberti expanded on this work by then proposing a framework by which the impact of various urban development patterns on the environment could be tested (Alberti, 2005). In developing their framework for evaluating the ecological impacts of different urban development patterns, Alberti (2005) relied on a methodology proposed by Turner and Gardner (1991) which was also the basis for PyLandStats (Bosch, 2019), an open-source Python library built on top of the Scientific Python (SciPy) stack that we used to develop the measures used in this study. The majority of these measures were calculated based upon clusters of patches (i.e., land uses). Patches were defined by dissolving contiguous SCAP land use codes and clipping the resulting patches using census tract boundaries. Examples of landscape metrics include a measure of land use diversity (cf. land use mix) such as a number of patches within each tract (*number_of_patches*) or land use configuration such as edge density (*edge_density*), and largest patch index (*largest_patch_index*).

3.2. Methodology

This study consists of three analyses to provide a visual and quantitative understanding of urban change with construction and demolition activity. The first analysis demonstrates the validity of *net construction* (hereafter, *net construction* will denote the *new composite value* described in section 3.2.1) as a dependent variable by examining temporal and geospatial heat maps against known local patterns. The second analysis uses a series of RF regression models to evaluate the predictive accuracy of *net construction* compared with independent models of construction and demolition. The final analysis utilizes feature importance scores, a by-product of random forest, to compare which variables appear most prominently and assess whether the roles of those variables in urban change differ from the findings in the literature.

Data were analyzed using an iterative methodology leveraging a variety of open-source tools and programming languages (see Table A2 in Appendix A). Derived from the cross-industry standard process for data mining (CRISP-DM) (Chapman et al., 2000) and the data analytics lifecycle model (DALM) (Song and Zhu, 2016), the workflow diagram (see Fig. A1 in Appendix A) demonstrates how the analysis progressed through each step and revealed the iterative process of analysis to achieve better results.

3.2.1. Net construction

Current literature that has utilized permit data has typically analyzed different types of development activity independently (Charles, 2014; Dye and McMillen, 2007; Lai and Kontokosta, 2019; Silverman et al., 2015; Steenberg et al., 2019; Stevenson et al., 2010; Thomas, 2010; Weber et al., 2006). This approach makes it difficult to identify a correlation between construction and demolition. In reality, each construction project involves multiple permits (e.g., plumbing, electrical, etc.), whereas demolition needs a single permit. Over the 15-year period in the study area, construction permits occurred at a frequency that is over 16 times more than demolition. Thus, we retained the full count, not eliminating duplicate permits for a single project, by standardizing all permit counts using Min-Max Scaling formula as:

$$x_i^{'} = \frac{x_i - min(x)}{max(x) - min(x)} \quad (1)$$

This resulted in a new range of values that fall between 0 and 1 for both demolition as well as construction and renovation which are treated as a composite value (see Table 2 for a summary). *Net construction* is defined as scaled values of demolition subtracted from

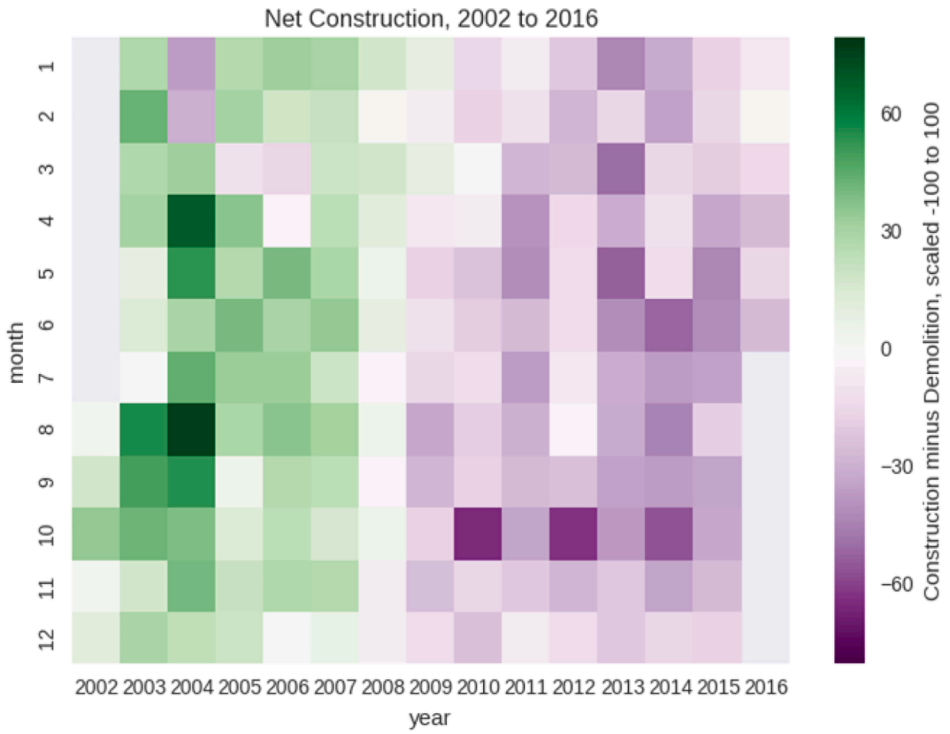

Fig. 2. Temporal heat map. Temporal heat map of net construction from 2002 to 2016. A clear delineation before and after the recession of 2008 reveals a slowdown in new construction (green) and an increase in demolition (purple) following the housing crisis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

construction and renovation.

$$x_{i,net} = x_{i,const} - x_{i,demo} \quad (2)$$

where $x_{i,const}$is the scaled value for all construction and renovation permits and $x_{i,demo}$is the scaled value for all demolition permits and $x_{i,net}$is the result net value. An increasing positive *net construction* value represents a higher rate of construction while an increasing negative value represents higher rates of demolition relative to construction and renovation activity.

To explore whether the new measure of *net construction* provides a more accurate depiction of neighborhood change, we examined the spatio-temporal pattern of *net construction* to evaluate emerging patterns against known spatial and temporal phenomena. The spatio-temporal heat map of *net construction* was created at a monthly time step.

3.2.2. Random forest regression models

RF's ability to handle large, complex datasets efficiently coupled with its ability to handle nonlinear relationships among the variables made it an ideal method for this study. Additionally, the output of feature importance scores (section 3.2.3) presents an opportunity to discover new relationships that are not typically discussed in the urban studies literature and could potentially pave the way for new policy solutions. We ran three RF models using 122 independent variables (section 3.1.2) against three dependent variables of construction, demolition, and the new composite measure of *net construction*.

RF is a type of ensemble method which is a class of learning algorithm that utilizes an aggregation of several weak learners in order to develop a single model that outperforms each individual learner (Hastie et al., 2009). To train a forest, a random sample is drawn from the full dataset to build a single decision tree. These samples are returned to the full dataset so that they can be used in subsequent trees which contributes to RF's ability to handle complexity. After selecting a sample for a tree, the algorithm then pulls a random assortment of features and begins testing a series of splits using each feature to find one that results in the lowest mean square error (MSE). Once the best feature is identified, values that fall below that threshold are split to one side and the remaining values are placed on the other branch. The process repeats until all samples are exhausted or a predefined depth is reached.

While there are a number of parameters that can be adjusted to refine the accuracy of a model during the training process, we focused on the number of trees (*n_estimators*) in the forest and the maximum number of features (*max_features*) that are tested when identifying the best way to split a tree in the Scikit-learn implementation. For other model parameters such as the maximum depth of the tree (*max_depth*) we accepted the default value. To identify the best values for *n_estimators* and *max_features* we utilized an iterative approach that permuted differing values for each parameter for the greatest stability and accuracy, resulting in using all of the features in conjunction with 400 trees.

To assess model performance, an out-of-bag (OOB) score was calculated. As each tree is built, approximately one third of the samples are withheld as a test sample while the remaining samples are used to train the tree. At the completion of the tree, the test samples are then used to evaluate the model's prediction accuracy and the OOB score is recorded so that the overall accuracy of the model can then be determined. To provide a common frame of reference with other regression-based research, we also report Pearson's r correlation (r) and Spearman's rank correlation (ρ) for the robustness tests with normally distributed values. We implemented RF using Scikit-learn, a Python-based machine learning library that is part of the Scientific Python stack (SciPy) (Pedregosa et al., 2011).

3.2.3. Feature importance score

A by-product of RF regression is feature importance scores which indicate the relative importance of each feature in determining the overall accuracy of the model. There are a variety of mechanisms used to calculate this score, but the Scikit-learn uses a Gini Importance score which is also referred to as the mean decrease in impurity and described by (Louppe, 2014) in the formula:

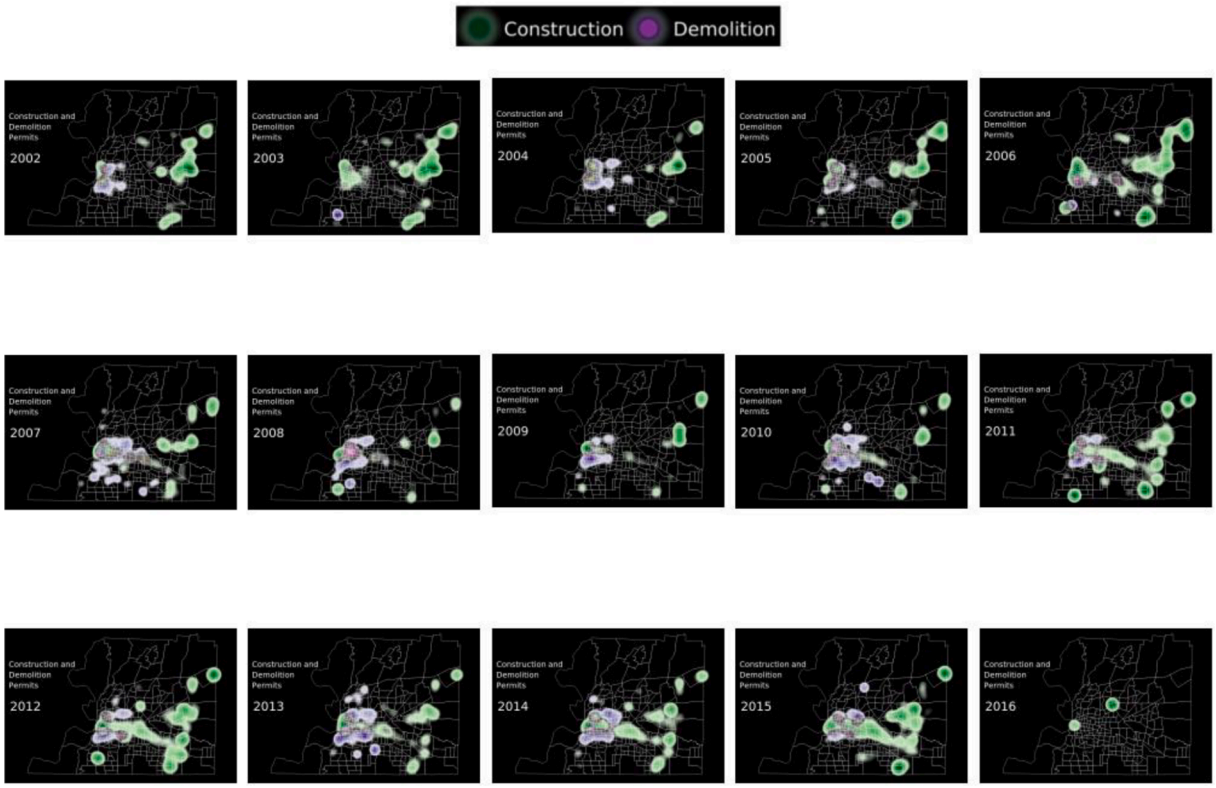

Fig. 3. Spatial heat map. Spatial heat map of construction and demolition permits, 2002 to 2016. Greater concentrations of construction (green) on the eastern edge of the county and demolition in the core (purple) match local conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$Imp(X_m) = \frac{1}{N_T} \sum_{t} \sum_{t \in T: v(S_t) = X_m} p(t) \Delta i(s_t, t) \tag{3}$$

where T represents all of the nodes in a given tree, X_m a given variable, s_t, a given split, and $p(t)$ the proportion of samples that are likely to reach node t.

This value represents the weighted impurity for a particular variable averaged across all trees in the forest. Scikit-learn normalizes these values by dividing the importance value for each feature by the total importance value for all features. For ease of comprehension, the final scores calculated by the model were adjusted to fall between 0 and 100.

Variables that occur towards the top of a tree more frequently are deemed more important because of their role in creating more accurate predictions. Feature importance values were compared for each of the three models to evaluate the similarities and differences in features that are identified as having the most influence on model accuracy.

4. Results

4.1. Net construction

A temporal heat map of *net construction* at monthly time steps exhibits general downslope trends and few outliers verified by several important events (Fig. 2, see Fig. A2 in Appendix A for heat maps depicting independent counts of construction and demolition). Most notably, the stark transition from net positive construction to net negative in 2008 coincides with the housing crisis and the start of the Great Recession. An additional validation point occurs with seemingly anomalous rates of demolition in January to February of 2004 (Fig. 2). These apparent outliers follow a severe derecho, or straight-line windstorm, in July of 2003 referred to locally as "Hurricane Elvis'' which caused significant damage throughout the city and likely led to a spike in demolition as property owners sought to repair or rebuild. In addition to understanding the temporal patterns that emerge from *net construction*, its spatial distribution of high rates of *net construction* is in line with where the activity is actually occurring (Fig. 3).

Spatial heat maps provide meaningful insight into how the county has changed over this 15-year span and demonstrates differing patterns in the years leading up to the housing crisis in 2008 and the years after. Prior to the housing crisis, construction and renovation activity was most pronounced in the eastern portion of the county, with a particular concentration around a major north–south corridor in the residential suburbs in 2005 and 2006. There was also a high level of construction activity in the western edge of the county, mainly around the central business district (CBD), although this activity is overlapping with a significant amount of demolition centered around the southern edge of the CBD.

These patterns change in the years following the recession, particularly around 2010, when the development in the eastern portion of the county nearly disappears, and a greater amount of demolition within the core of the city becomes more apparent. Development seems to increase around 2011 through 2012 and much of the construction activity is concentrated around the middle of the county along an east–west axis. While there is some data in 2016, the lack of a complete record for the year is apparent in the sparse map for that year.

With over 436 square miles, most of the county experienced net negative construction activity than net positive which had a total of 341 total square miles. Predictably, the tracts that experienced more demolition activity tended to have older housing stock with a median building age over 62 years (27 years for net positive), greater concentrations of non-white population (89% versus 44% for net positive), and smaller parcel size (median 9,283 square feet versus 11,798 for net positive).

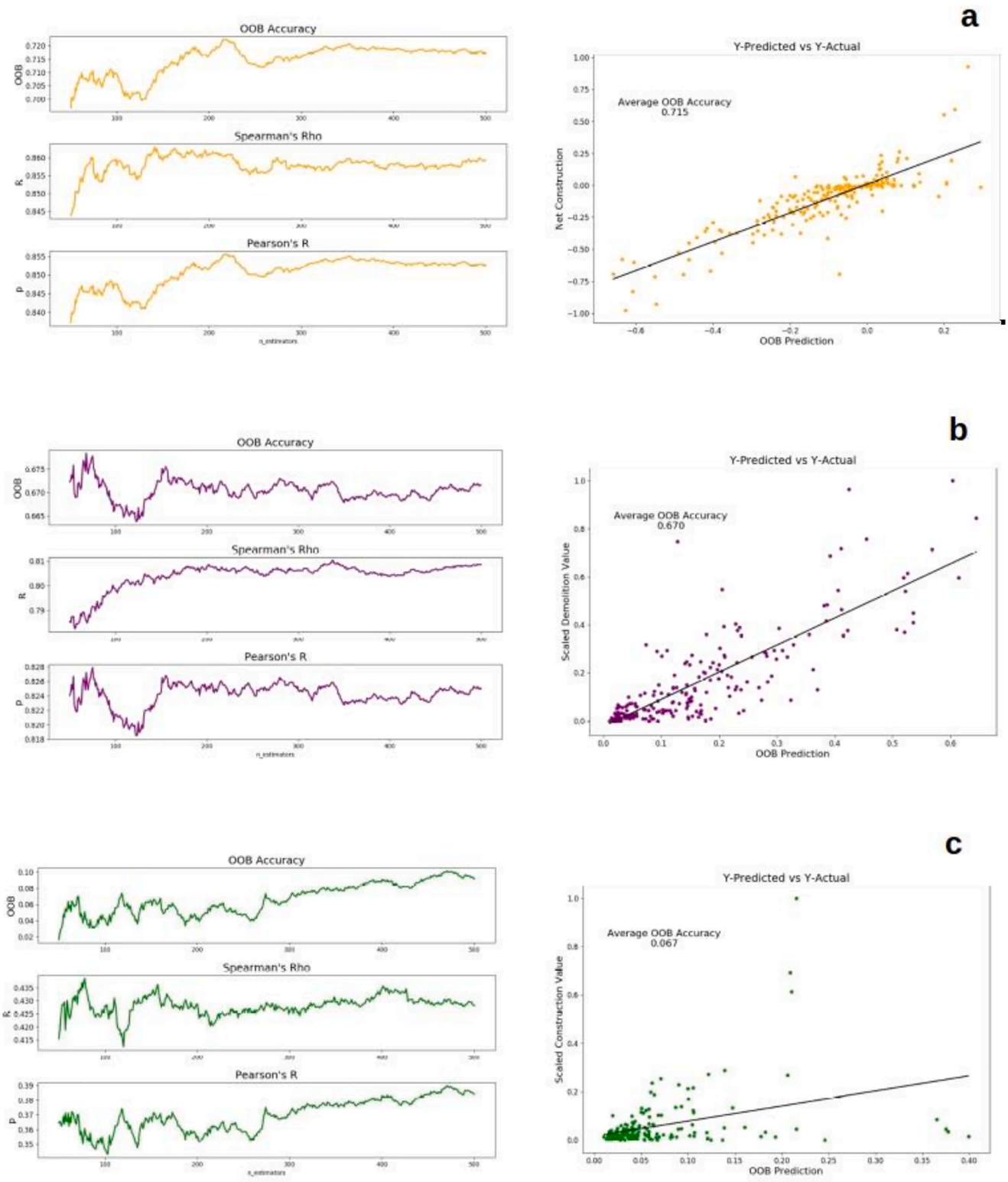

Fig. 4. Model performance and accuracy. Model performance and accuracy for Net construction (a), Demolition (b), and Construction (c). The line graphs on the left reveal the out-of-bag (OOB), Spearman's, and Pearson's scores for each of the three models. The scatter plot on the right shows the actual value for each model on the y-axis and the predicted score for each on the x-axis along with the line of best fit.

4.2. Model accuracy

The accuracy of the three models (hereafter, Net construction, Construction, and Demolition refer to RF models by the corresponding dependent variables used) varies widely with the best performance of the Net construction model, a maximum OOB accuracy value of 78%, followed by the Demolition model of 68% (Fig. 4). The lowest accuracy model was for Construction where the OOB accuracy was less than 10%. A scatter plot comparison between the actual versus predicted scores revealed the vast discrepancy between the various models (Fig. 4).

Table 3
Feature importance scores. Feature importance scores derived from the Net construction, Demolition, and Construction random forest models (10 highest values for each model in bold, NAICS: North American Industrial Classification System).

Category	Variable	Description	Net Importance	Demo Importance	Const Importance
Land and housing	number_of_patches	Number of patches within zone	**100.00**	0.00	0.00
	pct_hu_vcnt	Percent vacant housing	**39.63**	**31.58**	0.48
	pct_dev	Percent of developed parcels	**26.75**	**76.60**	2.63
Demographics	pct_nonwh	Percent non-white population	**75.66**	**83.82**	1.79
	pct_bach	Percent with Bachelor's degree	**49.47**	**57.29**	10.00
Accessibility	cmgrdn_dist	Average distance to nearest community garden for all parcels within a Census tract	**43.65**	**25.75**	**62.91**
	commcenter_dist	Average distance to nearest community center for all parcels within a Census tract	**33.68**	4.20	**100.00**
	pol_dist	Distance to nearest police station	11.68	1.50	**63.13**
	pvt_dist	Distance to nearest private school	5.17	1.64	**37.92**
Financial	pct_pov_tot	Total percent of population below poverty	**32.79**	**22.75**	2.47
	hhinc	Median household income	**25.64**	6.00	**17.88**
	mdngrrent	Median gross rent	3.87	2.62	**17.48**
	foreclose	Foreclosures	4.64	**13.72**	**19.34**
Employment	pct_unemp	Percent unemployed	**79.46**	**100.00**	1.92
	pct_manuf	Businesses by Sector: Percent of jobs in Manufacturing (NAICS 31–33)	4.02	**17.13**	2.21
	pct_ag	Businesses by Sector: Percent of jobs in Agriculture and Forestry (NAICS 11)	3.51	0.21	**33.93**
Transportation	mata_stop_sqmi	Transit stops per square mile	13.40	**33.28**	2.04
	mmcnxpsmi	Multimodal connections per square mile	9.17	2.65	**16.17**
	tt30more	Total working age population whose travel time to work is 30 min or more	5.30	1.99	**18.12**

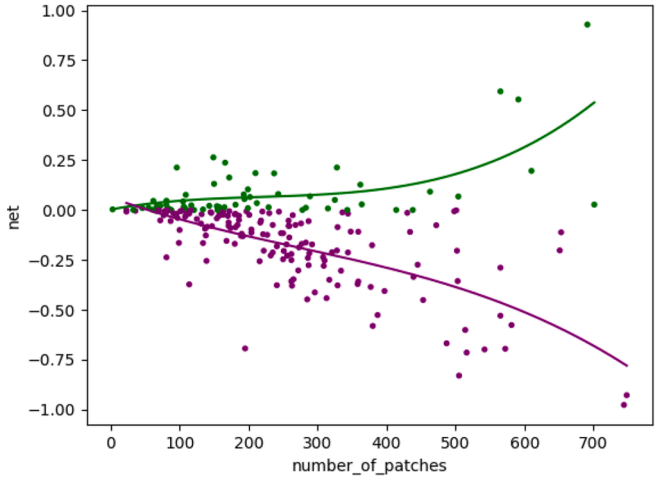

Fig. 5. Number of patches scatter plot. Scatter plot showing net construction value (y-axis) and number of patches (x-axis) with line of best fit for net positive (green) and net negative values (purple).

Although the *net construction* values were fairly tightly coupled, especially around the middle of the range where the score fell between −0.2 and 0.2, the Construction and Demolition values were much more dispersed. Predicted construction scores appeared to be relatively accurate at the low end of the range (i.e., between 0.0 and 0.05) but became much less predictable as the values increased. Such variation in accuracy is likely due to a disparity in the number of construction permits and the number of demolition permits which, when aggregated using census tract boundaries, had a standard deviation of over 1,400 and 48 for respectively (Table 2).

Similar to the OOB results, both Pearson's r correlation (r) and Spearman's rank correlation (ρ) showed the most accurate model of the three was Net construction where r was about 88% and ρ was about 86%. The Demolition model was the second most accurate, but unlike Net construction, r was less accurate at approximately 81%, whereas ρ was almost 83%. Finally, just like the OOB scores, the worst performing model of the three was Construction where r was slightly below 44% and ρ was just below 39%.

It is important to note that although the scores for both Spearman's rank and Pearson's r correlation were higher than the values reported by the OOB score for all three models, the OOB value still provides a more accurate account of how well the model performs. OOB is the result of testing unseen samples that are drawn from the full dataset against the predicted values and not simply a comparison between actual and predicted values as is the case with Spearman and Pearson.

4.3. Feature importance scores

When the top ten most important features are examined in each of the three models (Table 3), the features with the greatest influence on Net construction represent an intersection of the top variables for Construction and Demolition which further demonstrates its utility as a comprehensive measure of urban change. Net construction and Demolition shared a total of six variables; percent of population with bachelor's degree (*pct_bach*), percent developed parcels (*pct_dev*), total percent at or below poverty (*pct_pov_tot*), percent unemployment (*pct_unemp*), and percent non-white population (*pct_nonwh*). Net construction and Construction shared two: distance to nearest community center (*commcenter_dist*), and total household income (*hhinc*).

The fact that the Net construction model shared a number of features in common with both independent models demonstrates its ability to capture those characteristics that are unique to each of those models. However, the emergence of features that are exclusive to Net construction also hints at its ability to identify characteristics that only appear when construction and demolition are studied simultaneously. For example, the most important feature for Net construction, the number of patches within each tract (*number_of_patches*), the landscape metric analogous to land use mix, did not appear at all for either Construction or Demolition. When visualized in detail for *net construction*, the number of patches demonstrated a contradictory pattern between net positive and net negative values (Fig. 5).

The top features for Demolition and Construction were percent unemployed (*pct_unemp*) and average distance to nearest community center (*commcenter_dist*), respectively, which appeared as one of the top ten features for the Net construction model. This seems to further indicate that Net construction as a composite value not only was able to identify many of the influencing factors that relate to Construction and Demolition, but also able to capture additional characteristics that may only result when construction and demolition activity are analyzed simultaneously.

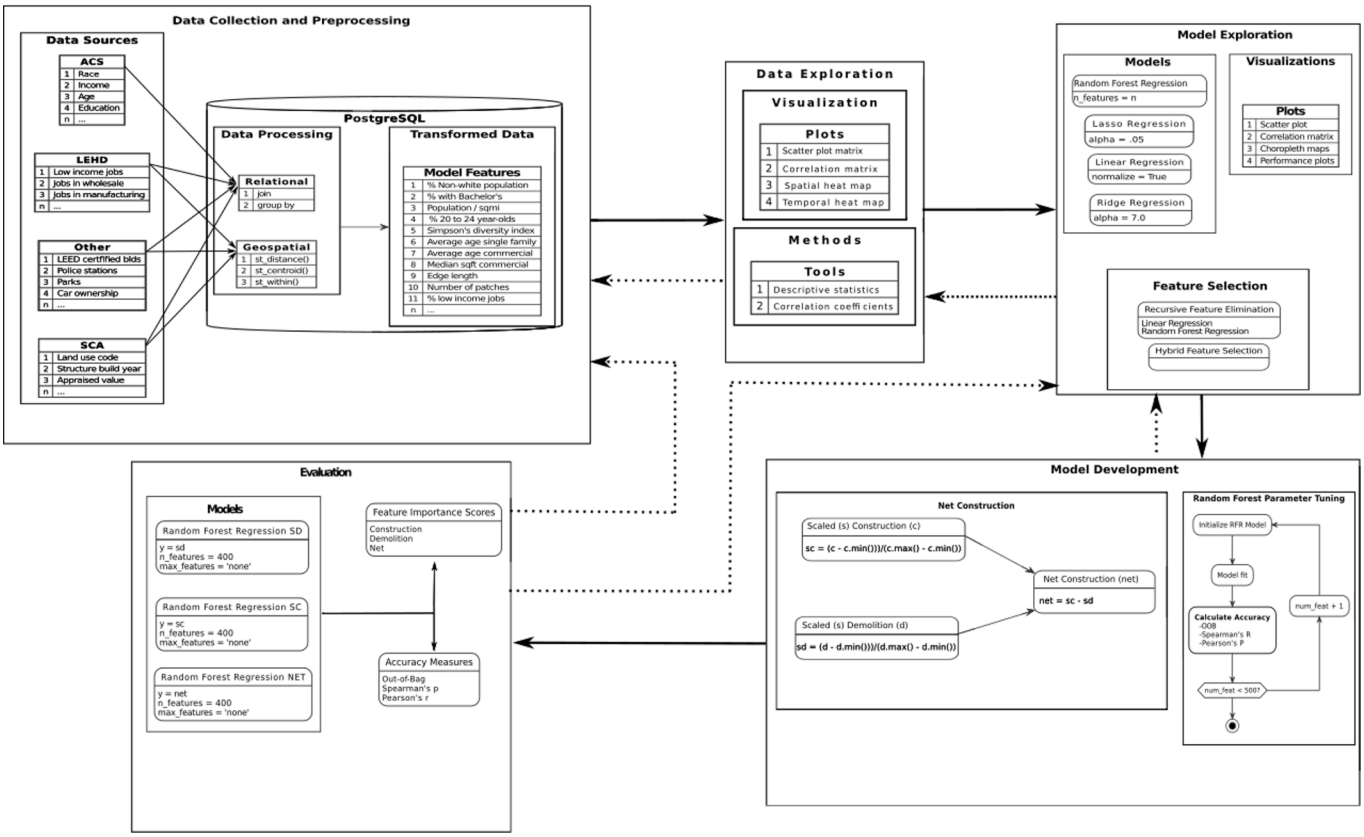

Fig. A1. Workflow diagram. Data were collected, processed, and analyzed in an iterative workflow. Although the analysis progressed in a linear fashion as represented by the solid arrows, previous steps would be revisited as new variables were added or after modifying existing variables (represented with dashed lines).

There was less intersection between the two independent models (i.e., Construction and Demolition) which only share two variables in common, average distance to community gardens (*cmgrdn_dist*) and foreclosures (*foreclose*). Another notable distinction between the independent models were the categories that contain the most important features. For Construction, the four most important features, average distance to nearest community center (*commcenter_dist*), distance to nearest police station (*pol_dist*), average distance to nearest community garden (*cmgrdn_dist*), and distance to nearest private school (*pvt_dist*), all relate to various measures of Accessibility while features from Demolition showed importance within the category of Land and housing and Demographics.

5. Discussion

This discussion contextualizes the results of this study within the current literature while also connecting the methodology and data to help inform and guide policy. In closing, this section will identify current gaps in the research and discuss future research opportunities.

5.1. Findings and contributions

This research has contributed to the debate surrounding community change in a number of ways. First, the effectiveness of *net construction* in assessing community change suggests that this measure is a practical indicator as an outcome variable to model urban development, providing a better understanding of the built environment's impact on community change. Not only did *net construction* provide an accurate depiction of urban change before and after the 2008 recession, but it also demonstrated better predictive capability with RF models when compared against the models with construction or demolition alone.

Another contribution is in the use of a supervised learning method to model urban change. RF regression was able to incorporate a wide variety of relationships and data types, helping to capture some of the complexity that drives community change. One advantage of this approach was the ability to incorporate a multi-disciplinary complement of input such as socio-economic variables, urban form characteristics, and landscape metrics into a single analysis to compare which of them are most critical through feature importance scores. The scores produced by the three models, Construction, Demolition, and Net construction, showed both similarities and differences with previous studies.

When Construction and Demolition were modeled independently, this research found similarities with studies such as Knaap et al. (2007) and Lowry and Lowry (2014), or Yin and Silverman (2015), which emphasized the role of density, proximity, or street networks in characterizing a community's urban form. However, the presence of numerous demographic and employment variables like percent unemployment (*pct_unemp*), percent non-white population (*pct_nonwh*), and percent with bachelor degrees (*pct_bach*) at the top of the importance list for Demolition demarks a notable departure from previous studies that have found little predictive capability between socio-economic variables and demolition (Hollander et al., 2019; Weber et al., 2006). While a more nuanced treatment of demolition types could help explain this discrepancy (Paredes and Skidmore, 2017), it is also conceivable that the use of random forest might help identify more complex influences that have not previously been discussed.

The presence of distance to community gardens (*cmgrdn_dist*) at the top of the list of important variables for all three models lends further support to the potential for ML models like RF to discover hidden influence over urban change. While there is ample discussion on the connection between abandonment, vacant lots, and community gardens (Anderson & Minor, 2017; Chin, 2021; Gobster et al., 2020), the fact that their existence or absence carries such significant influence when

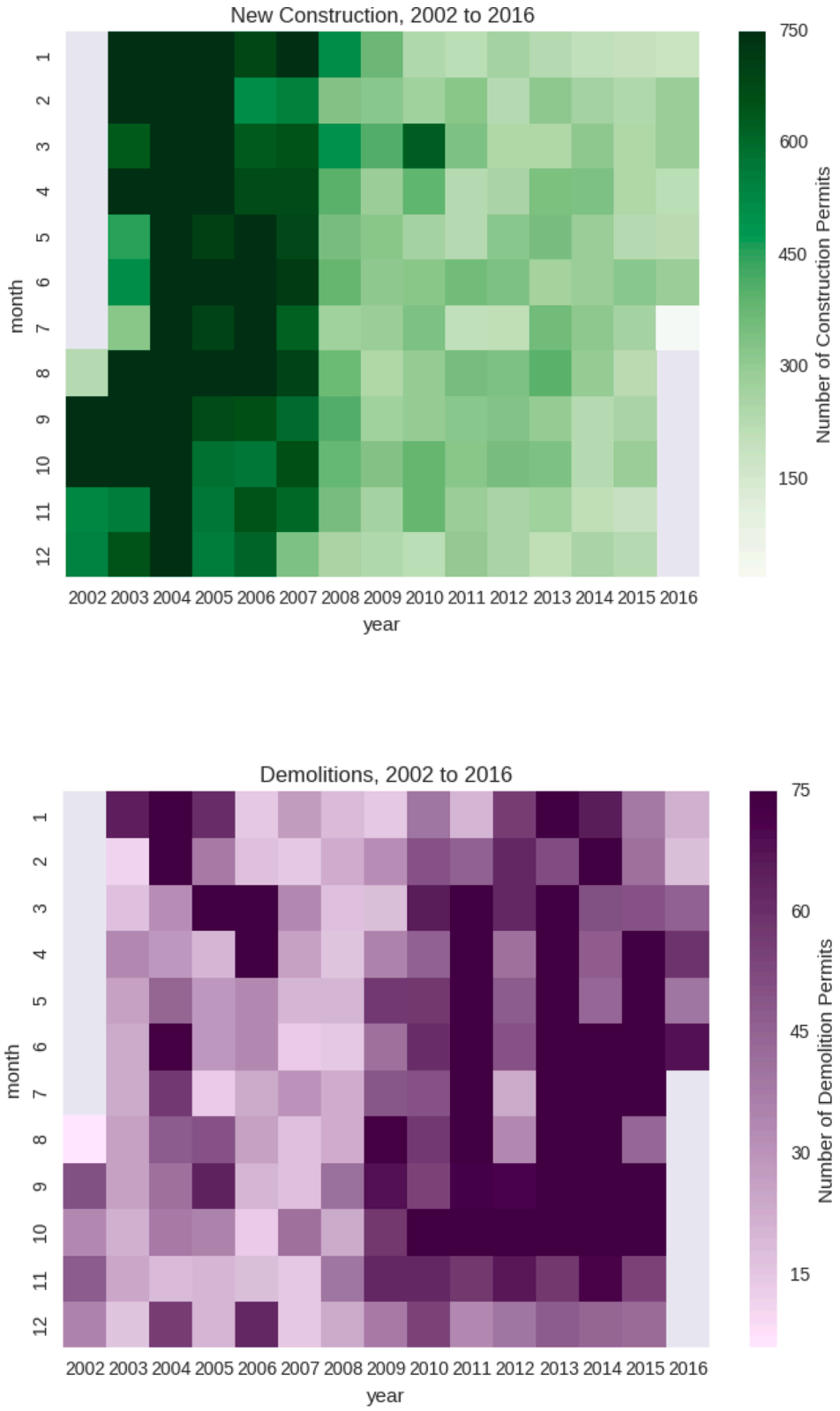

Fig. A2. Temporal heat maps for independent counts of construction (green) and demolition (purple) permits.

predicting Construction or Net construction values hints at the larger capability that ML offers for analyzing pressing urban challenges.

Of greater significance, however, is the ability of net construction to highlight the underlying characteristics of areas experiencing greater construction or demolition *simultaneously*. As discussed previously, community gardens correlate to variables like vacancy, so it stands to reason that they would appear in areas with greater rates of demolition. Whereas proximity to community gardens is important for areas with net negative scores, distance away from community gardens is important for areas with positive scores which would also be true for other location measures as well. These differences could be attributed to a contrast in land development patterns between the two locations. The

Table A1
Variable sources and descriptions.

U.S. Census Bureau; American Community Survey, 2012–2016	Average number of hours worked per week	Percent with Bachelor's degree	Percent unemployed
	Median household income	Percent vacant housing	Percent with Less than 1 year college
	Dwelling units per acre	Percent in labor force	Percent with More than 1 year, no degree
	Median years at current residence	Percent with Master's degree	Total population per square mile
	Median gross rent	% receiving Medicaid	Total working age population whose primary mode of travel to work is bicycle
	Median home price	Percent with Diploma	Total working age population whose primary mode of travel to work is single-occupancy automobile
	Percent population under 19 years of age	Percent non-white population	Total working age population whose primary mode of travel to work is public transit
	Percent population 20 to 24	Percent owner-occupied housing	Total working age population whose primary mode of travel to work is walking
	Percent population 25 to 34	Percent with Doctorate degree	Total working age population whose travel time to work is 15 to 19 min
	Percent population 35 to 49	Total percent of population below poverty	Total working age population whose travel time to work is 15 min or less
	Percent population 50 to 66	Percent with Professional degree	Total working age population whose travel time to work is 30 min or more
	Percent population 67 and over	Percent renter-occupied housing	
U.S. Census Bureau, Longitudinal Employer-Householder Dynamics	Percent of jobs in Agriculture and Forestry (NAICS 11)	Percent of jobs in Health Care (NAICS 62)	Percent of jobs in Public Administration (NAICS 92)
	Percent of jobs in Arts (NAICS 71)	Percent of jobs in Health Care (NAICS 62)	Percent of jobs in Real Estate (NAICS 53)
	Percent of jobs in Arts (NAICS 71)	Percent of jobs in Information (NAICS 51)	Percent of jobs in Retail (NAICS 44-45)
	Percent of jobs in retail, service, or other commercial industries	Percent of low and moderate income jobs	Percent of jobs in Transportation (NAICS 48-49)
	Percent of jobs in Construction (NAICS 23)	Percent of jobs in Manufacturing (NAICS 31-33)	Percent of jobs in Utilities (NAICS 22)
	Percent of jobs in Education (NAICS 61)	Percent of jobs in Management (NAICS 55)	Percent of jobs in Waste Management (NAICS 56)
	Percent of jobs in Education (NAICS 61)	Percent of jobs in Mining (NAICS 21)	Percent of jobs in Wholesale (NAICS 42)
	Percent of jobs in Finance (NAICS 52)	Percent of jobs in Other Services (NAICS 81)	
	Percent of jobs in Food Services (NAICS 72)	Percent of jobs in Professional Services (NAICS 54)	
2004, 2017 Shelby County Assessor's Certified Roll	Percent change in property values 2004–2017	Total number of living units	Median number of feet for residential parcel perimeter
	Average age of all structures	Number of patches within zone	Median total number of rooms in structure Table A1
	Average age commercial buildings	Median parcel acreage	Median square footage commercial structures
	Average age single family homes	Number of patches per unit of area	Median square footage total living area for residential structures
	Probability that two random adjacent cells are from the same class	Percent commercial square footage	Shannon Diversity index
	Measure of diversity using Simpson Diversity Index	Percent of developed parcels	Total edge length
	Edge length per area unit	Percent multifamily housing	
	Measure of "edginess", standardized and adjusted for landscape size	Percent vacant parcels	
	Proportion of total landscape comprised by largest patch		
Landscape Metrics	Contagion	Edge density	Landscape patch index
	Largest patch index	Number of patches	Patch density
	Shannon diversity index	Total edge	
Center for Applied Earth Science and Engineering Research, University of Memphis	Average distance to nearest childcare center for all parcels within a census tract	Average distance to nearest library for all parcels within a census tract	Percent coverage tree canopy
	Average distance to nearest community center for all parcels within a census tract	Distance to nearest middle school	Percent impervious surface
	Average distance to nearest elementary school for all parcels within a census tract	Multimodal connections per square mile	Distance to nearest police station
	Distance to nearest fire station	Total square miles of open space	Distance to nearest private school
	Average distance to nearest high school for all parcels within a census tract	Distance to nearest park	Percent of streets with sidewalks

(continued on next page)

Table A1 (*continued*)

	Average distance to nearest hospital for all parcels within a census tract	Total number of parks per capita	
Memphis Area Transit Authority	Miles of MATA routes per square mile	Transit stops per square mile	Bus ridership - Number passengers on per square mile
U.S. Green Building Council	Number of certified green buildings per square mile	Number of street intersections per square mile	
Center for Neighborhood Technology, Housing + Transportation Index	Number of cars per household	Employment Mix Index (Low 0-High 100)	Total annual vehicle miles traveled
Memphis and Shelby County Office of Sustainability, Mid-South Regional Greenprint	Bicycle friendliness index based on a combination of variables including low speed streets, completed bike lanes and greenways, and multiple street connections (0 = unfriendly to 238 = friendly)	Job access (0 = low accessibility to 10 = high accessibility)	Walkability index based upon the presence or absence and quality of footpaths and sidewalks, traffic and road conditions, land use patterns, building accessibility, and safety. (0 = low walkability to 60 = high walkability)
	Average distance to nearest community garden for all parcels within a census tract	Average distance to nearest farmers market for all parcels within a census tract	
Memphis Urban Area Metropolitan Planning Organization	Total miles of bicycle lanes per square mile	Greenways (Existing and Proposed) - Miles per square mile	
National Highway Traffic Safety Administration Fatality Analysis Reporting System	Number of pedestrian and bicycle injuries per capita		
Shelby County Regional GIS (ReGIS)	Total number of foreclosures		
U.S. Department of Housing and Urban Development, Housing Affordability Data System	Number of affordable housing units within 0.5 mile of green space	Percent of affordable housing stock	Transit accessibility (0 = low accessibility to 100 = high accessibility)
U.S. Environmental Protection Agency, Uniform Resource Locator	Number of brownfield sites per capita		
U.S. Geological Survey, National Hydrography Dataset	Wetland acres per square mile		

Table A2
Relevant tools. A variety of open source tools were selected to test different scenarios by creating terminal-based commands that supported running the same command repeatedly with different input, evaluating the output for differences.

Tool	Description	Use(s)	Reference
Python	Open source programming language commonly used throughout data science and machine learning.	- Data collection - Data processing - General	Python Software Foundation, 2020
Pandas	A data manipulation library that supports vectorized data processing, value filtering, plotting, and statistical functions that make data exploration and analysis quick and efficient.	- Exploratory analysis - Visualization - Data manipulation	McKinney, 2010
Matplotlib	Visualization library based upon Matlab that supports scientific plotting through a customizable API (application programming interface)	- Data visualization	Hunter, 2007
Seaborn	An additional visualization library built on top of Matplotlib that simplifies some of the more complicated classes within its API	- Data visualization	Waskom, 2020
Numpy	The foundation for most libraries within the Python Scientific stack (SciPy) with optimized numerical data structures like N-dimensional arrays (ndarrays) and a range of statistical methods	- Data analysis - Data manipulation - Basic statistical operations	van der Walt et al., 2011
Scikit-Learn	A Python-based library for machine learning and artificial intelligence. In addition to a range of ML algorithm implementations, it also contains numerous methods for preprocessing and data validation	- Machine learning - Data preprocessing	Pedregosa et al., 2011
SQL Alchemy	A Python-based library for connecting to, manipulating, and managing relational databases	- Database connection and maintenance - Executing queries and extracting data	Bayer, 2020
PyLandStats	Python package for calculating a range of landscape metrics	- Generating landscape metrics	Bosch, 2019
PostgreSQL	Open source database management system (DBMS)	- Data maintenance - Data processing - Analysis	The PostgreSQL Global Development Group, 2020
PostGIS	Open source extension for PostgreSQL for geospatial data management and processing.	- Merging geospatial and tabular data - Geospatial analysis	PostGIS, 2020
QGIS	Open source geospatial information system.	- Data processing - Geospatial visualization - Geospatial analysis	Open Source Geospatial Foundation Project, 2020
Click			Ronacher, 2014

(*continued on next page*)

Table A2 (*continued*)

Tool	Description	Use(s)	Reference
	A Python packaged used for generating command line interfaces (CLI) from Python modules.	- Create command line interface (CLI) for iterative approach for analysis and visualization	

former tends to occur in more urban locations and the latter in more suburban areas. A defining characteristic of these suburban tracts is a greater distance to both services and amenities, which is further supported by two transportation variables in the list of top ten features for the Construction model. Accounting for net construction, our modeling approach can capture this inverse relationship while the independent models are not.

This same inverse relationship can be seen with socio-economic variables as well. For example, areas that experienced net negative construction had a median percentage of non-white population of over 89% and a median percentage of the population with a bachelor's degree of 9%. In contrast, tracts with net positive construction had median values of 42% and 25% respectively. So, although demographic variables are not featured prominently in the Construction model, incorporating demolition into the analysis accentuates their importance. This finding suggests that the demolition or construction model alone does not serve as an accurate analysis tool to make an informed decision. Instead, the net construction model can be used to improve policy effectiveness as we illustrate in the next section.

Landscape metrics, which have been gaining in application in urban morphology as better data and computational capabilities become available (Angel et al., 2016; Taubenböck et al., 2019; Lemoine-Rodríguez et al., 2020), warrant further consideration. In particular, *number_of_patches*, the top feature for Net construction, relates to the quantity and diversity of adjacent land uses analogous to land use mix in urban study literature (Herold et al., 2003; Song et al., 2013). Originally drawn from landscape ecology, the concept of a patch typically refers to "communities or species assemblages surrounded by a matrix with a dissimilar community structure or composition (Godron, 1981, p. 734)." This concept has been extended to the urban environment by defining patches as collections of contiguous land with similar land uses and has been used to evaluate urban sprawl and fragmentation (Abrantes et al., 2019). Some of the earliest applications of landscape metrics within urban context can be found in transportation literature, most notably Hess et al. (2001) who found that measures like the number of patches provided a detailed mechanism for studying land use mix. In addition, patches have also been found to play a fundamental role in urban productivity and healthy residential development (Jia et al., 2019).

The relationship between the number of patches and *net construction* reveals a near mirror image between rates of positive and negative activity that increase and decrease respectively as the number of patches go up (Fig. 5). This seemingly contradictory pattern resembles the SLOSS (single large or several small) debate that once dominated conservation journals until Soulé and Simberloff (1986) advocated for both "bigness" and "multiplicity". It is the balance between a habitat's size and the complexity of its shape that is essential for determining how suitable it is for a habitat (Clifton et al., 2008). In the case of this study, both the size and shape of a parcel for any given land use seem to play a pivotal role in determining the level of either construction or demolition.

5.2. Policy implications

Net construction presents municipal leaders with an effective mechanism by which to measure and assess neighborhood stability or resilience. While *net construction* offers insight into neighborhoods that may be experiencing higher construction or demolition activity rates, the neighborhoods that hover around the net neutral range also warrant careful consideration. As *net construction* represents the relative difference between construction and demolition activities, scores that fall close to zero reflect locations where construction and demolition occur at roughly equal rates. Furthermore, by calculating a composite *net construction* score over a 15-year period that includes a major economic catastrophe such as the 2008 housing crisis, we were able to identify neighborhoods that remained stable throughout the crisis or rebounded in relatively short order. These "net neutral" neighborhoods embody Meerow et al. (2016)'s definition of urban resilience in their ability "to maintain or rapidly return to desired functions in the face of a disturbance" (p.45) and provide municipalities with a potential case study into the characteristics that may be critical to building healthy and resilient communities. Additionally, the concept of a net neutral neighborhood offers another perspective for a deeper understanding of the relationship between urban areas and "socio-economic-ecological" processes. Because these neighborhoods are identified using a routine dataset maintained by most municipal jurisdictions, it allows for greater comparisons from one region to another (Sapena et al., 2020).

One area of policy that could greatly benefit from an access to a diverse and rich set of data relates to demolitions. Cities typically regulate how structures should be demolished, but they often lack any policy to strategically guide or steer that demolition. In its most recent comprehensive plan, the City of Memphis discourages the removal of historically significant structures in order to encourage good design characteristics that are contextually relevant to the surrounding community (City of Memphis and Memphis and Shelby County Division of Planning and Development, 2019). But Mallach (2012) argued for a more deliberate and targeted approach to address a range of deleterious effects caused by vacant and abandoned structures. Incorporating a comprehensive body of data could help struggling cities develop guiding principles that could help balance these approaches while simultaneously promoting to ease the downward pressure on property values caused by dilapidated properties.

The challenges caused by the lack of reliable data inside city hall are amplified when the focus expands to include outside entities such as universities, non-profits, or private businesses. This study demonstrated how a fairly common and routine dataset collected by nearly all governments can be re-purposed to discover hidden relationships or highlight unseen characteristics. Increasing the transparency of local government, especially with regard to data and business processes, has been shown to pay dividends financially. Smart Procure, a database provider that collects and disseminates local and state purchasing data, initially relied on Freedom of Information Act (FOIA) requests to obtain routine data relating to what governments bought and how much they spent on the acquisition. A simple idea that leverages routine public information not only helped build a highly successful technology company but has also helped local governments improve their purchasing programs by increasing transparency throughout the process (Goldstein and Dyson, 2013).

5.3. Future research opportunities

The results of this study have shown some of the underlying characteristics that contribute to higher rates of development activity in one neighborhood versus another. However, there remain a few questions beyond the scope of this effort that could help improve our overall understanding of the complex nature of urban change.

The first relates to time and the dynamic change that characterizes contemporary urban development whereby the physical environment is continuously modified to fit the needs and demands of current residents. Although the permit data covered a 15-year time frame that occurred both before and after the housing crisis of 2008, the analysis largely treated these data as though they were static and represented a snapshot in time. While the accuracy achieved by the models demonstrate the

feasibility of the method, it is possible that an approach that incorporates time into the analysis would only increase the accuracy of the predictions. By aggregating the permit data into time blocks that overlap ACS sample years, it would be possible to extract greater detail around the specific characteristics that contribute to development activity. A more rigorous temporal approach with a more granular exploration may offer greater insight to understand the fluctuation of permitting over time along with how the underlying factors shifted.

Another area for further investigation is in the aggregating unit used for the analysis. Given the characteristics built into census tract boundaries, they were a logical unit for grouping non-census derived variables for summarization. One limitation to this approach, however, is that the resulting sample size is relatively low. Although RF models are able to handle low sample sizes with relatively high-dimensional feature spaces, the use of another aggregating unit like census block groups might offer greater detail.

Another approach to increasing sample size would be to expand the analysis beyond Memphis and Shelby County to incorporate multiple cities into the modeling frameworks. Two main key data we employed in the study, building permits and parcel data, are standard data sources that all municipal governments collect. Expanding this study to include additional cities would both increase the available sample size while simultaneously testing whether there are regional differences in factors that influence construction and demolition activity. Though there are bound to be regional differences in regard to some variables, we would propose that some of the key urban form measures likely remain unchanged (Lemoine-Rodríguez et al., 2020).

6. Conclusion

This research explored the use of *net construction* as a standardized measure for urban development using construction, renovation, and demolition permit data and identified the factors that influence community change. Both temporal and spatial visualizations showed promise in understanding community change and provided a more holistic view on how the built environment affects development patterns. Expanding upon this, we utilized RF regression as a method capable of handling a broad array of relationships and data types. RF empowered the analysis of the underlying characteristics that were most closely associated with community change indicated by *net construction*. Questions persist as modeling has its inherent limitations, and some determining factors of urban change are likely not incorporated into our model. Even with a complete set of data inputs, algorithms alone will not predict all urban changes. As an example, local politics and planning policies can play a key role in influencing development practices.

Many urban problems remain unsolvable using traditional analytic techniques but may, in fact, have a solution when approached from another perspective; the expansion of ML into urban geography presents such an opportunity. By increasing the volume and enriching the diversity of data fed into new models, the potential to discover new patterns or relationships will inevitably follow. However, despite the potential that ML offers for addressing some of the complex challenges that confront cities, it is also important to acknowledge the risk that comes with any technology that experiences wider adoption beyond its scientific origin (Janssen and Kuk, 2016). While numerous platforms are lowering barriers to ML, an incomplete understanding of the implications or assumptions built into an algorithm can lead to unintended consequences.

CRediT authorship contribution statement

Nathan Ron-Ferguson: Writing – original draft, Methodology, Investigation, Formal analysis. **Jae Teuk Chin:** Writing - review & editing, Project administration. **Youngsang Kwon:** Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

References

Abrantes, P., Rocha, J., Marques da Costa, E., Gomes, E., Morgado, P., & Costa, N. (2019). Modelling urban form: A multidimensional typology of urban occupation for spatial analysis. *Environment and Planning B: Urban Analytics and City Science, 46*(1), 47–65. https://doi.org/10.1177/2399808317700140.

Alberti, M. (1999). Urban patterns and environmental performance: What do we know? *Journal of Planning Education and Research, 19*(2), 151–163. https://doi.org/10.1177/0739456X9901900205.

Alberti, M. (2005). The effects of urban patterns on ecosystem function. *International Regional Science Review, 28*(2), 168–192. https://doi.org/10.1177/0160017605275160.

Alberti, M., Botsford, E., & Cohen, A. (2001). Quantifying the urban gradient: Linking urban planning and ecology. In J. M. Marzluff, R. Bowman, & R. Donnelly (Eds.), *Avian Ecology and Conservation in an Urbanizing World* (pp. 89–115). US: Springer. https://doi.org/10.1007/978-1-4615-1531-9_5.

Anderson, E. C., & Minor, E. S. (2017). Vacant lots: An underexplored resource for ecological and social benefits in cities. *Urban Forestry & Urban Greening, 21*, 146–152. https://doi.org/10.1016/j.ufug.2016.11.015.

Angel, S., Blei, A., Parent, J., Lamson-Hall, P., & Sanchez, N. G. (2016). Atlas of Urban Expansion: The 2016 Edition, Volume 1: Areas and Densities (2016th ed., Vol. 1). NYU Urban Expansion Program; UN-Habitat; Lincoln Institute of Land Policy. https://www.lincolninst.edu/sites/default/files/pubfiles/atlas-of-urban-expansion-2016-volume-1-full.pdf.

Auerbach, J., Blackburn, C., Barton, H., Meng, A., & Zegura, E. (2020). Coupling data science with community crowdsourcing for urban renewal policy analysis: An evaluation of Atlanta's Anti-Displacement Tax Fund. *Environment and Planning B: Urban Analytics and City Science, 47*(6), 1081–1097. https://doi.org/10.1177/2399808318819847.

Bayer, M. (2020). SQLAlchemy—The Database Toolkit for Python. https://www.sqlalchemy.org/.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag.

Boeing, G. (2018). Measuring the Complexity of Urban Form and Design (23(4); URBAN DESIGN International, pp. 281–292). Center for Open Science. doi: 10.1057/s41289-018-0072-1.

Boessen, A., Hipp, J. R., Butts, C. T., Nagle, N. N., & Smith, E. J. (2018). The built environment, spatial scale, and social networks: Do land uses matter for personal network structure? *Environment and Planning B: Urban Analytics and City Science, 45*(3), 400–416. https://doi.org/10.1177/2399808317690158.

Bosch, M. (2019). PyLandStats: An open-source Pythonic library to compute landscape metrics. *PLOS ONE, 14*(12), e0225734. https://doi.org/10.1371/journal.pone.0225734.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cao, Q., Luan, Q., Liu, Y., & Wang, R. (2021). The effects of 2D and 3D building morphology on urban environments: A multi-scale analysis in the Beijing metropolitan region. *Building and Environment, 192*, 107635. https://doi.org/10.1016/j.buildenv.2021.107635.

Carmona, M. (2019). Place value: Place quality and its impact on health, social, economic and environmental outcomes. *Journal of Urban Design, 24*(1), 1–48. https://doi.org/10.1080/13574809.2018.1472523.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc., 76.

Charles, S. L. (2011). Suburban Gentrification: Understanding the Determinants of Single-family Residential Redevelopment, A Case Study of the Inner-Ring Suburbs of Chicago, IL, 2000-2010. Joint Center for Housing Studies of Harvard University Cambridge, MA. http://140.247.195.238/sites/jchs.harvard.edu/files/w11-1_charles.pdf.

Charles, S. L. (2014). The spatio-temporal pattern of housing redevelopment in suburban Chicago, 2000–2010. *Urban Studies, 51*(12), 2646–2664. https://doi.org/10.1177/0042098013506045.

Chin, J. T. (2021). The shifting role of public–private partnerships in vacant property redevelopment. *Land Use Policy, 105*, 105430. https://doi.org/10.1016/j.landusepol.2021.105430.

City of Memphis, & Memphis and Shelby County Division of Planning and Development. (2019). Memphis 3.0: The Comprehensive Plan of the City of Memphis, Tennessee.

Clapp, J. M., & Wang, Y. (2006). Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics, 59*(2), 259–284. https://doi.org/10.1016/j.jue.2005.10.003.

Clifton, K., Ewing, R., Knaap, G., & Song, Y. (2008). Quantitative analysis of urban form: A multidisciplinary review. *Journal of Urbanism: International Research on*

Placemaking and Urban Sustainability, 1(1), 17–45. https://doi.org/10.1080/17549170801903496.
Coulton, C. (2012). Defining Neighborhoods for Research and Policy. *Cityscape: A Journal of Policy Development and Research, 14*(2), 231–236.
Coulton, C. J., Korbin, J., Chan, T., & Su, M. (2001). Mapping Residents' perceptions of neighborhood boundaries: A methodological note. *American Journal of Community Psychology, 29*(2), 371–383. https://doi.org/10.1023/A:1010303419034.
Dye, R. F., & McMillen, D. P. (2007). Teardowns and land values in the Chicago metropolitan area. *Journal of Urban Economics, 61*(1), 45–63. https://doi.org/10.1016/j.jue.2006.06.003.
Ewing, R., & Hamidi, S. (2015). Compactness versus Sprawl: A review of recent evidence from the United States. *Journal of Planning Literature, 30*(4), 413–432. https://doi.org/10.1177/0885412215595439.
Ghosh, A., Sharma, R., & Joshi, P. K. (2014). Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion. *Applied Geography, 48*, 31–41. https://doi.org/10.1016/j.apgeog.2014.01.003.
Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry, 56*(1), 114–137. https://doi.org/10.1111/ecin.12364.
Gobster, P. H., Rigolon, A., Hadavi, S., & Stewart, W. P. (2020). Beyond proximity: Extending the "greening hypothesis" in the context of vacant lot stewardship. *Landscape and Urban Planning, 197*, 103773. https://doi.org/10.1016/j.landurbplan.2020.103773.
Godron, M. (1981). Patches and structural components for a landscape ecology. *BioScience, 31*(10), 733–740. https://doi.org/10.2307/1308780.
Goldstein, B., & Dyson, L. (Eds.). (2013). *Beyond transparency: Open data and the future of civic innovation*. Code for America Press.
Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface, 2*(3), 349–360. https://doi.org/10.4310/SII.2009.v2.n3.a8.
Herold, M., Liu, X., & Clarke, K. C. (2003). Spatial metrics and image texture for mapping urban land use. *Photogrammetric Engineering & Remote Sensing, 69*(9), 991–1001. https://doi.org/10.14358/PERS.69.9.991.
Hess, P., Moudon, A. V., & Logsdon, M. G. (2001). Measuring land use patterns for transportation research. *Transportation Research Record, 1780*(1), 17–24. https://doi.org/10.3141/1780-03.
Hollander, J., Johnson, M., Drew, R. B., & Tu, J. (2019). Changing urban form in a shrinking city. *Environment and Planning B: Urban Analytics and City Science, 46*(5), 963–991. https://doi.org/10.1177/2399808317743971.
Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy, 82*, 657–673. https://doi.org/10.1016/j.landusepol.2018.12.030.
Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011.
Jia, Y., Tang, L., Xu, M., & Yang, X. (2019). Landscape pattern indices for evaluating urban spatial morphology – A case study of Chinese cities. *Ecological Indicators, 99*, 27–37. https://doi.org/10.1016/j.ecolind.2018.12.007.
Knaap, G.-J., Song, Y., & Nedovic-Budic, Z. (2007). Measuring patterns of urban development: New intelligence for the war on Sprawl. *Local Environment, 12*(3), 239–257. https://doi.org/10.1080/13549830601183412.
Lai, Y., & Kontokosta, C. E. (2019). Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems, 78*, 101383. https://doi.org/10.1016/j.compenvurbsys.2019.101383.
Lemoine-Rodríguez, R., Inostroza, L., & Zepp, H. (2020). The global homogenization of urban form. An assessment of 194 cities across time. *Landscape and Urban Planning, 204*, 103949. https://doi.org/10.1016/j.landurbplan.2020.103949.
Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. ArXiv: 1407.7502 [Stat]. http://arxiv.org/abs/1407.7502.
Lowry, J. H., & Lowry, M. B. (2014). Comparing spatial metrics that quantify urban form. *Computers, Environment and Urban Systems, 44*, 59–67. https://doi.org/10.1016/j.compenvurbsys.2013.11.005.
Mallach, A. (2012). Laying the groundwork for change: Demolition, urban strategy, and policy reform. Brookings Institution Report. September Washington DC Brookings Metropolitan Policy Program. https://www.brookings.edu/wp-content/uploads/2016/06/24-land-use-demolition-mallach.pdf.
Mazumdar, S., Learnihan, V., Cochrane, T., & Davey, R. (2018). The Built environment and social capital: A systematic review. *Environment and Behavior, 50*(2), 119–158. https://doi.org/10.1177/0013916516687343.
McKinney, W. (2010). pandas: A Foundational Python Library for Data Analysis and Statistics. 9.
Meerow, S., Newell, J. P., & Stults, M. (2016). Defining urban resilience: A review. *Landscape and Urban Planning, 147*, 38–49. https://doi.org/10.1016/j.landurbplan.2015.11.011.
Mora, H., Pérez-delHoyo, R., Paredes-Pérez, J., & Mollá-Sirvent, R. (2018). Analysis of social networking service data for smart urban planning. *Sustainability, 10*(12), 4732. https://doi.org/10.3390/su10124732.
Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences, 114*(29), 7571–7576. https://doi.org/10.1073/pnas.1619003114.
Open Source Geospatial Foundation Project. (2020). QGIS Geographic Information System (Version 3.14) [Computer software]. Open Source Geospatial Foundation Project. https://qgis.org/en/site/.
Ossola, A., Locke, D., Lin, B., & Minor, E. (2019). Greening in style: Urban form, architecture and the structure of front and backyard vegetation. *Landscape and Urban Planning, 185*, 141–157. https://doi.org/10.1016/j.landurbplan.2019.02.014.
Papadomanolaki, M., Verma, S., Vakalopoulou, M., Gupta, S., & Karantzalos, K. (2019). Detecting Urban Changes with Recurrent Neural Networks from Multitemporal Sentinel-2 Data. ArXiv:1910.07778 [Cs, Eess]. http://arxiv.org/abs/1910.07778.
Paredes, D., & Skidmore, M. (2017). The net benefit of demolishing dilapidated housing: The case of Detroit. *Regional Science and Urban Economics, 66*, 16–27. https://doi.org/10.1016/j.regsciurbeco.2017.05.009.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
PostGIS. (2020). https://postgis.net.
Python Software Foundation. (2020). Python Programming Language 3.6.10 Documentation. https://docs.python.org/3.6/.
Reades, J., De Souza, J., & Hubbard, P. (2019). Understanding urban gentrification through machine learning. *Urban Studies, 56*(5), 922–942. https://doi.org/10.1177/0042098018789054.
Reis, J. P., Silva, E. A., & Pinho, P. (2016). Spatial metrics to study urban patterns in growing and shrinking cities. *Urban Geography, 37*(2), 246–271. https://doi.org/10.1080/02723638.2015.1096118.
Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 67*, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.
Ronacher, A. (2014). Click Documentation (7.x). https://click.palletsprojects.com/en/7.x/.
Sabouri, S., Brewer, S., & Ewing, R. (2020). Exploring the relationship between ride-sourcing services and vehicle ownership, using both inferential and machine learning approaches. *Landscape and Urban Planning, 198*, 103797. https://doi.org/10.1016/j.landurbplan.2020.103797.
Sapena, M., Ruiz, L. A., & Taubenböck, H. (2020). Analyzing Links between Spatio-Temporal Metrics of Built-Up Areas and Socio-Economic Indicators on a Semi-Global Scale. *ISPRS International Journal of Geo-Information, 9*(7), 436. https://doi.org/10.3390/ijgi9070436.
Shafizadeh-Moghadam, H. (2019). Improving spatial accuracy of urban growth simulation models using ensemble forecasting approaches. *Computers, Environment and Urban Systems, 76*, 91–100. https://doi.org/10.1016/j.compenvurbsys.2019.04.005.
Silverman, R. M., Yin, L., Patterson, K. L., & Derudder, B. (2015). Municipal property acquisition patterns in a shrinking city: Evidence for the persistence of an urban growth paradigm in Buffalo, NY. *Cogent Social Sciences, 1*(1). https://doi.org/10.1080/23311886.2015.1012973.
Song, I.-Y., & Zhu, Y. (2016). Big data and data science: What should we teach? *Expert Systems, 33*(4), 364–373. https://doi.org/10.1111/exsy.12130.
Song, Y., Merlin, L., & Rodriguez, D. (2013). Comparing measures of urban land use mix. *Computers, Environment and Urban Systems, 42*, 1–13. https://doi.org/10.1016/j.com.
Soulé, M. E., & Simberloff, D. (1986). What do genetics and ecology tell us about the design of nature reserves? *Biological Conservation, 35*(1), 19–40. https://doi.org/10.1016/0006-3207(86)90025-X.
Steenberg, J. W., Robinson, P. J., & Duinker, P. N. (2019). A spatio-temporal analysis of the relationship between housing renovation, socioeconomic status, and urban forest ecosystems. *Environment and Planning B: Urban Analytics and City Science, 46*(6), 1115–1131. https://doi.org/10.1177/2399808317752927.
Stevenson, J. R., Emrich, C. T., Mitchell, J. T., & Cutter, S. L. (2010). Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal Mississippi following Hurricane Katrina. *Cartography and Geographic Information Science, 37*(1), 57–68.
Talen, E., Wheeler, S. M., & Anselin, L. (2018). The social context of U.S. built landscapes. *Landscape and Urban Planning, 177*, 266–280. https://doi.org/10.1016/j.landurbplan.2018.03.005.
Taubenböck, H., Gerten, C., Rusche, K., Siedentop, S., & Wurm, M. (2019). Patterns of Eastern European urbanisation in the mirror of Western trends – Convergent, unique or hybrid? *Environment and Planning B: Urban Analytics and City Science, 46*(7), 1206–1225. https://doi.org/10.1177/2399808319846902.
The PostgreSQL Global Development Group. (2020). PostgreSQL: The world's most advanced open source database. https://www.postgresql.org/.
Thomas, J. V. (2010). Residential construction trends in America's metropolitan regions. DIANE Publishing. http://books.google.com/books?hl=en&lr=&id=iCbbV0-JtGIC&oi=fnd&pg=PP1&dq=%225+-+Share+by+Unit%22+%22amount+of+permits+issued+by+central+cities+and+core+suburban+communities%22+%22In+fifteen+regions,+the+central+city+more+than+doubled+its+share+of%22+&ots=TMPmbr6Ht8&sig=TRNpgomIs5qDhNXgNYmlOwL_xtM.
Tribby, C. P., Miller, H. J., Brown, B. B., Werner, C. M., & Smith, K. R. (2017). Analyzing walking route choice through built environments using random forests and discrete choice techniques. *Environment and Planning B: Urban Analytics and City Science, 44*(6), 1145–1167. https://doi.org/10.1177/0265813516659286.
Turner, M. G., & Gardner, R. H. (Eds.). (1991). Quantitative Methods in Landscape Ecology: The Analysis and Interpretation of Landscape Heterogeneity. Springer-Verlag. https://www.springer.com/gp/book/9780387942414.
U.S. Census Bureau. (1994). Geographic Areas Reference Manual. U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census. https://www2.census.gov/geo/pdfs/reference/GARM/GARMcont.pdf.

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering, 13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37.

Waddell, P., & Besharati-Zadeh, A. (2020, November 26). A Comparison of Statistical and Machine Learning Algorithms for Predicting Rents in the San Francisco Bay Area. Transportation Research Board 98th Annual Meeting, Washington DC, United States. http://arxiv.org/abs/2011.14924.

Walde, I., Hese, S., Berger, C., & Schmullius, C. (2014). From land cover-graphs to urban structure types. *International Journal of Geographical Information Science, 28*(3), 584–609. https://doi.org/10.1080/13658816.2013.865189.

Waskom, M. (2020). *Seaborn* (Version 0.11.0) [Computer software]. *Zenodo.*. https://doi.org/10.5281/zenodo.4019146.

Weber, R., Doussard, M., Bhatta, S. D., & Mcgrath, D. (2006). Tearing the city down: Understanding demolition activity in gentrifying neighborhoods. *Journal of Urban Affairs, 28*(1), 19–41. https://doi.org/10.1111/j.0735-2166.2006.00257.x.

Yin, L., & Silverman, R. (2015). Housing abandonment and Demolition: Exploring the use of micro-level and multi-year models. *ISPRS International Journal of Geo-Information, 4*(3), 1184–1200. https://doi.org/10.3390/ijgi4031184.

Yoo, S., Im., J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning, 107*(3), 293–306. https://doi.org/10.1016/j.landurbplan.2012.06.009.