Achieving Fairness in Zoning Laws with Machine Learning

William Schimitsch[1] , **Farhad Mohsin**[1]
[1]College of the Holy Cross, Worcester, MA
wfschi25@g.holycross.edu, fmohsin@holycross.edu

Abstract

Zoning is a powerful regulatory tool used to determine the land use and development of a given area. Decisions regarding zone designations, then, have large implications for any given community. It is a problem where local zoning decisions are made by a small-sized zoning board, often through an opaque process. We note two aspects of the task of zoning; first, the end task is to assign a equitable assignment of zoning class (e.g., residential, commercial, mix-use etc.); and second, we have a large amount of data (geographical, demographic, and infrastructural) that is relevant to zoning. Thus we see this as a problem that could benefit from an algorithmic decision-making process that aims to be equitable. In this paper, we first explore zoning classification as a supervised learning task based on existing zoning data in a locality. We extensively collect publicly available data from different sources to train models on them and show how rather complex models are needed to learn to predict the classification accurately. Furthermore, we do a counterfactual analysis based on socioeconomic features to show how they seem to be important indicators for the classification problem in currently available data. We hope that our exploratory paper will lead to more work in applying computational techniques for fair zoning assignments.

1 Introduction

Land use and building regulations have been formally enacted in the United States of America since the early 20th century, when legislators addressed the challenges of increased immigration and industrialization [Chandler and Dale, 2001]. Zoning is the method of urban planning that grants cities and towns the authority to adopt regulations to direct specific use of land, buildings, and structures in order to minimize disruptive or incompatible land use [Chandler and Dale, 2001]. Zoning is usually done by a small board of appointed or elected individuals. The small size of zoning boards, which are typically less than a dozen members, can lead to complications regarding biases, conflicted interests, substitute or alternate members, community impact, and fair decision-making. Today, the country faces new issues, such as a housing and cost of living crisis, which has led to increased skepticism towards zoning across left- and right-leaning groups [Ketcham, 2024; Kahlenberg, 2023; Badger and Bui, 2019]. Furthermore, restrictive zoning bylaws and regulations, such as single-family zoning, can exacerbate racial inequities and stunt economic growth: this kind of “exclusionary zoning” establishes that certain individuals are “not welcome in a community unless [they] can afford a single-family home” [Kahlenberg, 2023; Badger and Bui, 2019; Kahlenberg, 2021].

The process of zoning classification, that is, discerning which parcels should have what land-use designation or ordinance, can be modeled, by definition, by a classification task, that can benefit from machine learning and other computational techniques. The motivation behind exploring zoning classification as a computational task is twofold: first, we aim to better understand the factors that, both directly or indirectly, influence zoning decisions; second, we aim to improve the fairness of existing zoning bylaws by directing our model’s decisions away from prejudice built into the data as well as those biases in existing infrastructure and land use regulation.

With the focus on assessing fairness of existing zoning classifications from a socioeconomic perspectives, we compiled a dataset for Worcester County, Massachusetts, in order to better understand the relation between the characteristics of a parcel of land and its zoning classification[1]. This dataset then enabled us to evaluate zoning processes through the lens a machine learning model for zoning classification and explore feature importance for the different features. Our work aggregated comprehensive datasets and identified underlying patterns among given zoning area, ultimately with the hope of developing fairer and more accurate zoning policies than current zoning legislative boards and ordinances.

Our Contributions

- We collect data from a multitude of sources following the framework prepared by [Lawrimore *et al.*, 2024] for the Worcester county in Massachusetts.
- We show that a gradient boosted tree model [Fried-

[1]Worcester was chosen as an example mostly due to the authors’ proximity to the region.

man, 2001], specifically an XGBoost model [Chen *et al.*, 2022] achieves rather high accuracy in predicting zoning on a test set. Our counterfactual analysis on the models show how socioeconomic features of the population turn out to be an important feature in predicting current zoning classifications.

- One of our major aims with this paper is to bring this important problem in notice of more researchers who work on computational methods on fair assignment problems, as we think the way to work towards a more equitable zoning classification is through computational.

2 Relevant Work

2.1 Zoning

Municipalities divide districts and neighborhoods based on a "comprehensive plan," which might prioritize, say, economic development or resource protection through certain land use restrictions. In this way, "zoning depends on planning and planning depends on zoning" [Chandler and Dale, 2001]. Considering fairness in zoning bylaws, it is important to recognize *who* actually *does* zoning: those responsible for zoning process are the members of a municipality's planning and zoning boards. These boards are groups of appointed or elected individuals that develop, recommend, and approve zoning ordinances. While the former implement and adopt a comprehensive plan for a given municipality, from which all subsequent zoning bylaws, decisions, and changes are derived, the latter consider applications for special permits or exceptions to the regulations and requirements specified by a given zoning ordinance. These groups can hold legislative, advisory, administrative, and judicial power regarding zoning [Chandler and Dale, 2001].

The small size of zoning boards, which are typically less than a dozen members, can lead to complications regarding biases, conflicted interests, substitute or alternate members, community impact, and fair decision-making. It is natural for board members to possess occasional conflicts of interest regarding certain zoning approvals or appeals: for example, board members tend to be local to the area they govern, which means that a zoning ordinance regarding a zone that encapsulates their place of residence or is nearby would be a clear conflict of interest. Such an occasion necessitates the use of alternate board members so that the size of the board does not drastically shrink for a given matter [Salkin, 2009]. Of course, this raises further issues, such as naming alternate board members, compensation, terms, etc [Salkin, 2009]. Another concern relates precisely to the size and composition of these commissions, which are often skewed toward 'white collar' jobs and misrepresent the wishes of the average citizen [Anderson *et al.*, 2008]. As a result, comprehensive zoning plans can prioritize economic growth over community impact and public good [Anderson *et al.*, 2008]. Other concerns regarding zoning include environmental sustainability and discouraged economic or societal development. The discontentment and disfunctionality of current zoning processes thus call for an updated, holistic, and modern approach to land-use regulation.

2.2 Machine Learning and Zoning

Widespread adoption of ML in real-world decision-making environments has led to increased concerns among scholars regarding fairness, bias, and prejudice. Indeed, studies have shown in multitudes of work how various issues can cause bias in ML models [Buolamwini and Gebru, 2018; Julia Angwin and Kirchner, 2016; Mehrabi *et al.*, 2021; Pessach and Shmueli, 2022; Chouldechova and Roth, 2020] Therefore, although a ML model might be able to understand a dataset well and accurately predict the output features, the model inherits the existing bias in the dataset.

Previously, Lawrimore *et al.* has used machine learning to predict zoning classification where zoning information may not be publicly available. However, we think that predicting based on existing zoning labels is subject to the bias inherent in the current zoning labels. Since zoning ordinances have direct social, political, and cultural effects, it is crucial to understand the fairness of the ML models at play here. Thus, zoning is a kind of "socio-technical system"—not only are the potential biases of ML models here important, but so too are the ways in which the model shapes the environment and human actors that dictate certain zoning ordinances [Chouldechova and Roth, 2020]. For example, some features associated with zoning are social vulnerability indexes(SVI) of the population. We might say that, with all geographical, infrastructural features being equal, just changing the SVI features should not lead to different classification, and we might be interested in fairness with respect to these features. In this work, we follow the framework proposed by Lawrimore *et al.* to collect various data regarding zoning, but with the separate goal of analyzing fairness.

2.3 Counterfactual analysis

Counterfactuals are something well-understood in computational social choice, as the notion of counterfactual scenarios have been used in auction design, and fair division methods. In ML, counterfactual analysis can be applied to explain feature importances. The term "counterfactual" denotes something that has not happened or is not the case. Therefore, counterfactual analysis dictates that the existing, real data are altered in some fashion, such that certain changes in the input data correlate to changes in the model's predictions [Dandl and Molnar, 2025]. In this way, the relationship between the data and the model's predictions can be better understood: it is possible to calculate the minimum or sufficient change to a certain feature or group of features that leads to a change in the model's prediction. In this work, we focus particularly on counterfactuals regarding to various socioeconomic features.

3 Dataset Collection

3.1 Features

Although data regarding the zoning bylaws of land parcels are widely available in the United States, there exists no centralized dataset containing information about these parcels that might have influenced the bylaw in place [Atlas, 2023]. Therefore, in order to compile a dataset for Worcester County, Massachusetts, we collected data from various national- and state-level sources. The resulting dataset contains 34 input features (independent variables), each providing information regarding the characteristics of a given parcel of land, and an output feature (dependent variable), which contains the classification of zoning ordinance for that parcel. We followed the framework developed by Lawrimore *et al.*, who pose a group of 39 predictor variables across four larger categories for their zoning classification task: population, community characteristics, built environment, and natural environment [Lawrimore *et al.*, 2024]. We adapted these features to our county-level task. Furthermore, we included other variables, such as public utility status and additional social vulnerability indicators, in order to better understand how certain levels of infrastructure and social characteristics affect zoning. Table 1 list all of the features we collect and use in our work.

These features are themselves not free from bias: data on infrastructure, such as information regarding roads, fire stations, schools, etc., are not completely independent of a parcel's zoning classification. In fact, these two can be interdependent and biased. That is, it is not certain whether municipalities adopt certain zoning ordinances because of certain existing built characteristics of the land, or if those characteristics were the result and reinforcement of an initial zoning ordinance. Therefore, our ML model should not necessarily preclude that a given parcel is, say residential, simply because there is no school nearby. Indeed, that there is no school in the area might be a point of discrimination within that municipality.

3.2 Data collection

The data for the 34 input features was available across approximately 20 different sources. Gathering and compiling this data, even just for one Massachusetts county, required time and resources to manage the complex relationship between different features. Specifically, sources provide data at varying scopes: data was available for parcels, census blocks, towns, county regions, counties, states, or the nation depending on the source. Due to these different scopes and spatial geometries, we leveraged geospatial tools to aggregate data from different sources. You can see more technical details in Appendix B.

Next, we collected population data from the 2020 U.S. Census Bureau survey for census blocks in Massachusetts to consider the population and population density of different areas [Bureau, 2022]. We also included the areas and lengths of individual parcels to factor in the size constraints of zoning ordinances [of Geographic Information and of Coastal Management, 2019].

Both community services and demographic characteristics are important for understanding the needs of a given neighborhood. We gathered geo-spatial data from the Massachusetts Bureau of Geographic Information (MassGIS), in conjunction with other state departments, for railroads, hospitals, police stations, fire stations, schools (pre-kindergarten through high school), colleges, and parks to calculate the distance of each parcel to the nearest respective facility [Staff, 2023; Massachusetts Department of Public Health, 2024b; Massachusetts Department of Public Health, 2024a; Program, 2022; Agency, 2022; of Elementary Secondary Education School, 2024; Massachusetts Bureau of Geographic Information, 2024; Institute, 2025]. For all distance calculations, we retrieved the minimum distance to a given geometry (i.e., park, rail line, hospital, etc.) within a range of 10 kilometers.

Because zoning can be influenced by political or prejudiced beliefs, it is important to account for a community's characteristics, particularly its level of social vulnerability. For this, we used the CDC/ATSDR Social Vulnerability Index, which incorporates 16 U.S. Census variables from the 5-year American Community Survey [for Disease Control *et al.*, 2022], to compile data on socioeconomic status, household composition, racial and ethnic minority status, housing type, and transportation for regions within Worcester County, enabling analysis of how zoning may or may not consider social vulnerability.

We used geo-spatial data from the U.S. Census Bureau and MassGIS, along with state utility and environmental agencies, to compute distances and binary features related to roads, buildings, and public infrastructure, which we believe capture the built environment's impact on an area [Bureau, 2024; of Geographic Information, 2024; Massachusetts Department of Public Utilities and Cable, 2021; of Environmental Protection WQTS database, 2024].

To analyze similarities between parcels, we collected data on physical land characteristics—such as slope, water proximity, land cover, and agricultural use—since these natural features are less influenced by human decisions than community or built environment traits. Using data from the USGS's 3D Elevation Program, National Hydrography Dataset, and Protected Areas Database [Survey and of Geographic Information, 2023; Survey, 2023; Survey, 2024], as well as data from MassGIS, NOAA's Office of Coastal Management, the U.S. Division of Agriculture, and the U.S. Census Bureau [of Geographic Information and of Coastal Management, 2019; of Agriculture, 2023; Bureau, 2022], we calculated metrics like average slope, distances to water and cropland, presence of protected lands, and land/water area per census block.

4 Classification Tasks and Analysis

Zoning ordinances are available in Massachusetts in a centralized dataset compiled by MassGIS [of Geographic Information and of Coastal Management, 2019]. Zoning is a local regulatory process respective to each municipality, which means there are 1,600+ unique land use classifications across the state of Massachusetts. Therefore, MassGIS generalized these into 16 encompassing classes: Unknown, Open land, Commercial, Industrial, Forest, Agriculture, Recreation, Tax exempt, Mixed use (primarily residential), Residential (single family), Residential (multi-family), Residential (other), Mixed use (other), Mixed use (primarily commercial), Right-of-way, and Water [of Geographic Information and of Coastal Management, 2019]. So, we train models for this multi-class classification task, with a one-versus-rest strategy. We maintain a 80-20 stratified train-test splitting strategy to create our test set.

4.1 Models

Gradient boosted models [Friedman, 2001] has in recent times boasted great performance in tabular data-based ML

Table 1: Input features for XGBoost zoning classification

Population	**Community Characteristics**	**Built Environment**	**Natural Environment**
Population Total	Distance to Railroad	Distance to Road	Slope
Population Density	Distance to Hospital	Distance to Building	Distance to Lake
Parcel Area	Distance to Police Station	Distance to Water Supply	Distance to River
Parcel Length	Distance to Fire Station	Natural Gas Provider (y/n)	Distance to Miscellaneous Waters
	Distance to School	Cable TV Provider (y/n)	Protected Land (y/n)
	Distance to College		Distance to Protected Land
	Park (y/n)		Land Cover
	Distance to Park		Crop (y/n)
	Housing Units		Distance to Cropland
	Social Vulnerability Index		Land Area
	Socioeconomic Status		Water Area
	Household Characteristics		
	Racial & Ethnic Minority Status		
	Housing Type & Transportation		
Data Sources			
U.S. Census Bureau (2020); MassGIS	CTPS; Mass. DESE; NCES; MEMA GIS Program; Mass. DPH; OEMH; CHIA; DMH; Esri; ATSDR	U.S. Census Bureau (2020); MassGIS; Mass. DEP; Mass. DPU; Mass. DTC	USGS 3DEP; USGS NHD; USGS PAD-US; MassGIS; NOAA; OCM; USDA

tasks. This is largely due to the availability of optimized libraries such as XGBoost (Extreme Gradient Boosting) [Chen *et al.*, 2022]. Lawrimore *et al.* used Random Forests in their work in predicting zoning for unknown parcels, we saw XGBoost outperformed Random Forests in our test set. We also want to stress that our goal in this work was not to get the best possible model, but rather a reasonably well-performing model so that we can analyze the feature importance for various features. We present comparative result between linear models and XGBoost in Table 2, which indicates that the prediction task is sufficiently complex that simple models like linear models are not as successful as XGBoost. We just report the macro-level metrics in Table 2. In Table 3, you will see how the XGBoost model performs for different classes along with the distribution of labels in the test set (which is similar in ratio to the distribution in the full dataset).

Table 2: Performance comparison of classifiers

Classifier	Accuracy	Precision	Recall	F1 Score
Linear SVM	0.37	0.29	0.37	0.33
Logistic Regression	0.34	0.29	0.34	0.31
XGBoost	**0.74**	**0.75**	**0.71**	**0.73**

4.2 Counterfactual testing

In Table 4, we report the results of our counterfactual experiments. We first generate counterfactual data instances, where we change specific features. For this purpose, we first collect the features in super-features consisting of multiple features, as explained below:

- SVI: Socioeconomic status, household characteristics, minority status, housing type and transportation, social vulnerability index.
- Geography: Land area, water area, average slope.
- Utility: Water supply, gas availability, cable availability.
- School: Distances to nearest school and college.
- Safety: Distances to police, fire station, and hospital.

For each experiment, we modify the test dataset for the features in a specific category. Care is taken so that only realistic values are generated. Then, we calculate the counterfactual accuracy for the generated counterfactual dataset. A low counterfactual accuracy indicates high importance of a group of feature in the prediction. As seen in Table 4, the lowest F1-score (which sometimes goes to 0 for some categories due to 0 recall) is due to change in the SVI indexes in many cases. In fact, you can see how this specially affects the Residential classes. This reinforces our idea that existing social vulnerabilities are further perpetuated by current zoning practices. And if the current practices are considered as basis for future zoning decisions, these bias may persist.

5 Discussion and Future Work

In this work, we aim to introduce the problem of zoning ordinances as a problem of interest to the fair division and assignment research community. We adopt the framework that Lawrimore *et al.* uses for predictive analysis to instead analyze fairness concerns in terms of socioeconomic

Table 3: XGBoost classification performance metrics by zoning class

Zoning Class	**Precision**	**Recall**	**F1-score**	**Occurrences**
Open land	0.73	0.61	0.67	93,465
Commercial	0.74	0.68	0.71	20,819
Industrial	0.75	0.72	0.74	12,347
Forest	0.79	0.71	0.75	9,228
Agriculture	0.78	0.71	0.75	16,007
Recreation	0.77	0.69	0.73	5,057
Tax exempt	0.82	0.81	0.82	57,455
Mixed use, primarily residential	0.76	0.66	0.71	12,694
Residential - single family	0.77	0.82	0.80	231,863
Residential - multi-family	0.67	0.55	0.61	51,403
Residential - other	0.78	0.70	0.74	3,559
Mixed use, other	0.75	0.69	0.72	3,204
Mixed use, primarily commercial	0.76	0.68	0.72	751
Right-of-way	0.76	0.85	0.81	159,000
Water	0.83	0.79	0.81	8,864
Accuracy			0.76	685,716
Macro average	0.77	0.71	0.74	685,716
Weighted average	0.76	0.76	0.76	685,716

Table 4: Counterfactual accuracy for different counterfactuals

	Original	SVI	Geography	Utility	School	Safety
Open land	0.67	**0.49**	0.56	0.61	0.53	0.51
Commercial	0.71	**0.2**	0.54	0.56	0.35	0.21
Industrial	0.74	**0.13**	0.47	0.33	0.36	0.29
Forest	0.75	0.36	**0.3**	0.66	0.46	0.43
Agriculture	0.75	**0.43**	0.5	0.63	0.58	0.44
Recreation	0.73	**0.25**	0.41	0.53	0.49	0.32
Tax exempt	0.82	**0.63**	0.75	0.78	0.68	0.68
Mixed use, primarily residential	0.71	**0**	0.35	0.57	0.39	0.31
Residential - single family	0.8	**0.72**	0.74	0.77	0.74	0.73
Residential - multi-family	0.61	0.36	0.41	0.54	0.39	**0.28**
Residential - other	0.74	**0.06**	0.5	0.6	0.33	0.26
Mixed use, other	0.72	**0**	0.39	0.37	0.44	0.22
Mixed use, primarily commercial	0.72	**0.12**	0.47	0.63	0.2	0.31
Right-of-way	0.81	**0.73**	0.75	0.79	0.77	0.75
Water	0.81	**0.38**	0.56	0.78	0.73	0.74
macro	0.76	**0.59**	0.66	0.71	0.65	0.62

indexes. We show from our counterfactual explanations, how social vulnerability related features are important for the current zoning classifications. For future work, we want t Future work could focus on developing an explicitly fair zoning algorithm. We can consider modifying current predictive classification systems to incorporate fairness constraints, using techniques such as envy-free classification [Balcan *et al.*, 2019]. However, with a dataset with high label imbalance, where single-family residences are highly prevalent, which restricts housing diversity and access, it is also worth designing a zoning framework from scratch that addresses equity directly in the assignment process. Additionally, geographic constraints—such as proximity to protected land or infrastructure limitations—must be treated carefully to distinguish valid environmental concerns from those that may inadvertently reinforce exclusionary zoning practices. Ultimately, we hope this work motivates further interdisciplinary discussion and innovation at the intersection of land use policy, machine learning, and algorithmic fairness.

References

[Agency, 2022] Massachusetts Emergency Management Agency. Fire stations. https://www.mass.gov/info-details/massgis-data-fire-stations, 2022.

[Anderson *et al.*, 2008] Jerry L. Anderson, Aaron E. Brees, and Emily C. Reninger. A study of american zoning board composition and public attitudes toward zoning issues. *The Urban Lawyer*, 40(4):689–745, 2008.

[Atlas, 2023] Land Use Atlas, 2023.

[Badger and Bui, 2019] Emily Badger and Quoctrung Bui. Cities start to question an american ideal: A house with a yard on every lot. *The New York Times*, 2019.

[Balcan *et al.*, 2019] Maria-Florina F Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-free classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 02 2018.

[Bureau, 2022] U.S. Census Bureau. 2020 u.s. census data release for massachusetts. https://www.mass.gov/info-details/massgis-data-2020-us-census, 2022.

[Bureau, 2024] U.S. Census Bureau. Tiger/line shapefiles. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html, 2024.

[Chandler and Dale, 2001] Michael Chandler and Gregory Dale. *Zoning basics*. Champlain Planning Press, 2001.

[Chen *et al.*, 2022] Tianqi Chen, Michael Benesty, Yuan Tang, Nan Zhu, Jiaming Yuan, Hyunsu Cho, Rory Mitchell, Hongliang Liu, Tong He, Vadim Khotilovich, Bing Xu, Sergei Lebedev, Scott Lundberg, and Egor Smirnov. *XGBoost Documentation*. Distributed (Deep) Machine Learning Community, 2022. Version 3.0.0.

[Chouldechova and Roth, 2020] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020.

[Dandl and Molnar, 2025] Susanne Dandl and Christoph Molnar. *Interpretable Machine Learning*, chapter 15 Counterfactual Explanations. 2025.

[for Disease Control *et al.*, 2022] Centers for Disease Control, Prevention/Agency for Toxic Substances, Disease Registry/Geospatial Research, Analysis,, and Services Program. Cdc/atsdr social vulnerability index. https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html?CDC_AAref_Val=https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html, 2022.

[Friedman, 2001] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[Institute, 2025] Environmental Systems Research Institute. U.s. recreational areas. https://www.arcgis.com/home/item.html?id=4f14ea9215d1498eb022317458437d19, 2025.

[Julia Angwin and Kirchner, 2016] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[Kahlenberg, 2021] Richard D. Kahlenberg. The 'new redlining' is deciding who lives in your neighborhood. *The New York Times*, 2021.

[Kahlenberg, 2023] Richard D. Kahlenberg. Only Zoning Reform Can Solve America's Housing Crisis. *The Wall Street Journal*, 2023.

[Ketcham, 2024] John Ketcham. Washington can encourage pro-growth zoning. *The Wall Street Journal*, 2024.

[Lawrimore *et al.*, 2024] Margaret A. Lawrimore, Georgina M. Sanchez, Cayla Cothron, Mirela G. Tulbure, Todd K. BenDor, and Ross K. Meentemeyer. Creating spatially complete zoning maps using machine learning. *Computers, Environment and Urban Systems*, 112:102157, 2024.

[Massachusetts Bureau of Geographic Information, 2024] National Center for Education Statistics Massachusetts Bureau of Geographic Information. Colleges and universities. https://www.mass.gov/info-details/massgis-data-colleges-and-universities, 2024.

[Massachusetts Department of Public Health, 2024a] Bureau of Hospitals Massachusetts Department of Public Health, Department of Mental Health. Non-acute care hospitals. https://www.mass.gov/info-details/massgis-data-non-acute-care-hospitals, 2024.

[Massachusetts Department of Public Health, 2024b] Center for Health Information Analysis Massachusetts Department of Public Health, Office of Emergency Medical Services. Acute care hospitals. https://www.mass.gov/info-details/massgis-data-acute-care-hospitals, 2024.

[Massachusetts Department of Public Utilities and Cable, 2021] Massachusetts Department of Telecommunications Massachusetts Department of Public Utilities and Cable. Public utility service providers. https://www.mass.gov/info-details/massgis-data-public-utility-service-providers, 2021.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021.

[of Agriculture, 2023] U.S. Division of Agriculture. National agricultural statistics service cropland data layer. https://nassgeodata.gmu.edu/CropScape/, 2023.

[of Elementary Secondary Education School, 2024] Massachusetts Department of Elementary Secondary Education School. Massachusetts schools (pre-k through high school). https://www.mass.gov/info-details/massgis-data-massachusetts-schools-pre-k-through-high-school, 2024.

[of Environmental Protection WQTS database, 2024] Massachusetts Department of Environmental Protection WQTS database. Public water supplies. https://www.mass.gov/info-details/massgis-data-public-water-supplies, 2024.

[of Geographic Information and of Coastal Management, 2019] Massachusetts Bureau of Geographic Information and National Oceanic Atmospheric Administration Office of Coastal Management. Massachusetts 2016 land cover/land use. https://www.mass.gov/info-details/massgis-data-2016-land-coverland-use, 2019.

[of Geographic Information, 2024] Massachusetts Bureau of Geographic Information. Building structures (2-d). https://www.mass.gov/info-details/massgis-data-building-structures-2-d, 2024.

[Pessach and Shmueli, 2022] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), February 2022.

[Program, 2022] Massachusetts Emergency Management Agency GIS Program. Police stations. https://www.mass.gov/info-details/massgis-data-police-stations, 2022.

[Salkin, 2009] Patricia E. Salkin. Providing for alternate members on planning and zoning boards: Drafting effective local laws. *Planning & Environmental Law*, 61(8):3–7, 08 2009. Copyright - Copyright American Planning Association Aug 2009; Last updated - 2024-10-07; SubjectsTermNotLitGenreText - Social.

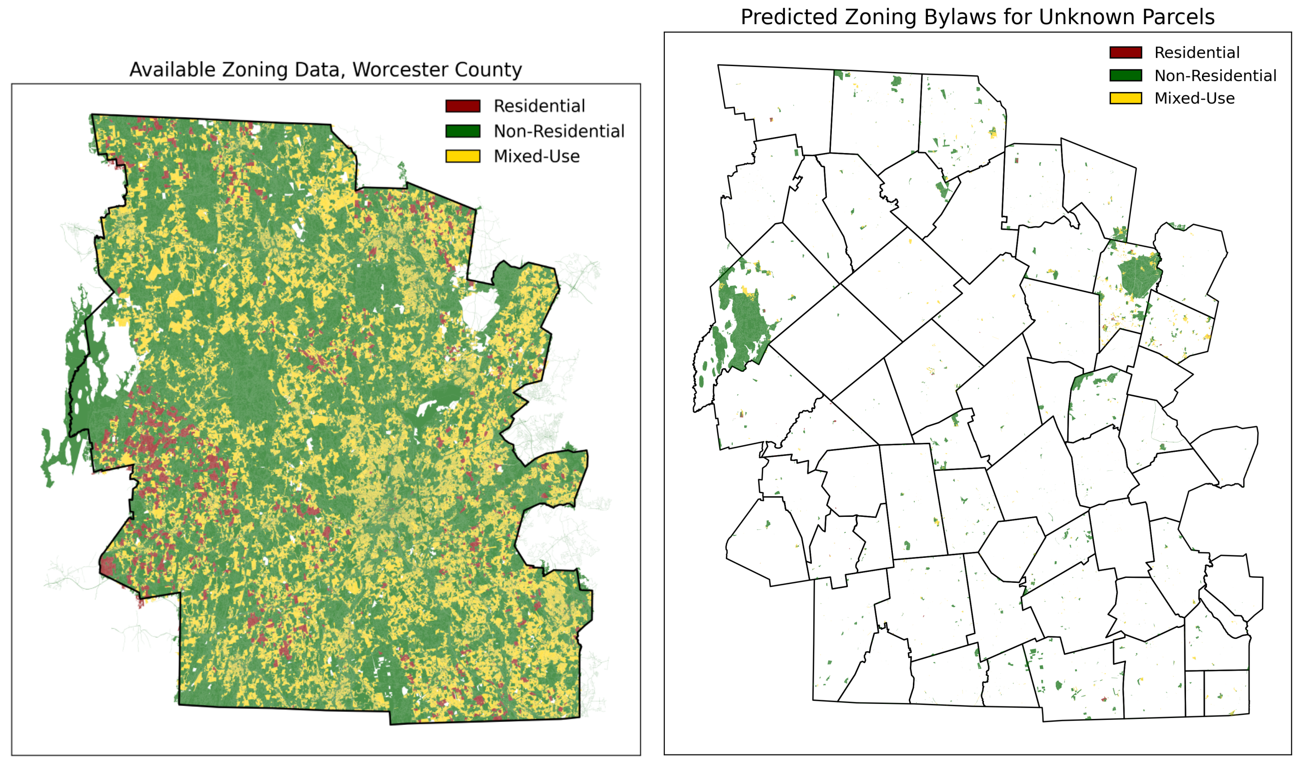
[Staff, 2023] Central Transportation Planning Staff. Amtrak and mbta commuter rail trains. https://www.mass.gov/info-details/massgis-data-trains, 2023.

[Survey and of Geographic Information, 2023] U.S. Geological Survey and Massachusetts Bureau of Geographic Information. Lidar terrain data - slope. https://www.mass.gov/info-details/massgis-data-lidar-terrain-data, 2023.

[Survey, 2023] U.S. Geological Survey. National hydrography dataset. https://www.usgs.gov/national-hydrography/national-hydrography-dataset, 2023.

[Survey, 2024] U.S. Geological Survey. Protected areas database of the united states (pad-us) 4. https://www.usgs.gov/data/protected-areas-database-united-states-pad-us-4, 2024.

A Predicting Zoning for Unknown Parcels



(a) Available data on zoning ordinances for municipalities in Worcester County, Massachusetts

(b) Parcels without available zoning ordinances for municipalities in Worcester County, Massachusetts

Figure 1: Zoning ordinance data coverage in Worcester County, Massachusetts

The main goal in [Lawrimore *et al.*, 2024] was to predict zoning ordinance for unknown parcels. As part of our work, we followed their framework and did the same for unknown parcels in the Worcester county. The result can be shown in Figure 1.

B Details on Data Collection

The data for the 34 input features was available across approximately 20 different sources. Gathering and compiling this data, even just for one Massachusetts county, required time and resources to manage the complex relationship between different features. Specifically, sources provide data at varying scopes: data was available for parcels, census blocks, towns, county regions, counties, states, or the nation depending on the source. Due to these different scopes and spatial geometries, we leveraged GeoPandas, an open-source Python library for managing geo-spatial data, and its "spatial join" functionality in order to merge divergent data. With this, we could read geo-spatial data (from shapefiles, geodatabases, GeoPackages, etc.) into GeoPandas' GeoDataFrame object and subsequently merge data and calculate distance or binary features (i.e, "Distance to Park" and "Park (y/n)"). Some data, such as cropland and elevation data, were only available as raster data, so we first processed them into a set of geometries, based on the pixel values, and then compiled these geometries into a GeoDataFrame.

One of the primary questions that zoning regulations attempt to answer is "how to affect what is wanted and avoid what is not" [Chandler and Dale, 2001]. Urban development and housing needs are therefore at the forefront of zoning concerns because certain infrastructure is required to support certain neighborhoods and some land should not be inhabited because they are incompatible (i.e., industrial areas) or negatively impact environmental resources. Therefore, we collected population data from the 2020 U.S. Census Bureau survey for census blocks in Massachusetts to consider the population and population density of different areas [Bureau, 2022]. We also included the areas and lengths of individual parcels to factor in the size constraints of zoning ordinances [of Geographic Information and of Coastal Management, 2019].

Both community services and demographic characteristics are important for understanding the needs of a given neighborhood. We gathered geo-spatial data from the Massachusetts Bureau of Geographic Information (MassGIS), in conjunction with other state departments, for railroads, hospitals, police stations, fire stations, schools (pre-kindergarten through high school), colleges, and parks to calculate the distance of each parcel to the nearest respective facility [Staff, 2023; Massachusetts Department of Public Health, 2024b; Massachusetts Department of Public Health, 2024a; Program, 2022;

Agency, 2022; of Elementary Secondary Education School, 2024; Massachusetts Bureau of Geographic Information, 2024; Institute, 2025]. For all distance calculations, we retrieved the minimum distance to a given geometry (i.e., park, rail line, hospital, etc.) within a range of 10 kilometers. If there existed no such geometry for a parcel, it was assigned a maximum distance, just over 10 km, to indicate there was no nearby geometry. In addition, we coded a binary park feature, denoting whether or not a given parcel contains a park or area of recreation. Together, we aim to represent the kind of social area a parcel of land is and understand whether it possesses characteristics typical of either rural, suburban, or urban areas.

The process of zoning can be an exclusionary process that is infused with political or prejudiced beliefs calls for additional measurements of a community's characteristics, particularly index of social vulnerability. The CDC/ATSDR social vulnerability index helps public services ”identify communities that may need support before, during, or after disasters” [for Disease Control *et al.*, 2022]. Just as certain community characteristics, like housing density, can impact zoning decisions, so can certain disadvantaged groups be impacted by zoning. Therefore, we compiled social vulnerability data available for different regions of Worcester County, including social vulnerability index, socioeconomic status, household characteristics, racial and ethnic minority status, and housing type and transportation. The CDC/ATSDR calculated these summary variables using 16 U.S. Census variables from the 5-year American Community Survey [for Disease Control *et al.*, 2022]. Along with these variables, we pulled the number of housing units per census block as a measure of housing density [for Disease Control *et al.*, 2022]. This allows us to analyze how zoning decisions do or do not factor in measures of social vulnerability.

Transportation and building infrastructure have major implications for zoning bylaws: these are necessary to support the daily lives of citizens as well as build communities. From the U.S. Census Bureau, we retrieved geo-spatial information regarding major roads [Bureau, 2024]. MassGIS provides data on building structures and, in conjuction with the state Department of Public Utilities, Department of Telecommunications and Cable, and Department of Environmental Protection, information regarding public utilities and water supply [of Geographic Information, 2024; Massachusetts Department of Public Utilities and Cable, 2021; of Environmental Protection WQTS database, 2024]. These data allowed us to calculate nearest distances, in the same manner as those features in the community characteristics categories, to resources and buildings as well as code binary features regarding public utility infrastructure. Such features, we believe, are able to represent the impact of the built environment of a given area.

One way to understand the similarities between parcels is through the analysis of the land's physical characteristics—natural environment is generally less affected by human decision-making than community characteristics and built environment. That is, we can compare parcels with similar geography and hydrography to understand what *other* features (i.e., population, community characteristics, and built environment features) cause similar natural environments to be regulated by zoning processes differently (or not). These geographical and hydrological features include: average slope, or variation in elevation, provided by the U.S. Geological Survey's (USGS) 3D Elevation Program and MassGIS, distances to lakes, river, and miscellaneous waters, provided by the USGS's National Hydrography Dataset, and land cover to assess urban growth, inventory wetlands, provided by MassGIS and the National Oceanic and Atmospheric Administration's Office of Coastal Management, ”to assess coastal intertidal areas, and adjacent uplands, and delineate wildlife habitat” [Survey and of Geographic Information, 2023; Survey, 2023; of Geographic Information and of Coastal Management, 2019]. We calculated the average slope of a parcel by converting the raster pixel data (supplied in a JPEG 2000 file) to a GeoPandas GeoDataFrame and calculated the hydrographical distances in the same manner as previously mentioned. In addition, we pulled data for protected areas from USGS's Protected Areas Database of the United States (4.0) in order to calculate a binary protected land feature and the nearest distance to protected land for each parcel [Survey, 2024]. By definition, these areas are free from zoning regulations because there are already regulations on how the land should be conserved and treated. Therefore, it is crucial to know where exactly these lands lie geo-spatially. Agriculture is also an important land characteristic to factor in, as municipalities often want to protect these resources. We pulled cropland data layers from the U.S. Division of Agriculture, which we then masked from raster data to a GeoDataFrame to discern whether a parcel contained cropland as well as how far it lies from cropland [of Agriculture, 2023]. Finally, we considered the total land and water area contained in each census block from the U.s. Census Bureau for another measure of a region's geographical characteristics [Bureau, 2022].