Essays on Policy Evaluation: From Experiments to Machine Learning

Brendan Shanks

University of Munich

Essays on Policy Evaluation: From Experiments to Machine Learning

INAUGURAL-DISSERTATION

zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2021

vorgelegt von

Brendan Shanks

Referent:	Prof. Dr. Joachim Winter
Korreferent:	Prof. Davide Cantoni, PhD
Promotionsabschlussberatung:	14. Juli 2021

Datum der mündlichen Prüfung: 05.07.2021
Berichterstatter: Joachim Winter, Davide Cantoni, Stephan Heblich

To Marie

Acknowledgments

Though an often confusing, solitary endeavour, writing this dissertation has benefited from the interaction with and support of fantastic colleagues and friends at the university. Whether I needed coding aid, an extra pair of eyes to read over a draft, or just a break from the toils of the graduate lifestyle, I was always able to find someone who was willing to lend a hand.

I first and foremost want to thank Joachim for his supervision over the years. He was always available for feedback and advice whenever I needed it. And most importantly to me he gave me the space to explore the topics that most interested me: including the ones that did not work out. As a young, wide-eyed masters student I started out wanting to study macroeconomics. To think that were it not for taking his microeconometrics course I may have become a macroeconomist!

I also want to thank Davide, not only for agreeing to be my second supervisor, but for the feedback over the years at empirical seminars and for all the hard work he has done in preparing us for the job market. This dissertation, Chapter 1 in particular, has benefited immensely from his input.

A special thanks goes out to Stephan who hosted me in Toronto during my visit there and for agreeing to be a part of my dissertation committee. It was a fantastic experience to meet and be in the presence of such a high density of scholars interested in urban economics.

I am grateful for my co-authors, Kristina, Michael, Timm, and Vojta, who broadened my research interests and constantly challenged me to improve. The journey would have been much less rich without the repeated interactions and fruitful collaboration with them.

The ride was made all the more enjoyable because of the fantastic GRK cohort over in the *Rückgebäude* during the first year: Hung-Ni, Lukas, Corinna, Franziska, Michael, Valerii, and Felix. Whenever I felt lost or disoriented they would always bring me back and point me in the right direction. The Mensa lunch breaks were often a welcome distraction.

My colleagues at the Chair of Empirical Economic Research, Christoph, Corinna, Fabian, Nadine, Pavel, and Tobias, made the transition across Ludwigstraße painless and enjoyable. It was great to have such a tight-knit group to support me through the years.

This dissertation would not have been half as good without the sustained support I received from Ines, Julia, and Manu. They not only saved me from many headaches but were always there if I just needed to vent.

I would not have made it this far without the support of my parents. The journey was often a squiggly one but they made sure it never went backwards.

And finally Marie. I cannot express how thankful I am for your unwavering support and encouragement over the years.

Brendan Shanks
Munich
March 2021

Contents

List of Figures

List of Tables

Preface

He mentions the phrase "identification problem,"
which, though no one knows quite what he means, is
said with such authority that it is totally convincing.
— Edward E. Leamer (1983)
"Let's Take the Con Out of Econometrics"

The importance of econometrics to economics is hard to overstate. In fact, the first recipients of the Nobel Memorial Prize in Economic Sciences were two founders of the field—Ragnar Frisch and Jan Tinbergen—"for having developed and applied dynamic models for the analysis of economic processes." It's what separates theory from evidence; presumption from fact. It allows economists to test hypotheses, identify causal effects, measure structural parameters.

As a field, econometrics, and applied microeconomics along with it, has changed remarkably since the days of Frisch and Tinbergen. It has also faced its fair share of critiques and moments of existential crisis. Leamer (1983), in his oft-referenced commentary on the field of econometrics, disparaged the "sad and decidedly unscientific state of affairs we find ourselves in." His primary criticism was the false sense of objectivity many practitioners projected: that there was a "truth" to be found in the functional form, set of variables, or model specification. Econometricians often ignored or were ignorant of the sometimes subconscious influence of their own assumptions and priors. He suggested the need for sensitivity analyses, what applied economists now usually refer to as robustness checks, to test whether different modelling assumptions change the inferences being made. The set of different assumptions would stem in part from the researcher, and from "anticipat[ing] the opinions of his consuming public."

Five years after Leamer levied his critique LaLonde (1986) followed with his influential remarks. He was troubled by the realization that the standard econometric techniques used to estimate the impact of an employment program with observational data were unable to replicate experimental results. He concluded that results derived from econometric methods should be compared to experimental ones to verify their validity. Given the infeasibility of applying experiments to all questions of economic interest—on account of the potential costs, ethical concerns, and practical limitations—this left applied econometricians questioning the potential contributions of their methods.

Fortunately for the field, and for economics in general, econometrics has made great strides since the critiques of Leamer and LaLonde. More thought has been turned to the quality of the research method being used and the assumptions specific econometric models impose. A standard toolkit of methods is now well-established in applied microeconomics: difference-in-differences, instrumental variables, regression discontinuity design, and experimental methods, for example. This "credibility revolution," as documented by Angrist and Pischke (2010), has re-focused researchers' questions towards issues of design: do I have exogenous variation in the treatment of interest? Can I disentangle what channels the effect works through? Are there institutional quirks that may allow me to identify the impacts of certain policies? Only after the design has been settled on does the researcher choose the econometric method best suited to measure the effect she is after.

As LaLonde suggested, experiments, either in the lab or field, where the source of randomization is controlled by the researcher, have been a key part of this revolution. But arguably the biggest shift has come from practitioners seeking out quasi-experimental sources of variation: differences in some variable of interest on account of geography, history, politics, policy, or some combination thereof. This has, both out of necessity and curiosity, expanded the set of questions and fields in which economists apply their methods. To paraphrase John Maynard Keynes, the applied microeconomist must be legal scholar, sociologist, psychologist, historian—in some degree.

These trends have now been coupled with an increase in the number of sources and size of data, and with improvements in computing power. This has created new opportunities for economists to address difficult to answer economic questions, both old and new. At the same time, the tools required to fully utilize these advances has evolved. Machine learning techniques are now gaining application in economics. They complement the econometric toolkit by aiding in prediction problems (measuring economic growth through satellite data) and identifying hidden patterns and clusters (defining product and customer market segments). The approach taken by these methods, as explained by Mullainathan and Spiess (2017), is inductive: given the large amount of data at hand,

what rules best explain their distribution and pattern?

But machine learning need not be separate from causal inference. Indeed, ML techniques can be used to investigate heterogeneity in treatment effects in subsets of the variable space that would be infeasible to conduct manually. This is not to say that ML can replace well-thought out and well-designed research studies; only that there are often dimensions of interest (which students benefit the most from a school lunch program?) that can now be tackled from different perspective. And much more is possible. As remarked by Athey (2018), "it is perhaps easier than one might think to make predictions about the impact of ML on economics, since many of the most profound changes are well underway." It seems the applied microeconomist must also be—in part—computer scientist.

This dissertation comprises three chapters that are grounded in this evolution of econometrics. They are meant to be self-contained, with their own appendices where appropriate, though the second chapter builds off the results from the first. A consolidated bibliography for all three papers is provided at the end.

The first chapter concerns land use and zoning regulations. It aims to answer the following question: how exactly do land use regulations impact urban development, specifically housing development? These seemingly-innocuous rules turn out to have large impacts on development and urban form. To arrive at this conclusion, I employ a by-now-established tool from the current econometric toolkit: regression discontinuity. By exploiting the discrete change in the level of regulation when crossing municipal borders in Massachusetts, I can compare how development differs between locations that are separated by less than 200m, subject only to different rules on building. As there is a dearth in widespread measures of the level of land use regulation, I derive my own by applying machine learning techniques—or natural language processing techniques more specifically—on the text from the zoning bylaws across the state. I take the approach of exploring for hidden "clusters" of towns, defined as having similar distribution of words in their zoning bylaws. Using existing measures of regulation that cover a subset of my data, I then determine which of these clusters captures highly-regulated towns. I find that stringent zoning codes strongly decrease the stock of housing, and that this is due to larger lot sizes (land consumption per house). Other potential explanations for the reduced housing stock, such as towns building fewer new houses or restricting the supply of developable land, are not supported by the data. For policy makers whose goal is to increase the supply of housing in high-growth regions, regulations that encourage higher consumption of land, such as minimum lot sizes and arduous shape requirements, should be scrutinized.

The second chapter is a descriptive paper, and an extension of the first. It has two

goals: to investigate the suitability of off-the-shelf ML methods in measuring the level of regulation and documenting the characteristics of high-regulated jurisdictions. The measurement aspect should inform researchers who want to use text-based methods on other legal texts (sections of bylaws or state laws) to derive a fuzzy measure of regulation stringency. In this particular setting, decomposition methods (such as principal components analysis and the latent Dirichlet allocation model) and ridge regression outperform decision tree-based methods (such as random forest). By documenting what towns enact burdensome land use regulations I seek to shed light on the causes of regulations. As this is still a poorly understood issue, I hope to help future researchers by establishing some facts in this regard. I find that historical density and land use patterns are the best predictors of current day regulation. Surprisingly, given some theories on the origins of land use regulations, historical demographic characteristics are weak predictors of present zoning. Lastly, I chart how town-level demographics have evolved since the implementation of widespread local-level zoning. The most striking has been the divergence of towns in the lowest quintile of regulation stringency in the share of non-white residents. This, and other results, provide some new stylized facts that can be useful in understanding the origins of such regulations.

The third and last chapter employs the experimental method to address an economically important question: can entrepreneurship be taught? We partner with a non-profit organization operating in Uganda and conduct a field experiment by randomizing admittance to entrepreneurship academies run at local universities. Though self-employment is widespread in Uganda, and in other sub-Saharan African countries, they are not of the sort that drives economic growth. The key issue here—in addition to efficacy—is whether training university students to become entrepreneurs creates new high-growth businesses, or simply diverts talented youths away from well-paying, formal sector employment. The chapter is based on a registered pre-results review written in June 2020 for ongoing field work. It discusses the research design and hypotheses being tested before post-treatment data is collected. Baseline data before participation in the academies for both control and treatment groups are available. Of three planned waves of academies, two have already taken place and a third is scheduled for later in the year, conditional on Covid-19 developments. The trajectory of the disease may also influence the scheduled data collection (described in the chapter) should there be further cancellations of university courses. However, current data collection efforts are promising. For example, to date we have been able to conduct the midline survey with over 92% of wave I treatment and control individuals through phone surveys, even as some individuals have return to their home villages during university closures.

Chapter 1

Land Use Regulations and Housing Development

Evidence from Tax Parcels and Zoning Bylaws in Massachusetts

Abstract: Land use regulations come in a wide variety of forms and govern how development occurs. They restrict housing development resulting in housing supply being less responsive to demand shocks. Yet little is known on what facets of residential development are most impacted, hindered by lack of comprehensive data on land use regulation stringency. I address this shortcoming by compiling a novel measure of land use regulation based on applying natural language processing techniques to over 40,000 pages of zoning bylaw texts. Utilizing a spatial regression discontinuity design around municipal borders, I find that stringent land use regulations reduce housing supply primarily through increasing the land usage per house. Strongly regulated localities do not compensate by developing more land overall. These results highlight how regulations like minimum lot sizes and setback requirements pose barriers to housing development in high-growth regions.

1.1 Introduction

Land use regulations (LUR) have been in place across US cities from the early 20th century, but did not take off until the 1970s (Gyourko et al., 2013). They dictate how land can be used, govern what can be built, where and how, and determine the role of local residents in the decision making process.[1] They are generally set at a local level, such as a township or municipality. These regulations are meant to address externalities from potential market failures: separating polluting sources from residential areas, reducing urban sprawl, coordinating development with transportation, amongst others. Recent work suggests, however, that the costs of restricting development outweigh the benefits (Turner et al., 2014; Hsieh and Moretti, 2019). These costs include reduced aggregate housing supply (Saiz, 2010; Hilber and Vermeulen, 2016) and housing markets that are less elastic to labour (Diamond, 2016) and immigration shocks (Saiz, 2007). However, there is a lack of evidence on how LUR impact housing development and therefore lessen supply. This is important for policy makers looking to understand the implications of LUR. Knowledge of these effects can also provide clues for *what* sorts of regulations are most consequential.

Moreover, it is infeasible to consider every potential regulation individually. First, the set of all possible regulations to choose from is large. Second, local stakeholders often have influence on the development process. Third, many regulations may seem different, but restrict development similarly. For example, restrictions on lot shape and lot width both serve to increase land usage per house. A common way to mitigate these issues is to create indices of LUR stringency meant to capture the overall regulatory burden. The most well-known of these indices is the Wharton Residential Land Use Regulation Index (WRLURI, Gyourko et al., 2008), derived from survey responses of 2649 US localities in 2005. Yet to answer questions on the impact of LUR that require measures of regulation stringency in out-of-sample jurisdictions, or in different years, there is no obvious solution.

This paper presents a novel measure of land use regulatory intensity derived from applying a machine learning method called the Latent Dirichlet Allocation model on over 40,000 pages of zoning bylaw texts from close to all municipalities in Massachusetts.[2] This paper focuses on the current regulatory environment in the state, but the technique can be applied in other jurisdictions on bylaws from different time periods. With this new index I then investigate how stringent LUR are manifested in housing development. First, I ask whether stringent LUR reduce the density of the housing stock at

[1] Some examples of LUR are minimum lot sizes, mandating a certain number of parking lots based on a building's size, establishing buffer zones around wetlands, and zoning (sorting usages across space).

[2] My measure has a coverage rate of 97.2% in Massachusetts compared to the 22.5% from the WRLURI.

the municipal level. Next, I ask how restrictive regulations are reflected in the housing market, in housing attributes, and in land use. This allows me to provide evidence on what *sets* of rules are most likely to be the most restrictive. Finally, I investigate whether these restrictions are reflected in local house prices.

To answer these questions, I apply several natural language processing (NLP) techniques to municipal zoning bylaws in Massachusetts.[3] To benchmark the resulting regulatory measures obtained from these text-based methods, I compare the NLP-derived results with the WRLURI, which is often used in the literature, as well as an index created from the Pioneer/Rappaport Housing Regulation Database (PRHRD). I find that an index created from a Latent Dirichlet Allocation (LDA) model (a latent finite-mixture machine learning model) best captures the variation in the land use regulatory environment. My LDA-based measure has a correlation coefficient with the other indices of around 0.6. This index is a natural complement to the existing measures: the current survey-based indices aid in interpreting the uncovered latent factors while the LDA method allows for expanding spatial and temporal coverage. Furthermore, it only requires the text from municipal zoning bylaws. I call this standardized measure the Natural Language Processing Zoning Stringency Index (NALPZ).

I then employ the NALPZ index in a spatial regression discontinuity design with municipal borders as cut-offs to evaluate the impact of stringent LUR on the pattern of housing development. This strategy exploits variation in the regulatory environment at 727 borders across 271 towns and controls for local housing demand, amenities, and tastes by comparing spatially close houses, only subjected to different LUR. Three main sources of data are utilized: i) housing characteristics from tax parcel data for all of Massachusetts, ii) Lidar data to calculate building heights,[4] and iii) land use data.

As there are various facets of housing development that can be affected by LUR, I group my outcomes into three main categories to add structure. The first group of outcomes are related to the housing market: building age and the year a house was last sold. They capture the rate of new housing development and turnover in the housing market. The second group of outcomes reflect housing attributes: building (livable) area, lot size, and building height. This group addresses how the shape and size of the houses themselves are influenced by stringent LUR. The final group of outcomes concern land use: the rate of conversion of undeveloped land to residential usage as well as the share of residential land of all developed land. These outcomes speak to how LUR shape the spatial pattern of residential development.

[3] Massachusetts is an ideal location for this analysis as it has previous measures of LUR to aid in benchmarking and interpretation.

[4] Lidar (light detection and ranging) is a remote sensing method for measuring distances, often used to derive high-resolution maps.

Though I lack exogenous variation in *specific* LUR, by evaluating the impact of the *overall* restrictiveness of LUR on different features of housing development, I can speak to what types of regulations are most likely binding. For example, regulations that cap the overall rate of development, such as growth controls, are more likely to impact the housing market rather than the attributes of the houses themselves. On the other hand, parcel-specific regulations, such as floor-area-ratios, are more likely to manifest themselves in the shape and appearance of houses.

This paper also speaks to the effect of LUR on local housing prices. Though LUR increase housing prices when considering average prices across municipalities in a regions (Hilber and Vermeulen, 2016), their effect at local levels depends on the degree of substitutability between nearby houses in different towns. From the tax parcel data, I observe the price the house was last sold for, as well as the most recent tax assessed value, broken down into building and land components. I use these outcomes in the spatial RDD to test how LUR impact house prices at the local level.

The results reveal several interesting findings. First, at the aggregate level, LUR strongly restrict development. The density of residential houses at town borders is lower in more restrictive municipalities. A standard deviation increase in NALPZ reduces housing density by 20% of the mean. Second, when considering the different groups of outcomes, I find that housing attributes react most strongly to restrictive regulations. For example, stringent LUR leads to significantly larger lot sizes (27% larger for each standard deviation increase in the regulatory index).[5] House sizes and heights are essentially not impacted by these regulations. Third, though the housing market and land use respond to LUR, the effects are economically small compared with housing characteristics. Houses are slightly older and a marginally higher fraction of developed land is allocated to residential use in more regulated towns. The rate of land conversion (undeveloped to residential) is not influenced by LUR. Fourth, house prices, after controlling for building and lot sizes, are about 5–6% higher for every standard deviation increase in the index. However, this is mostly explained by school district quality. This provides evidence for nearby houses across municipal borders being substitutes for one another.

I show that controlling for school quality measures, such as per pupil spending and graduation rates, school district fixed effects, or municipal property tax rates does not change the interpretation of my results. Several tests confirm that the results are not being driven by unobserved amenities or pre-existing differences in demographic characteristics. To test for the role of unobserved amenities, I estimate local amenity values with a canonical urban general equilibrium model (Ahlfeldt et al., 2015) and examine whether these vary differentially across municipal borders. Highly regulated towns do

[5]The average difference in the NALPZ across borders in Massachusetts is 0.74.

have higher amenity levels on average, but this relationship disappears when comparing neighbouring census block groups across borders, as my identification strategy does. Further, the demographic composition of census blocks at town borders pre-widespread LUR does not vary with current regulatory stringency.[6]

Taken together these results suggest that regulations that increase land usage *per* house are primarily responsible for constraining housing density and supply. For policy makers whose aims include increasing the availability of housing, regulations such as shape restrictions, setback requirements, and minimum lot sizes should be scrutinized.

This paper contributes to two strands of literature. The first strand of literature deals with the measurement of land use regulations.[7] I contribute to this literature by using natural language processing techniques on zoning bylaws to measure regulatory stringency. The novel use of a machine learning algorithm to measure LUR builds on previous work. The two most relevant works for this paper are the Massachusetts Regulation Database compiled by the Pioneer Institute for Public Policy Research and Rappaport Institute for Greater Boston (PIRI, 2005) and Gyourko et al. (2008). LUR are high-dimensional, coming in various guises, making it difficult to summarize the regulatory environment with a comprehensive measure.[8] This has made fully mapping the impact of more restrictive zoning on the pattern of housing development challenging. Remarkable attempts have been made to create measures of regulatory intensity. Researchers at the PIRI (2005) have compiled an extensive database of municipal land use regulations for towns around Boston (PRHRD). Gyourko et al. (2008) sent surveys to 6,896 jurisdictions across the US to establish the WRLURI. The thoroughness of these endeavours has resulted in a very detailed picture of their respective regulatory environments.

Though these have been necessary undertakings in helping our understanding of LUR, they are not without their disadvantages. First, the survey based methods suffer from nonresponse (Gyourko et al. (2008) had a 38% response rate). Even if the pattern of nonresponse is random, we lack measures of overall LUR stringency for many jurisdiction. Second, they are very resource intensive. Compiling the PRHRD required a research team consisting of a project manager, senior researcher, and twelve research assistants. This makes it difficult to administer in other localities in different time periods. Third, they are limited in the time dimension, usually depicting the regulatory environment at

[6] A census block corresponds roughly to a city block. A census block group contains 31.6 blocks on average.

[7] A broader overview of this literature is given in Section 1.3.

[8] An illustration of the numerous sorts of restrictions posed by LUR can be found in the PIRI (2005) database. They gather data on the regulatory environment around Boston through bylaws and surveys, and generate 119 variables meant to describe it. As many of these regulations affect development in similar ways, for example by reducing density, focusing on just a small subset of all regulations can lead to incorrect conclusions.

one point in time.[9]

I extend this previous work by creating an index of regulatory restrictiveness (NALPZ) that covers the vast majority of municipalities in Massachusetts (341 of 351, compared to 187 from the PRHRD and 79 covered by the WRLURI). Existing measures of LUR aid in the interpretation of the estimated latent categories, making my measure a complement to the previously established LUR measures.

As measures of land use restrictiveness are vital to many studies that require estimates of city-level housing supply elasticities, this technique provides a viable method to increase both the geographic coverage and time frequency of a LUR index. Researchers estimating Rosen-Roback style models (Diamond, 2016; Hsieh and Moretti, 2019) require measures of the responsiveness of local housing markets. Other papers use LUR measures to study their direct effects on welfare (Turner et al., 2014). A large body of research is interested in related questions that require measures of LUR or housing-supply elasticities derived from LUR measures (Dettling and Kearney, 2014; Hilber and Turner, 2014; Albouy, 2016; Aladangady, 2017; Stroebel and Vavra, 2019). Virtually all these papers use either the WRLURI as a measure of regulatory intensity, or housing supply elasticities from Saiz (2010), who in turn uses WRLURI to separate the contribution of geographic and regulatory restrictions on housing supply.

The second strand is concerned with the consequences of stringent land use or zoning regulations.[10] This paper contributes to this literature in two dimensions. First, it identifies how restrictive LUR manifest themselves in housing development. Second, it speaks to the substitutability of nearby houses across municipal borders. Previous work has established a credible link between the implementation of stringent LUR and the reduction in aggregate housing supply and the increase in low-density developments (Mayer and Somerville, 2000; Saks, 2008; Glaeser and Ward, 2009; Turner et al., 2014; Diamond, 2016; Jackson, 2016; Hsieh and Moretti, 2019), and increasing house prices (Ihlanfeldt, 2007; Turner et al., 2014; Hilber and Vermeulen, 2016; Severen and Plantinga, 2018). Other research has found that LUR result in additional negative externalities: they exacerbate geographic sorting and inequality (Saks, 2008; Diamond, 2016; Ganong and Shoag, 2017; Hsieh and Moretti, 2019), increase price volatility (Glaeser et al., 2008; Jackson, 2018), as well as encourage land use conversion (Irwin and Bockstael, 2004; Sims and Schuetz, 2009; Shertzer et al., 2018).[11] This paper makes an important contribution by consider-

[9] It is possible to ask local officials about *when* regulations were implemented, as the PIRI (2005) do. But it is more challenging to gather the data repeatedly over several years.

[10] Though often used interchangeably, land use regulations subsume zoning regulations. LUR also include environmental regulations, for example.

[11] In an address to the Urban Institute in 2015, Jason Furman, chair of then President Barack Obama's Council of Economic Advisers, claimed that "excessive or unnecessary land use or zoning regulations

ing how stringent LUR are reflected in the pattern of housing development. The results on the substitutability of neighbouring houses subjugated to differing levels of land use restrictiveness are consistent with the model of Helsley and Strange (1995), where house price differences are insignificant between closely substitutable towns.

The remainder of the paper is structured as follows. Background on land use regulations in Massachusetts is given in Section 1.2. Section 1.3 outlines the construction of the regulation index. Section 1.4 describes the data, Section 1.5 the empirical strategy, and Section 1.6 the main results. Section 1.7 discusses balancing, specification, and robustness checks. The last section offers conclusions. Specifics on the natural language processing techniques, model used to estimate local amenities, and additional details can be found in the Appendix.

1.2 Institutional Setting

Though municipalities in Massachusetts have utilized LUR since the 1930s, this was more the exception than the norm, as they needed to get approval from the state if they wanted to deviate from the state-level development guidelines. This changed with the introduction of the Massachusetts Zoning Act in 1975. The goal of this act was to "facilitate, encourage and foster the adoption and modernization of zoning ordinances and by-laws by municipal governments." Effectively, it made it easier for municipalities to implement their own zoning bylaws without significant state interference. There is corroborating evidence that LUR were not restricting house construction nationwide until the 1970s (Gyourko et al., 2013).

After the Act was passed, municipalities began implementing their own LUR almost immediately and in quick succession. Survey data collected by the PIRI (2005), and plotted in Figure A.16, show the cumulative share of towns that have implemented a specific category of LUR by each year. Before 1975, only regulations regarding subdivisions (dividing a parcel of land into smaller parcels) were common. Afterwards, various types of LUR were introduced in different municipalities in quick succession.

Comparing zoning districts across municipalities is difficult, as there is no standardized classification of the district types. Furthermore, as municipalities set their own zoning regulations, land that is zoned for single-family residential in one town may be subjugated to a completely different regulatory environment than similarly zoned land in another town.

---

have consequences that go beyond the housing market to impede mobility and thus contribute to rising inequality and declining productivity growth." (Furman, 2015)

Figure A.16 also provides some examples of different categories of LUR.[12] For example, subdivision regulations are concerned with the division of a parcel of land into smaller units (*eg* a developer buying a parcel of land and building more than one house on the parcel with the intention of selling them as individual units) while wetland regulations deal with issues surrounding development around stagnate bodies of water (*eg* how close buildings can be to wetlands). Zoning bylaws generally cover some of the categories of LUR, but others (such as septic regulations) are often their own section in the municipal bylaw code.

1.3 Measuring Land Use Regulations

LUR are notoriously difficult to measure. This is primarily because they operate in a high-dimensional space; LUR are written in a variety of different ways, which are often not consistent across municipalities, making direct comparison of bylaws across towns a challenging task.

Numerous strategies have been employed to measure the regulatory environment for land use or zoning. One strategy has been to focus on the implementation of or amendment to a single law (Zhou et al., 2008; Kahn et al., 2010; Severen and Plantinga, 2018). Other research has addressed this issue by calculating the share of LUR policies employed from a set of possible categories, and using this as a proxy for LUR intensity (Mayer and Somerville, 2000; Quigley and Raphael, 2005; Geshkov and DeSalvo, 2012).

Another approach is based on the assumption that the construction market is relatively competitive. These papers then use hedonic regressions or calculate price-to-cost ratios to infer the extent that LUR are impacting the housing market (Glaeser and Gyourko, 2003; Glaeser et al., 2005). Utilizing the same data source as I, other researchers gather information on LUR directly from zoning bylaws, normally by hand (Evenson and Wheaton, 2003; PIRI, 2005; Brooks and Lutz, 2019). Yet another more recent approach from Brueckner et al. (2017) combines data on prices for parcels of land in China with floor-area-ratio limits in an Alonso-Muth-Mills model to infer the stringency of regulation.

The most informative method—in terms of quantity and quality of information—has been to conduct surveys with local officials who are responsible for setting land use policy, often in combination with gathering data from primary sources (Levine, 1999; PIRI, 2005; Gyourko et al., 2008). These efforts to establish a unified database on LUR have been substantial undertakings; research teams had to either scour town websites

[12] There are various other categories that LUR can belong to, and Figure A.16 is in no way exclusive.

and digitize the information or interview town officials in order to gather the necessary data to compile these resources, sometimes both.

Though these previous undertakings in measuring LUR provide an important and necessary starting point, the spatial regression discontinuity design described in Section 1.5 benefits from a measure of regulatory stringency that has close to universal coverage over a large region. I address this need by creating a regulatory index that is based on the text from zoning bylaw documents from nearly all the towns in Massachusetts.

Specifically, I compile a corpus of municipal bylaws by gathering the documents directly from the various municipalities' websites. I then consider several different natural language processing (NLP) techniques and compare their results with the survey responses from the Pioneer/Rappaport Housing Regulation Database (PRHRD) and Gyourko et al. (2008).[13] NLP are a natural fit in this particular setting, as they are primarily used on unstructured text data. Importantly, the set of NLP techniques considered can all be classified as "unsupervised" methods, as they use no information from the survey data used to assess fit. This is to avoid over-fitting any derived measure to the limited number of towns available for benchmarking.

As there is no *ex ante* best technique to measure the level of regulation, I consider several different NLP methods. The first method considered is the simplest: the number of meaningful[14] words the zoning section of a bylaw contains. I refer to this as the "document length" measure. In addition to being the simplest, it also allows for testing the somewhat intuitive hypothesis: that the longer the zoning bylaw of a town is, the more stringent its LUR.

The second method employs a group of NLP techniques called either "dictionary methods" or "sentiment analysis". These methods utilize pre-defined dictionaries of terms that belong to a specific category (*eg* the category "land" would contain words such as "valley"). The more words in a document that belong to a specific category, the more that category represents the document. The implicit hypothesis being tested here is that words from one (or more) of these dictionaries are found more often in more (or less) regulated towns' bylaws.

The final method involves estimating a Latent Dirichlet Allocation (LDA) mixture-model, an unsupervised machining learning technique (Blei et al., 2003) that assumes that the distribution of observed words derives from latent, unobserved categories (nor-

[13] The measure of regulatory stringency (WRLURI) constructed by Gyourko et al. (2008) covers 79 towns in Massachusetts. Since the PIRI (2005) does not have a summary measure of the regulatory environment, I use their detailed data on the regulatory environment for 187 towns around Boston to create a measure of regulatory intensity.

[14] *ie* not "filler" words such as conjunctions or pronouns.

mally called "topics"). Given a pre-defined number of latent categories, it estimates the probability a specific word was drawn from a specific category and assigns probabilities to each document over the distribution of the latent categories. Under the assumption that one (or more) of these latent categories determines regulatory stringency, I test whether the derived probabilities correlate with the survey based measures.

Of the NLP measures considered, I find the topic probabilities from the LDA model best capture the regulatory environment in Massachusetts, as judged by its high correlation to the two LUR existing indices. I call this the Natural Language Processing Zoning Stringency Index, or NALPZ.

This section is broken down as follows. First, I describe the data I need to employ the NLP techniques I use, as well as the data I later use to benchmark the measures I derive. Second, I characterize in greater detail the NLP methods I introduced above. Finally, I investigate how well the measures derived from NLP techniques are able to explain the variation in regulatory intensity found in the survey data.

Measurement Data

Zoning Bylaws

I compile a corpus of municipal bylaws by gathering zoning bylaw documents directly from the various municipalities' websites. These documents are either in PDF or word format. Generally, the text is semi-machine readable, but in a few cases, the PDFs are scans of a paper version of the zoning bylaws. To extract the text from these documents I apply Optical Character Recognition software to the PDFs to extract the text. It is important to note that though the ordering of the text is not necessarily maintained, the methods I use do not utilize the ordering of the words, only their (relative) frequency. Of the 351 towns in Massachusetts, I am able to get bylaw documents for 341 of them, which are either up-to-date or only up to a couple of years old. The ten towns where zoning bylaw data is missing are smaller on average and have more rudimentary websites and digital services, and represent only 0.49% of the population of Massachusetts (2010 US Census).

Once the text is extracted, it needs to be preprocessed before it can be used. This involves, among other things, tokenizing the text (separating the text on whitespace and other characters into tokens), filtering (language-specific common words and stopwords are removed, such as "the", "or", "that"), and lemmatization (all words are reduced to their base form, "measurement" and "measured" become "measure").[15]

[15]The individual words are referred to as "tokens" in keeping with the literature. This highlights the fact that they may have been altered (*ie* through lemmatization) and filtered, and thus do not correspond

Additionally, I filter out some custom, context-specific stopwords. This includes website specific stems from downloaded documents (e.g. "http", "com") as well as legal-text specific words (e.g. "doc", "sec"), which do not provide much valuable information,. Then, I filter out tokens that are either mentioned in almost every document (e.g. "month", "equipment", "waste"), or mentioned in just a couple (usually the names of towns). Finally, I weight the tokens using the term frequency-inverse document frequency method, which gives more weight to tokens that appear in fewer documents.[16]

Then I count the occurrences of each token (raw and weighted) in each document. This results in a document-term matrix, where each row is a document (here: town zoning bylaw) and each column a token. The cells refer to the number of times a word appears in a specific document (raw or weighted).

The top raw and weighted words are shown in Figure A.1.

Dictionaries

With the processed documents, I apply various text analysis methods, investigating whether they map into the WRLURI or the PRHRD data. One common method is the "dictionary-based method" or "sentiment analysis". Essentially, using predefined dictionaries consisting of words belonging to a specific topic, one counts the occurrences of the dictionary words appearing in each document (raw or weighted).

I use dictionaries from the Harvard IV-4 Categories to perform sentiment analysis on the zoning bylaw documents. I use nine dictionaries labelled "active", "aquatic", "building", "land", "legal", "nature", "object", "place", and "region". Examples of words belonging to each of these categories are given in Appendix A1.

Pioneer/Rappaport Housing Regulation Database

Though previous data on the regulatory environment for land use across Massachusetts, indeed across the US, has been scant, there have been a couple of notable undertakings. One such project was carried out through a joint project by the Pioneer Institute and the Rappaport Institute in Massachusetts.

They collected data on LUR for most of the towns in a 50 mile radius around Boston (187 of 351 towns in Massachusetts). To built their database, they gathered data in 2004 from municipal websites, through Ordinance.com, and through phone calls if necessary. The data they gathered belonged to one of four categories: i) Zoning, ii) Subdivision,

directly to the words in the original text.

[16] Details on the weighting scheme are described in Appendix A1

iii) Wetlands, or iv) Sewage Disposal. The end result is the Pioneer/Rappaport Housing Regulation Database (PRHRD)

Besides being a thorough description of the regulatory around Boston, it also allows me to test the external validity of the NLP-derived measures and to explore how the results I find later in the empirical section vary when using a different measure of regulatory stringency. As the PIRI (2005) do not create an index themselves, I use their data to construct a regulation index through principal component analysis.

Wharton Residential Land Use Regulation Index

Though the database compiled by the PIRI (2005) is very thorough for the area around Boston, it is not the most commonly used measure of land use regulatory restrictiveness, mainly due to its limited geographic coverage and the fact that it does not come with a uni-dimensional measure of LUR. Gyourko et al. (2008) address these two issues by focusing on breadth rather than depth.[17]

They sent surveys to 6,896 municipalities (with 2,649 responses) across the US to gather information on the state and local regulatory environment. Applying factor analysis to these responses, they create an index, the Wharton Residential Land Use Regulation Index (WRLURI), that is meant to capture the stringency of land use control.

The WRLURI is also a measure of regulation intensity that is often used in the urban economics literature[18] (See Saks, 2008; Glaeser and Kahn, 2010; Saiz, 2010; Turner et al., 2014; Diamond, 2016; Ganong and Shoag, 2017; Hsieh and Moretti, 2019, for example), making it a particularly useful index to compare with. Gyourko et al. (2008) have provided their data along with this index for other researchers to use.

Though more commonly used than the PRHRD, only 79 municipalities in Massachusetts have a WRLURI value. Therefore, it is important to note that while a comparison with this index can be worthwhile, it is based on a (probably) non-random subsample of all towns in Massachusetts. Should the effect of more stringent LUR be heterogeneous, results using the WRLURI and any NLP-derived measure could differ without indicating a problem with the identification strategy.

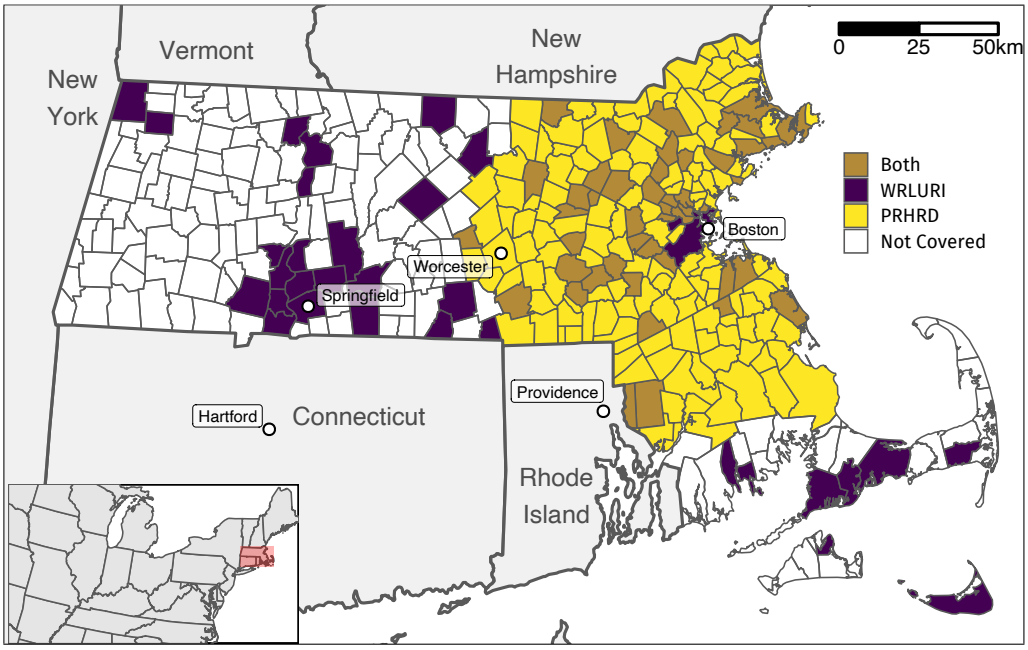
A map of the data coverage of both the PRHRD and WRLURI coverage in Massachusetts can be seen in Figure 1.1, where the yellow (light) municipalities indicate data available

[17] These issues do *not* imply any major shortcomings of the work of the PIRI (2005) or Gyourko et al. (2008), but rather refer to a fundamental trade-off between scope and detail with limited resources, where the former focus on detail and the latter on scope. See Gyourko and Molloy (2015, pp. 1298) for a discussion on this point.

[18] It is also used in other fields, notably in labour and public economics, when modelling housing supply response is vital.

from the PRHRD and purple (dark) WRLURI measures from Gyourko et al. (2008). Brown (grey) shaded towns are covered by both datasets, of which there are 47.

Figure 1.1: Current Data on Land Use Regulations in Massachusetts



Notes: Data come from Gyourko et al. (2008) and the PIRI (2005). WRLURI is the Wharton Residential Land Use Regulation Index from 2005. PRHRD is the Housing Regulation Database of Massachusetts compiled by the Pioneer Institute and the Rappaport Institute in 2004. I create an index from their data using Principal Component Analysis.

Natural Language Processing Methods

The following subsection will cover the various NLP techniques employed, and the measures derived from them that will be compared to the indices from the survey data mentioned in Section 1.3. In what follows, d will refer to a document (*ie* a town's zoning bylaw text), v a unique token (processed and meaningful words), x_{vd} the count of tokens v in document d, and tf-idf$_{vd}$ the term frequency-inverse document frequency weighted version.

Document Length

The first NLP technique consists of simply counting the number of tokens in each document. This is done for both the raw and tf-idf weighted tokens. It is conceivable that a longer bylaw document (*ie* more tokens) could be an indication of stricter LUR. Formally,

these measures are calculated as follows:

$$\begin{aligned} \mathrm{L}_d &= \sum_{v=1}^{V} x_{vd} \\ \widetilde{\mathrm{L}}_d &= \sum_{v=1}^{V} \text{tf-idf}_{vd} \end{aligned} \tag{1.1}$$

where V refers to the set of all unique tokens in the corpus.

As can be seen in the left panel of Figure A.2, there is a large variance in the length of these bylaw documents. Unsurprisingly, the largest municipality in the dataset, Boston, has the longest zoning bylaw. The distribution of the weighted counts also varies, though the normalization results in the distribution becoming closer to normal.

These two measures, raw token counts and tf-idf weighted token counts, are the first two NLP-derived candidate measures.

Dictionary Methods

The second technique involves tallying the number of tokens (raw or tf-idf weighted) that belong to a specific category (dictionary) for each document. For example, a dictionary labelled "Legal" contains tokens such as "advocate", "prison", and "ordinance". This is similar to calculating document length, but only considers tokens in the respective dictionaries. Formally:

$$\begin{aligned} \mathrm{C}_d^c &= \sum_{v \in \mathcal{D}_c} x_{vd} \\ \widetilde{\mathrm{C}}_d^c &= \sum_{v \in \mathcal{D}_c} \text{tf-idf}_{vd} \end{aligned} \tag{1.2}$$

where c indexes the chosen dictionary, and $\mathcal{D}_c$ represents the tokens in the dictionary.

The distributions of the measures for the various dictionaries are shown in Figure A.3. In the left panel the raw counts are divided by the document length, so the values can be interpreted as the share of tokens in the document that belong to the respective category. The more tokens a document has in a particular category, the more that category represents the document. Though the scale changes, the results vary little between the raw and unweighted categories. Words from topics such as "object" or "land" appear little in the bylaws, whereas words belonging to "active" and "place" appear regularly. These topics with high occurrences are also the topics that display the most variation.

These category-specific token counts are the second group of NLP-derived measures I consider.

Latent Dirichlet Allocation

The next NLP method I consider is a multinomial mixed-membership model with latent topics, called the Latent Dirichlet Allocation (LDA) model, first described by Blei et al. (2003). This estimator assumes that each document is a mixture over K topics (document-topic distribution), and the topics in turn have a probability distribution over each word v (topic-word distribution). The dispersion of the probabilities from these two distributions are governed by priors. Concretely, the document-topic probabilities are assumed to be distributed as:

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha) \tag{1.3}$$

and the topic-word probabilities as:

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\delta) \tag{1.4}$$

We observe the words in a document d with N_d words overall, resulting in $\boldsymbol{w}_d = (w_1, \ldots, w_{N_d})$. Furthermore, we assume that each word was generated from one of the K latent topics. Conditional on the document-term and topic-word probabilities, we can now describe the generating process of each word (i) for every document (d):

$$z_{id} \sim \text{Multinomial}(\boldsymbol{\theta}_d) \tag{1.5}$$

$$w_{id} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{id}}) \tag{1.6}$$

In the simplest case of $K = 2$, a document d would belong to topics 1 and 2 with probabilities θ_{d1} and θ_{d2}, respectively. If one topic, k, is more indicative of more stringent LUR, then θ_{dk} would be a measure of this. A visual depiction of the assumed data generating process is shown in Figure A.4. It is important to stress that this is a unsupervised machine learning method, which means that there is no "outcome" variable. The LDA method finds patterns in the text in order to assign topic probabilities to each document.

Conditional on K, I derive a measure of regulation intensity, S_K, by choosing the vector of document-topic probabilities that correlates the strongest with the WRLURI. Specifically:

$$\text{S}_K = \underset{\boldsymbol{x} \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}}{\arg\max} \; corr(\widetilde{\boldsymbol{x}}, \text{WRLURI}) \tag{1.7}$$

where $\widetilde{x}$ represents the normalization of the variable x to have the same moments as the standard normal distribution.[19] I do this because the magnitude of the unstandardized

[19] This is done by ranking the vector $\boldsymbol{x}$ and then using the quantile function of the standard normal to get a normalized score.

measure is not very informative; rather, it is the relative rank of the S_K's that have meaning. Furthermore, though the ranking of the documents by topic probabilities is invariant to choices on the priors, the distance between these probabilities is not. Note that the vector $\boldsymbol{\theta}$ is now index by k rather than d.

The LDA model has several parameters that must be set by the researcher. The parameters on the two prior (Dirichlet) distributions control the dispersion of the document-topic and topic-words probabilities. As I use the the $\boldsymbol{\theta}$'s to construct my regulation index, the prior on the document-topic distribution is more relevant (α). However, though the prior affects the dispersion of the topic probabilities, the relative topic probability rankings are stable. And since I standardize the measure in the end, the priors do not affect the creation of my index. Thus, for the parameter on the topic-word prior distribution I use the standard in the literature, $\delta = 0.1$ (Griffiths and Steyvers, 2004) and for the prior on the document-topic distribution I choose a prior that results in a wide coverage of probabilities.[20] I find that setting the Dirichlet parameter $\alpha = 1/K$ works well in practice.

The most important parameter to choose in my setup is the number of topics, K. Since the LDA is an unsupervised method, there is no outcome with which to measure goodness-of-fit. Thus I investigate how well the number of topics fits my data through 5-fold cross-validation. I first look at a measure called "perplexity", which can be thought of as a likelihood of the estimated model *given* validation data not used in the estimation, where a lower number indicates better fit. Then I look at how well the document-topic probabilities correlate with the WRLURI; in other words, I calculate $\rho = \max_{\boldsymbol{x} \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}} corr(\widetilde{\boldsymbol{x}}, \text{WRLURI})$ for each fold of each K considered. I plot the results from the cross-validation in Figure A.5. The first panel shows the perplexity measure, with each point indicating a fold and the solid line passing through the average of the five folds. In the second panel, I plot the ρ from above. Though the model is only estimated with the test data for each fold, I calculate the correlation for the entire sample.

Going only off of the perplexity measure, using a model with $K = 25$ would seem to be most appropriate. However, though introducing more topics may help with fit, the results become more difficult to interpret. Thus I focus on the results from the second panel to choose K. From it, an LDA model with three topics seems the most preferred. Not only is the average correlation with WRLURI the highest, but the variation among the different folds is lower.

Using the LDA model with three topics (i.e. $K = 3$), I obtain the distributions of

[20] *ie* not clustered around 0%, 100%, or $100/K\%$. The last would imply that a document is nearly equally well described by any of the latent topics.

the three topics (β_{1v}, β_{2v}, and β_{3v}, where β_{kv} is the probability that word v is drawn from topic k) over all tokens. The tokens with the highest probability per topic are shown in Figure A.6. It is important to note that though the model returns a distribution of words for three separate topics, the labelling and interpretation of the topics is left ambiguous. Though I refrain from attaching any labels to the topics, there are still evident patterns to the word groupings. For example, Topic 2 has words that deal with renewable energy sources ("photovoltaic" and "wind") and deal with hedonic uses ("marijuana" and "adult"). Topic 1 concerns itself more with vocabulary describing cities ("urban" and "sidewalk") as well as terms associated with development ("ratio" and "affordable"). Topic 3 is more mixed and is concerned with amenities ("entertainment") and types of development ("mixed", "cluster"). There is obviously lots of overlap, making it difficult ex ante to assign meaning to the latent topics.

Turning to the ability of the document-topic probabilities to capture variation in LUR, I find that Topic 2 is most correlated with a high level of restrictiveness (higher WRLURI) whereas Topics 1 and 3 are negatively correlated with regulatory stringency. This suggests using θ_{d2} as a NLP-derived measure of LUR for comparison to both the other derived measures and the PRHRD regulatory index.

To get at the words that best differentiate between the most "regulated" topic (2) and the two less regulated "topics" 1 and 3, I plot the largest absolute logarithm differences between the the probability that a word belongs to Topic 2 against Topic 1 or 3 in Figure A.7.[21] Words at the top in blue are more likely to appear in Topic 2 (stronger regulations), whereas those at the bottom in red have a higher probability of being in Topic 1 or 3 (weaker regulations). For example, the words "city" and "mixeduse" are more descriptive of Topic 1 or 3, whereas "Annual Town Meeting" and "photovoltaic" are more indicative of Topic 2.

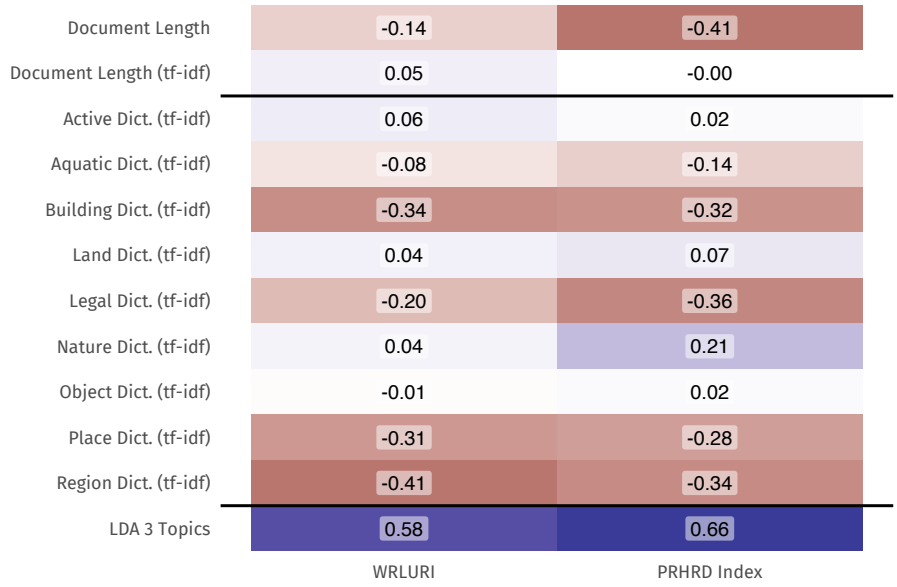
Natural Language Processing Zoning Stringency Index

With several NLP-derived measures in hand, I now turn to comparing them with the two regulatory indices described earlier. The results are shown in Table 1.1. It shows the correlation between the NLP-derived measures discussed in this section, namely the document length measures, the dictionary-based measures, as well as topic probabilities from the LDA model, and the regulatory indices WRLURI and the PCA index from the PRHRD data.

The top two rows correlate the length of the documents (based on the sum of raw or weighted tokens per document) with the regulatory measures. Somewhat surprisingly,

[21] $\log_2\{\beta_{2v}/(\beta_{1v}+\beta_{3v})\}$

Table 1.1: Correlations of Natural Language Processing Regulation Indices Candidates and Existing Indices



Notes: Data come from Gyourko et al. (2008) and the PIRI (2005). WRLURI is the Wharton Residential Land Use Regulation Index from 2005. PRHRD is the Housing Regulation Database of Massachusetts compiled by the Pioneer Institute and the Rappaport Institute in 2004. I create an index from their data using Principal Component Analysis.

the length of the bylaw documents does not seem to be strongly related to the regulation indices, either raw or unweighted. In fact, in the raw case, shorter zoning bylaw documents come from towns with more stringent LUR.

Turning to the middle section of rows, I then show the correlation of the tf-idf weighted counts of tokens belonging to one of the dictionaries with the regulatory measures.[22] I find that none of the nine dictionaries used to score the documents results in a measure that is strongly correlated with the regulation indices. Towns that use words from the dictionaries "region", "place", and "building" appear to be somewhat *less* regulated, with correlation coefficients between -0.28 and -0.41.

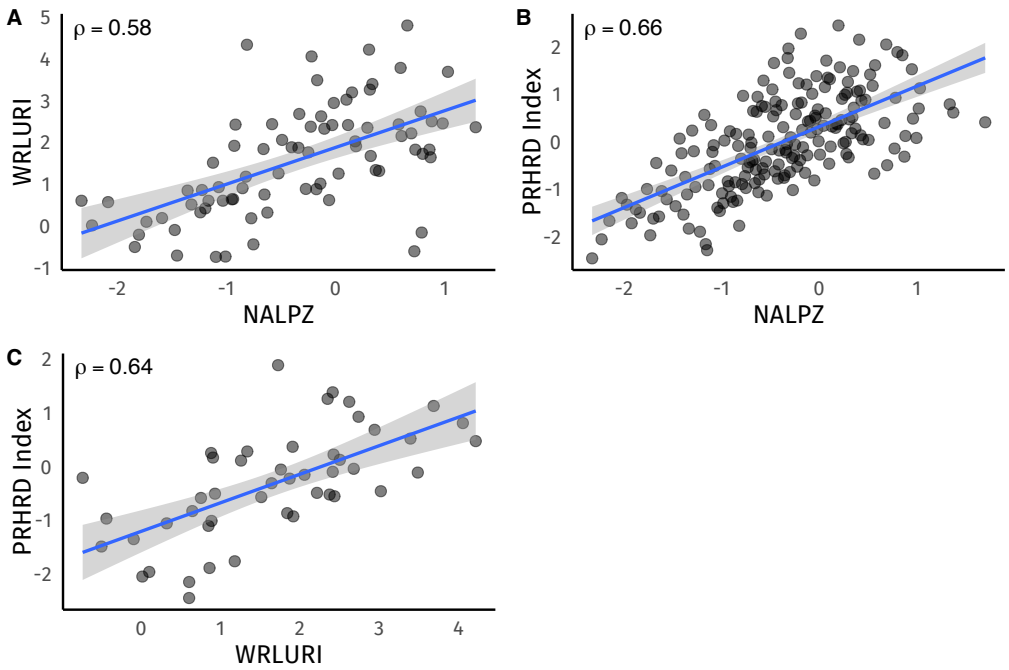
However, turning to the bottom row, it is apparent that the topic probabilities from the LDA model are strongly related to the two indices considered here. S_2 is strongly correlated with the aggregate LUR indices. I name this measure the Natural Language Processing Zoning Stringency Index (NALPZ). It has a correlation coefficient of 0.66 with the PCA index from the PRHRD data and 0.58 with the WRLURI. The mixture-model estimates latent topics that are strong predictors of LUR by finding patterns in the text contained in the bylaw documents.

Given the ability of NALPZ to capture the regulatory environment seemingly well, I

[22]The results using raw counts, unreported, are quite similar.

further examine the relationship between it and the other indices of regulation. These are shown in the scatter plots in Figure 1.2. Panel A and B display the relationship between NALPZ and the two survey-based indices. As can be seen, NALPZ does a good job of explaining these other two measures, especially given that it uses *none* of that information in its creation. Moreover, NALPZ is significantly easier to derive, requiring only the raw text given in the bylaws. The last plot, C, gives a sense of how the WRLURI measure varies with the PRHRD index.

Figure 1.2: Comparison of Natural Language Processing Zoning Stringency Index (NALPZ) with Existing Measures of Land Use Restrictiveness

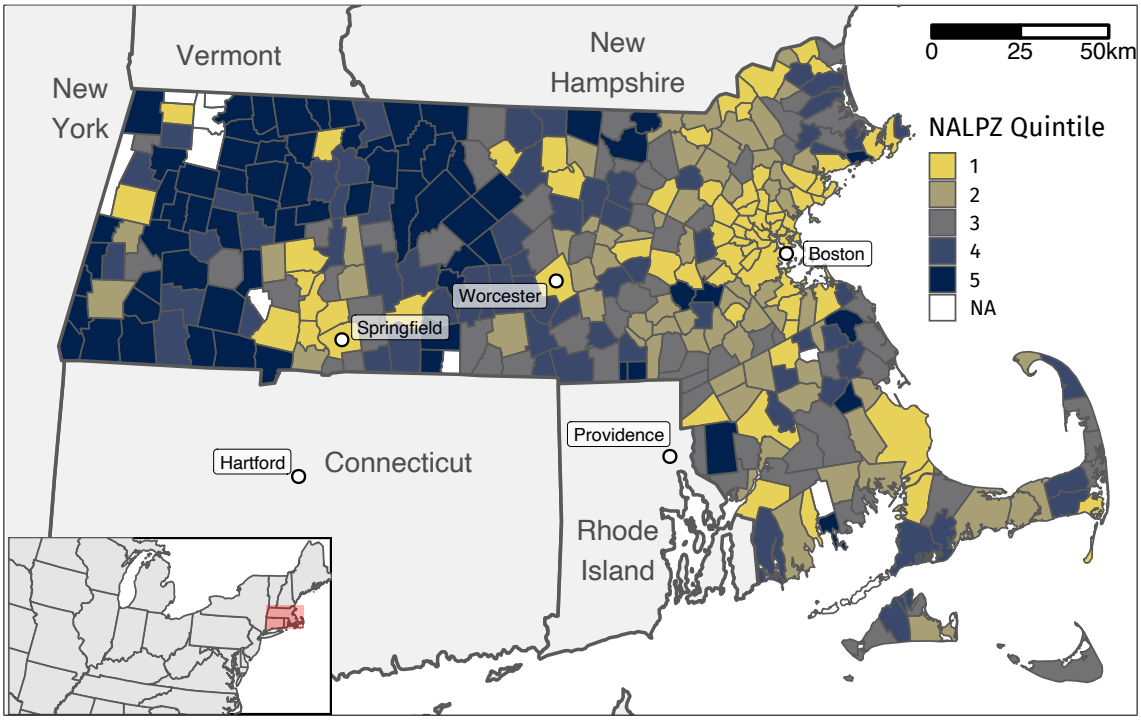


Notes: WRLURI (Gyourko et al., 2008) is the Wharton Residential Land Use Regulation Index from 2005. PRHRD (PIRI, 2005) is the Housing Regulation Database of Massachusetts compiled by the Pioneer Institute and the Rappaport Institute in 2004. I create an index from their data using Principal Component Analysis.

It is also worth briefly discussing the disadvantages of using NALPZ as a measure of the regulatory environment. The primary concern is that it is a "black box"; what exactly results in a higher or lower index value is not entirely clear. A town may be regulated more stringently either through implementing more LUR or making the current LUR more intensive. However, I am able to investigate *which* words are more often found in documents with a higher regulation index value, given the estimated topic-specific word probabilities. Furthermore, the ease of implementation makes this measure a nice compliment to the survey-based measures to help expand geographic coverage and potentially add a time dimension to current regulation measures.

A map of the spatial distribution of NALPZ is shown in Figure 1.3. As has been found already by Gyourko et al. (2008), larger (in terms of land) municipalities with lower population densities tend to have a stricter regulatory environment than average, as can be seen by the cluster of dark blue towns in western Massachusetts. Larger metropolitan areas, such as those around Boston, Worcester, and Springfield, are less regulated in general.

Figure 1.3: Natural Language Processing Zoning Stringency Index (NALPZ) Across Massachusetts



Notes: NALPZ is an index of land use regulation stringency derived from a Latent Dirichlet Allocation model applied to the zoning bylaw text of towns across Massachusetts.

This paper focuses on the *impact* of stringent LUR on housing development. However, Appendix A2 further explores the related question of the causes of stringent regulation. It discusses the spatial pattern of LUR in Massachusetts and what pre-existing town characteristics best predict a town's regulation restrictiveness.

1.4 Data

The NALPZ from Section 1.3 is the main explanatory variable. It captures the relative stringency of a town's LUR. To measure the impact of restrictive regulation, I compile data on current housing characteristics, land use over several decades in Massachusetts. Several additional data sources are included for the tests performed in Section 1.7. These

include census data on demographics, bilateral travel times between areas, school district quality measures, and property tax rates.

Massachusetts Standardized Assessors' Parcels

To effectively employ the spatial RDD design, the data I use need to have two vital pieces of information. First, they need to contain information on the characteristics of single-family houses across Massachusetts, and second, the houses need to be precisely geocoded to compare units within close geographic proximity.

The Massachusetts Standardized Assessors' Parcels database fulfills these two criteria. It contains every parcel of taxable land in the state of Massachusetts, except for Boston (which maintains its own records and database). The data are collected from each municipalities' tax assessors, under guidelines set forth by the Massachusetts Bureau of Geographic Information (MassGIS), that are then compiled by MassGIS.

The database consists of several tables, two of which are used for this analysis. The first is the Assessor Data Extract that contains all the information that tax assessors gather on each taxable parcel in their respective municipality. Some data included in this table are the assessed tax value of the structures (buildings) and the land, along with some information about the characteristics of the structures (*eg* year built, number of rooms, lot size). It also has information on the size of the parcels of land and of the building itself. Importantly, the data also contain the "use code" for the parcel. This code classifies each property based on its primary use such as residential, commercial, agriculture, etc.

The second table is the Tax Parcel Attribute file. This file contains the polygon for each parcel of land in the state of Massachusetts, each with an unique location identifier, as well as a shapefile indicating its precise location within the state. The Assessor Data Extract can be matched to the Tax Parcel Attribute file through the location identifier. Though most of the matches between the two files are one-to-one, in some cases several assessment units are matched to one tax parcel. For example, units in a condominium may have different owners, and thus be taxed independently, but share the same geographic location.

For my analysis, I only consider single-family houses. This is primarily done to keep in line with previous literature, and due to the fact that they represent the vast majority of residential structures in Massachusetts (here: 73% of taxable residential parcels). I am able to identify these parcels by the use code. I also filter out parcels that are located in municipalities without a NALPZ value, as well as parcels that are bordering a town without a NALPZ value (as they would have no comparison units).

From the tax parcel data, I am also able to estimate whether it is owner occupied. The data contain the address of both the property itself, as well as that of the owner. From this, I compute the Levenshtein distance between them. Parcels where this value is less than eight I code as owner-occupied.[23]

Of the 2,319,906 parcels overall, 951,971 remain after filtering. Summary statistics on the tax parcels are given in Table 1.2.[24] Last sale price is not reported for every tax parcel, thus I do not exclude units from the baseline sample when this is the only detail missing. Furthermore, in analyses using last sale price, I exclude parcels where this is less than $5,000 to include only arm's length transactions.

Building Heights

An important aspect of development is the height of buildings. This attribute is not contained in the tax assessment data, unfortunately. I overcome this lack by calculating building height from highly accurate lidar data in combination with a digital elevation model for Massachusetts and a shapefile containing all buildings in the state.

Lidar refers to light detection and ranging, a remote sensing method using reflected light to measure distances. The National Oceanic and Atmospheric Administration provides large quantities of lidar data covering all of Massachusetts. The main attributes of the data are its x, y, and z coordinates. This describes where the data point was measured in terms of longitude (x) and latitude (y), as well as its height above sea level (z). Together with the building shapefiles I am able to assign the highest point above sea level for each building. To derive elevation, I use the digital elevation model to get the elevation of the base of the houses. The height is then simply the difference between these two values.

I restrict the building height sample to buildings located on tax parcels that are coded as single-family residential. This allows the results for the tax parcel level and building level to comparable.

Land Use

Data on land use is provided by MassGIS, which classifies land into 37 different categories based on aerial images. There is data available for the years 1971, 1985, and 1999. The change in the coverage of broad categories of land use from 1971 to 1999 can be

[23] Small differences arise between these two addresses even when they are identical. This is generally due to abbreviating words such as "street" (str.) or "avenue" (ave.) in one of the addresses but not the other.

[24] The number of observations varies by attribute because not all municipalities gather information on every attribute.

Table 1.2: Summary Statistics

	Mean	SD	Min	Max	N
		Panel A: Tax Parcels			
Tax Assessment Data					
Building Value ($'000)	202.26	187.16	0.10	15,931.00	1,192,943
Land Value ($'000)	186.56	208.71	0.10	23,337.90	1,192,943
Other Value ($'000)	5.74	18.20	0.00	1,965.80	1,192,943
Total Value ($'000)	394.57	349.69	4.50	27,573.40	1,192,943
Other Parcel Characteristics					
Lot Size (m^2)	3,737.40	13,894.39	37.23	4,318,363.84	1,192,943
Year Built	1,955.96	38.31	1,800.00	2,018.00	1,192,943
Build Area (m^2)	254.64	148.15	50.07	6,611.17	1,192,943
No. of Rooms	6.78	7.99	1.00	8,020.00	1,099,449
Owner Occupied	0.84	0.36	0.00	1.00	1,192,827
Last Sale Year	2,000.22	15.05	1,900.00	2,019.00	1,192,943
Last Sale Price ($'000, 2018 prices)	482.63	416.02	50.01	18,508.42	713,602
Geographic Characteristics					
Dist. to Boston (km)	47.56	38.01	0.00	190.45	1,192,943
Dist. to Coast (km)	26.94	37.46	0.00	180.50	1,192,943
Dist. to Nearest Border (km)	1.45	1.31	0.00	13.52	1,192,943
		Panel B: Buildings			
Building Height (m)	11.53	5.17	3.00	25.00	1,302,984
Dist. to Nearest Border (km)	1.54	1.37	0.00	13.56	1,302,984
		Panel C: Towns			
Regulation Index					
NALPZ[1]	-0.03	0.91	-2.33	2.33	303
2010 Census Characteristics					
Population ('000)	17.88	22.79	0.17	181.47	303
Housing Units ('000)	7.61	9.57	0.11	74.64	303
Housing Density (units/km^2)	217.40	392.59	2.48	3,147.92	303
Rural (%)	32.43	37.61	0.00	100.00	303
Female (%)	51.70	2.07	39.81	60.17	303
Under 18 Years (%)	22.13	4.04	6.83	31.98	303
Over 64 Years (%)	15.00	4.44	7.89	39.80	303
Non-White (%)	7.43	9.06	0.45	56.87	303
Married (%)	68.59	7.12	26.16	83.27	303
School Quality Characteristics					
Expenditure Per Pupil ($'000)	15.90	4.85	9.52	59.81	296
Graduation Rate (%)	92.79	5.99	65.50	100.00	291
Municipal Taxes					
Residential Property Tax Rate (%)	15.24	3.85	2.75	24.34	303

Notes: Last sale price data drops observations for which the price is under $50,000, as it is most likely reflects the transfer of property rather than the true market price. Number of tax parcels and buildings differs as some properties have multiple buildings and some buildings are missing height data.

[1] Natural Language Processing Zoning Stringency Index

seen in Figure A.17. It shows the change in land use within 1km of municipal borders. Motivated by the fact that a significant amount of forest and agriculture land has been converted to residential use from 1971 to 1999, I use the share of developed land as an outcome of interest. This enables me to speak to the impact of LUR on conversion of land use and the spatial class of restrictions. I also calculate the share of developed land used for residential purposes, to test whether LUR influence the spatial pattern of housing development.

US Census Data

Census Blocks: Demographic and housing attribute data is gathered from the US Census Bureau at the block level—the smallest level of aggregation available—for the years 1970, 1990, 2000, and 2010. Data on demographics includes share of individuals under 18, over 64, non-white, married, female. Housing data consists of the average number of rooms, average rent, and average house price. This data is used to test whether the demographic composition of adjacent neighbourhoods in different towns are similar or not (pre-Massachusetts Zoning Act and widespread LUR). Thus, I only keep blocks that i) are in a town with a NALPZ measurement, ii) are neighbouring a town that also has a NALPZ value, and iii) are physically adjacent to a neighbouring town. This means that they share a border with a neighbouring town. I do this as my main empirical strategy uses houses in close to municipal borders. The assumptions regarding similar pre-treatment demographics thus must also hold at this level.

Census Block Groups: Additional census data is gathered at the block group level, which consists of a collection of blocks, which is used to create unobserved amenity values in Section 1.7. From the 2010 census I get the block group population.

American Community Survey: More block group level data in 2010 comes the ACS. This includes median household income and median housing prices.

Open Source Routing Machine/Open Street Maps

Bilateral travel times between all block group pairs is also required for the amenity estimations. These are calculated from Open Source Routing Machine (OSRM). Concretely, I generate the population-weighted centroids for each block group and calculate the bilateral commuting times between every pair using OSRM, which utilizes OpenStreetMap route networks.

School Districts

I estimate several specifications which control for one or more measures of school district quality. The first comes from Niche.com. They are an organization that provides information and rankings on neighbourhoods, schools, universities, and workplaces across the US. Importantly for this paper, they rank each school district in Massachusetts based on a variety of criteria, such as grades, parent/student surveys, and facility quality.

Additional measures of school (district) quality come from ClearGov.com. They aim to provide clarity to citizens on how tax revenues are spent within different communities. Notably, they also publish statistics at the school district level. From here I get measures of spending per pupil as well as graduations rates.

Since parents choosing a house based on school district quality will have access to the same, publicly available information, I am able to condition on the same information set as would be available these parents.

Finally, there are smaller towns which belong to a unified school district containing several other towns. This allows me to compare towns *within* some school districts with different degrees of regulatory intensities, by including school district fixed effects.[25] This specification is quite demanding of the data as this reduces the sample size significantly.

Municipal Property Tax Rates

In another specification, I also control for municipal property taxes. Should housing characteristics change in response to higher or lower property taxes, which is in turn systematically related to LUR, controlling for the rate would rectify this.

This data is gathered from the Division of Local Services of the Massachusetts Department of Revenue. I use the tax rate from 2018. This is the same year most of the tax assessment data was gathered.

1.5 Empirical Strategy

My baseline empirical method is a spatial regression discontinuity design (RDD). I use this to estimate the effect of more stringent land use regulation on the development and characteristics of single-family houses.

[25] It should be noted that because only smaller towns share school districts with neighbouring towns, the sample that allows for the identification of the effect of LUR in this specification is not representative of all towns in Massachusetts.

To implement this method, I need two key pieces of information: i) the nearest neighbouring town for each house (and by extension the nearest border), and ii) the distance to that town. As the tax parcels are all precisely geotagged, I am able to calculate this information from the tax database when combined with the towns shapefile from MassGIS.

With this information I create a "segment" identifier. This will be used in the spatial RDD framework to ensure that I am comparing tax parcels with other parcels in the nearest neighbouring town. In other words, I am comparing parcels that share a municipal border and are arguably in the same neighbourhood (especially when considering smaller bandwidths).

The baseline specification for the spatial RDD is as follows:

$$y_{ism} = \beta \text{NALPZ}_m + f(\text{geography})_{ism} + \pi_s + u_{ism} \tag{1.8}$$

where y_{ism} is the outcome of interest for parcel i, bordering segment s, in municipality m. NALPZ_m is the standardized Natural Language Processing Zoning Stringency Index, π_s are segment fixed effects, and $f(\text{geography})_{ism}$ is a function of the geographic location, the running variable. As the treatment—the level of regulation—varies at the town level, I estimate cluster-robust standard errors at the town level.

The main coefficient of interest is β. This captures the effect of a one standard deviation increase in the NALPZ on the considered outcome (y_{ism}). The main outcomes at the tax parcel level are the year the house was build, year last sold, and the size of the building and of the lot; from the building data the outcome of interest is the height of the building.

Following Dell (2010) and Dell et al. (2018), I begin by modelling $f(\text{geography})_{ist}$ with the latitude and longitude of each parcel, additionally controlling for the distance to Boston and to the nearest coast. As high-order polynomials have been shown to be unstable in RDD settings (Gelman and Imbens, 2018), I restrict my sample to a particular bandwidth around each town border and run a kernel regression (with respect to the running variable) with triangular weights. This involves modelling the running variable flexibly, allowing the effect of the distance to the nearest border to vary differently to each side of a border segment. Putting this all together results in:

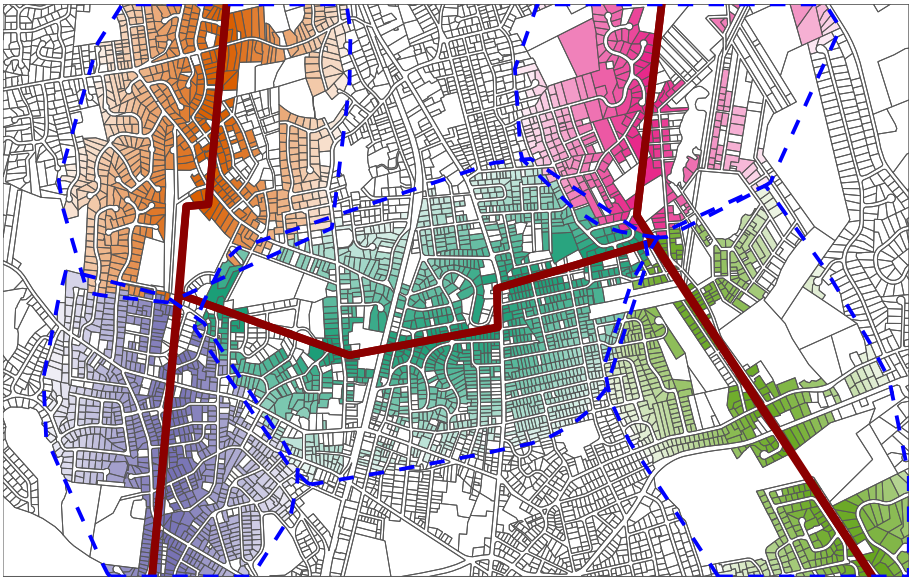
$$\begin{aligned} f(\text{geography})_{ism} =& \alpha_{sm}\text{distance to segment}_{ism} + \delta_1 \text{lat}_{ism} + \delta_2 \text{long}_{ism} \\ &+ \delta_3 \text{distance to boston}_{ism} + \delta_4 \text{distance to coast}_{ism} \end{aligned} \tag{1.9}$$

where α_{sm} are the town-border segment specific coefficients on the running variable. In practice, I find that after conditioning on the segment fixed effects (π_s) and the town-

border segment specific running variable, the additional geographic controls do not influence the results much. Thus, they are omitted from the reported results.

In the main results I show coefficient estimates with different bandwidths around town borders. This highlights the sensitivity of the estimates to including or excluding more observations further way from these borders. An example of this empirical strategy with a bandwidth of 500m is illustrated in Figure 1.4.

Figure 1.4: Parcel Spatial RDD Example



Notes: Example of the spatial RDD strategy with a bandwidth of 500m. Colours correspond to the matched town border. Red lines indicate town borders. Colour intensity indicates weight value. Uncoloured parcels are either not in the sample (*eg* not residential parcels) or outside of the bandwidth.

The empirical strategy here is similar to the one used by Turner et al. (2014) and Severen and Plantinga (2018), though I allow the distance gradient of the outcome variable to vary by town-border segment.

Aggregate Spatial Regressions

When investigating the impact of stringent LUR on the density of housing supply or on land use, where the unit of observation is a region or an aggregation of the tax parcel data, I estimate a simpler spatial regression. Here, the specification controls for segment fixed-effects to ensure comparisons are done between spatially near areas. Just as in the baseline specification, various bandwidths are considered. This results in the following estimating equation:

$$y_{sm} = \beta \text{NALPZ}_m + \pi_s + u_{sm} \tag{1.10}$$

where the indices are the same as the main specification. y_{sm} is a summary measure for some outcome in segment s, municipality m. This is an aggregate version of the baseline empirical model. The unit of observation is now a town-segment area rather than tax parcels *within* these geographical areas.

In the case where the outcome is housing supply, y_{sm} corresponds to either i) the number of single-family homes within the considered bandwidth, or ii) the density of houses per square kilometre within the bandwidth. When the outcome is land use, y_{sm} corresponds to either i) the share of land developed for residential use from 1971–1999, or ii) the fractional of developed land that is residential.

House Price Regressions

To test whether LUR affect house prices as well, I estimate a model similar to Equation 1.8, but I additionally control for lot sizes, and in some specifications for building sizes, for comparability. This results in the following specification:

$$\begin{aligned} y_{ism} =& \beta \text{NALPZ}_m + \alpha_{sm} \text{distance to segment}_{ism} \\ &+ \gamma_1 \text{building size}_{ism} + \gamma_2 \text{lot size}_{ism} + \pi_s + u_{ism} \end{aligned} \tag{1.11}$$

where γ_1 and γ_2 are the price effects for building size and lot size respectively. Here, the outcomes of interest are the price the house was last sold for and the total tax assessed value of the property (building and land together). I also present results where the outcomes are the assessed value of the building per square meter of building and the assessed value of the land per square meter of land.[26]

Identification Strategy

The empirical specifications above both exploit the same variation. Namely, the discrete change in the regulatory environment at the border between two municipalities. The primary assumption of this strategy is that no other observed or unobserved feature that varies at municipal borders affects housing development pattern *and* is systematically related to the stringency of LUR.

The main specification addresses several potential issues by only considering units that are geographically close. This controls for potential geographic confounders like local housing demand, access to well paying jobs, and preferences for certain regions.

In Section 1.7 I address other potential issues as follows. First, I test whether there existed differences in demographic characteristics in neighbouring blocks in different

[26] These two specifications do not add controls for the size of the building or lot as they are already normalized.

towns *before* the Massachusetts Zoning Act and widespread LUR. I supplement this by estimating amenity values by calibrating a standard urban spatial general equilibrium model in the spirit of Ahlfeldt et al. (2015), and testing whether these vary discretely at town borders systematically with the level of LUR. Second, I investigate whether my results are sensitive to alternative specifications of the running variable and weighting method. Third, I directly control for municipal characteristics, like school quality and property tax levels, that vary at town borders along with the regulatory environment.

Graphical Evidence

On account of there being multiple cut-offs (town border-segments) with a non-binary treatment variable (NALPZ index), it is not straightforward to graphically inspect the discontinuity. To nonetheless present suggestive evidence I plot the primary outcome variables residualized by border-segment fixed effects, binned into 100m intervals, by whether the housing unit belongs to the less or more regulated side of its matched border-segment.[27] The results are shown in Figure A.20.

Two things stand out. First, there is a significant discontinuity for some outcomes, such as for lot sizes and tax assessed land values. Second, the gradients for the outcome variables are not uniform. For example, there is a discontinuity for year built, but the distance to the border gradient appears quite similar. Looking at the house prices outcomes on the other hand, the gradients are quite different and in fact go in different directions. In less regulated towns the more expensive housing is located at the border region of a town whereas in more regulated towns the housing becomes more expensive as you approach the town centre. This fact highlights the importance of allowing the outcome gradients to vary by town border-segment.

1.6 Results

The results are broken down as follows. First, I present evidence for how the aggregate housing supply near the municipal borders is affected in towns with more restrictive development policies. This speaks to whether restrictive land use policies result in lower density housing overall. Second, I highlight results from my baseline spatial RDD specification, looking at various housing development outcomes while considering different bandwidths around the borders. This consists of two primary groups of outcomes: i) housing market outcomes (the age of the structural as well as the year the house was

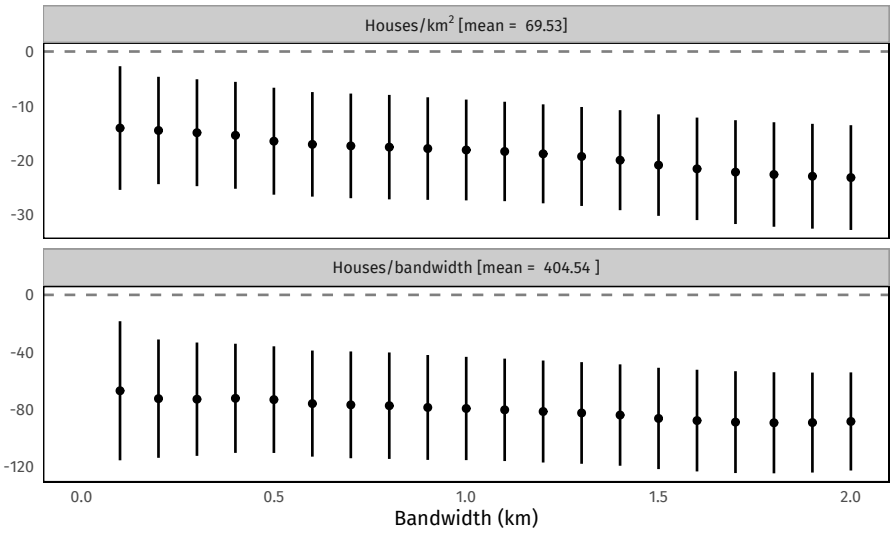
[27] As towns mostly have multiple neighbours, it may both be considered "less regulated" and "more regulated" but each housing unit can only be in one category on account of it being matched exclusively to one border-segment.

last sold), and ii) housing characteristics (building size, building height, and lot size). Third, I test for the role of LUR in altering land use by showing results for the effect of stringent regulations on the conversion of undeveloped land to residential use as well as on the fraction of developed land used for residential purposes. Fourth, I investigate whether stringent LUR are capitalized into house prices at the local, neighbourhood level.

Overall Housing Supply

To get an overview of how stringent LUR affect the housing supply, I plot the results of estimating Equation 1.10, which captures the effect of these regulations on the number of single-family homes. I consider two outcome measures for overall housing supply: i) the density of houses per square kilometre, and ii) the raw count of houses within a specified distance to the border. The results of these regressions are shown in Figure 1.5. I estimate the model for bandwidths between 100m and 2km, at 100m intervals. The coefficients are interpreted as the respective change in the outcome variable for a standard deviation increase in the NALPZ.

Figure 1.5: Spatial Regression of Housing Supply and Density on NALPZ



Notes: β coefficient from Equation 1.10 for various bandwidths plotted. Respective outcome is regressed on Natural Language Processing Zoning Stringency Index NALPZ and border segment fixed effects. Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level. Mean outcome refers to the average outcome value with a bandwidth of 100m.

As can be seen from the figure, the results are remarkably stable across the entire range of bandwidth options. They indicate (at a bandwidth of 100m) that the housing density is 14 houses per square kilometre lower and that the average number of houses within 1km of the border is 67 units less for every s.d. increase of regulatory stringency.

These estimates correspond to an effect size of about 20% and 17% of the mean outcome level, respectively.

This provides evidence that these restrictive land use policies have led to a reduction in the housing supply, at least at the peripheries of these towns, by reducing the density of development.

Housing Market

I now turn to the results of the baseline spatial RDD. I have plotted the results for each outcome individually, considering various bandwidths for sample selection. The plots are given in Figure 1.6. I show the results for bandwidths varying from 100m to 2km. Each point represents the β estimate, along with its 95% confidence interval from clustering the standard errors at the town level, from Equation 1.8 for the tax parcel and building level regressions, and from Equation 1.10 for the land use regressions. NALPZ is standardized, so the coefficients correspond the change in the respective outcome variable to a standard deviation increase in regulatory stringency.

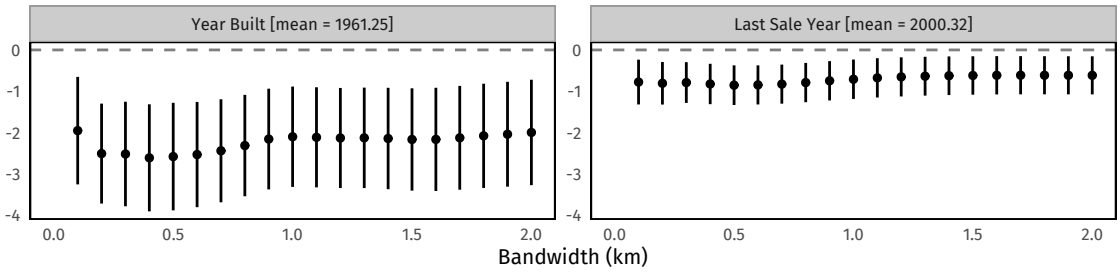
The results in Figure 1.6a speak to the effect of LUR on the housing market. The rate of development of new housing within a neighbourhood should be similar if the regulatory environment is the same. The results for this outcome, shown in the left plot, indicate that houses in more regulated towns are older on average. At the smallest bandwidth, it suggests that a one standard deviation increase in NALPZ corresponds to homes being two years older on average. As the average house age in sample is 57 years, this represents 3.4% of the mean.

The right side of the panel shows results with the year the house was last sold as the outcome. This speaks to a related question of whether LUR reduce the efficiency of the housing development and the real estate market (Mayer and Somerville, 2000; Glaeser et al., 2008). For example, if LUR make the housing supply less elastic to changes in housing demand, the selling and buying of the current housing stock may take place less often as incumbent residents hold on to their homes. Overall, there is not a major difference in the last time a home was sold with respect to LUR. At the smallest bandwidth, the results suggest a one standard deviation increase in NALPZ corresponds to houses last being sold about 0.8 years further in the past on average. This is an effect size of roughly 4.3% of the mean.

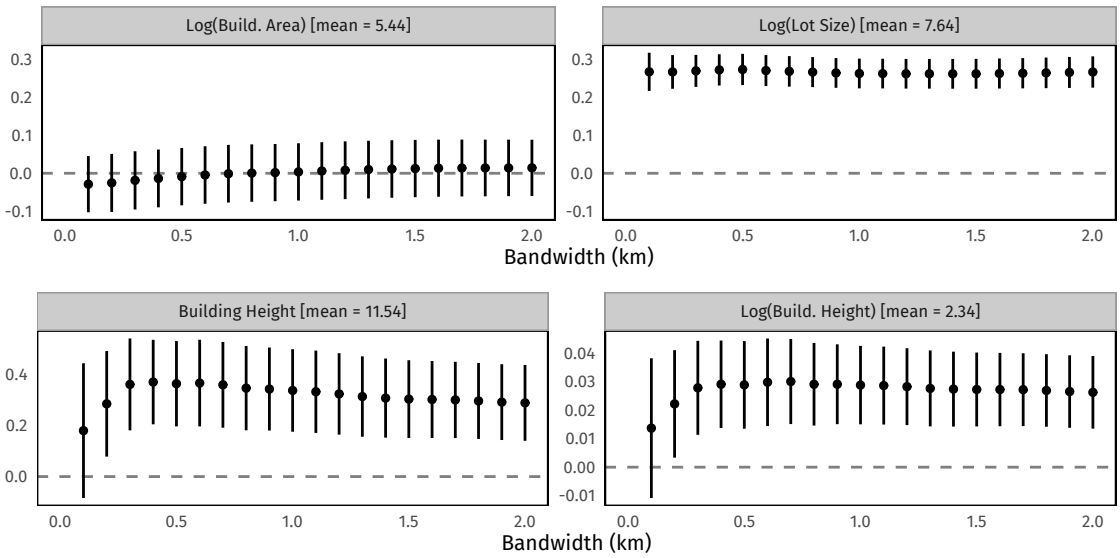
Taken together the results provide some evidence for LUR resulting in a less responsive housing market, though the effects are economically small.

Figure 1.6: Main Results

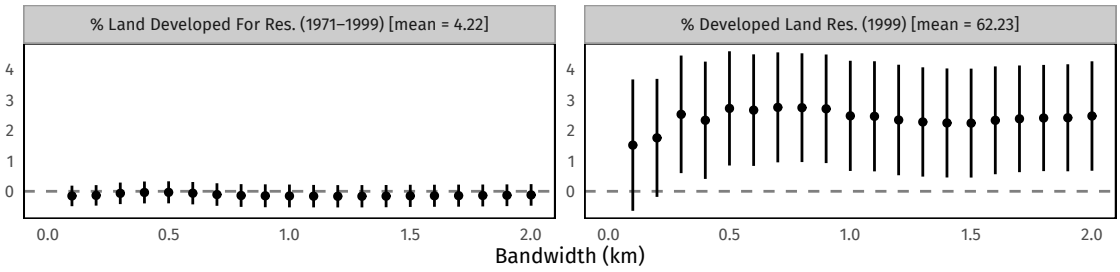
(a) Housing Market Outcomes: Spatial RDD of Respective Outcome on NALPZ



(b) Housing Attributes Outcomes: Spatial RDD of Respective Outcome on NALPZ



(c) Land Use Outcomes: Spatial Regression of Respective Outcome on NALPZ



Notes: Panel (a) and (b): plots the estimated β coefficient from Equation 1.8. Respective outcome is regressed on the Natural Language Processing Zoning Stringency Index (NALPZ), border segment fixed effects, and border segment-specific distance controls. Panel (c): plots the estimated β coefficient from Equation 1.10. Respective outcome is regressed on the NALPZ and border segment fixed effects. Both: Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level. Mean outcome refers to the average outcome value with a bandwidth of 100m.

House Attributes

The results for the housing attribute outcomes are shown in Figure 1.6b. As seen in the upper-left plot, there is virtually no difference between the size of livable space in houses in more or less strongly regulated municipalities in the same neighbourhood. Regardless of the bandwidth considered, no coefficient is significantly different than zero, with the point estimate very close to zero, relatively precisely estimated. Turning to the upper-right plot, the lot sizes, on the hand, are strongly influenced by restrictive zoning policies. Considering the most narrow bandwidth, a standard deviation increase in land use restrictiveness results in lot sizes being roughly 27% larger.

This result provides evidence for a specific channel that leads to lower density development overall: stringent LUR increases the amount of land used per house, resulting in lower density and fewer houses. An important follow-up question, that will be addressed when looking at how land use has changed with respect to regulatory stringency, is whether municipalities compensate for the increase in land *per* lot by allocating more land overall to residential use.

The bottom two plots presents results for building height as the outcome. As there is no guidance from the literature on the functional relationship between land use regulation and building height, I present results with the height in level and logarithmic terms. However, regardless of the specification there is no strong relationship between the stringency of LUR and building height. Though most specifications are significantly different than zero, the results are precise enough to rule out a standard deviation increase in NALPZ increasing building height by 5% in the log-level specification. With an average house height of 11.5m, this is an economically insignificant result.

Considering all the results together, development in towns with stricter LUR tends to be on larger parcels of land, without correspondingly larger or taller buildings.

Land Use

Finally, the results testing for land use regulations that spatially restrict development are highlighted in Figure 1.6c. They plot the relationship between land use (development) patterns and NALPZ.

The left plot presents the effect of stringent LUR on the conversion of undeveloped land in 1971 to residential use in 1999. This tests whether restrictive development policies impact the rate of land being developed. Regardless of the bandwidth the estimated coefficients are essentially zero with the 95% confidence intervals also bound very close around null. This suggests that the rate of conversion of undeveloped land to residential use was not differential between municipalities with varying degrees of LUR intensity.

The results in the right plot tell a similar story. The estimated effect of restrictive LUR on the share of residential land of all developed land is marginally significant for all but the two smallest bandwidths, with effect sizes around 2.4% of the mean.

Two conclusions can be drawn from these results. First, stringent LUR do not appear to alter the amount of land allocated to residential purposes. Second, stringent municipalities do not develop more land for residential use to compensate for the lower housing density due to larger lot sizes.

Spatial RDD by Density Group

Motivated by the the large effects of LUR on parcel lot sizes, I further investigate how the distribution of lot sizes changes with respect to the different quartiles of the NALPZ index, shown in Figure A.18. As is evident, there is a significant degree of heterogeneity in the distribution of lot sizes over the NALPZ quartiles. To aid in interpreting the distribution of lot sizes, the cut-offs between high, mid, and low density lot sizes are shown with the dashed vertical lines, as defined by MassGIS.

To better understand how LUR impact the lot sizes *within* the density groupings, I re-estimate the baseline spatial RDD empirical model for each density subsample with lot size as the outcome. The results are shown in Figure A.19.

A word of caution interpreting these results. Most likely, LUR result in more development happening in one density category rather than another (*eg* more regulated municipalities may encourage the building of more low-density homes). Therefore, these should be considered descriptive rather than causal. What is immediately clear, however, is that conditional on density grouping, the effect of more stringent LUR is strongest amongst low-density housing. This suggests that the results are being driven by larger lots becoming even larger.

House Prices

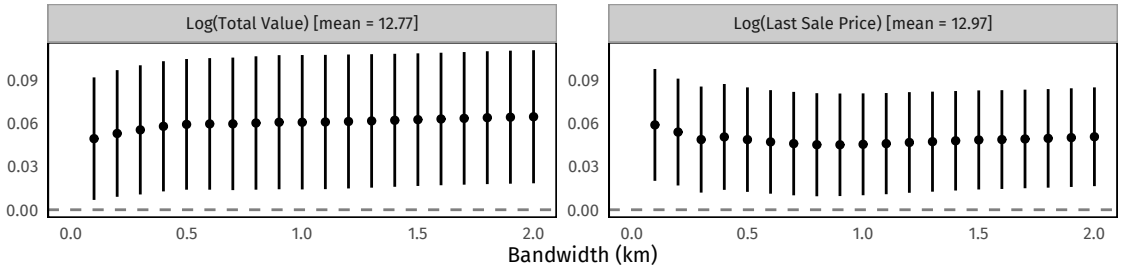
Figure 1.7 shows the results of the house price regressions. These results speak to the question of whether more stringent regulation is reflected in the house prices, which depends on the degree of substitutability between neighbouring houses.

The top panel (1.7a) displays the results when the outcome is the total tax assessed value of the property or the last sale price of the house. The estimated effect of LUR on house prices is quite similar regardless of the metric used to measure house prices: a standard deviation increase in the NALPZ increases house prices roughly 5–6%.

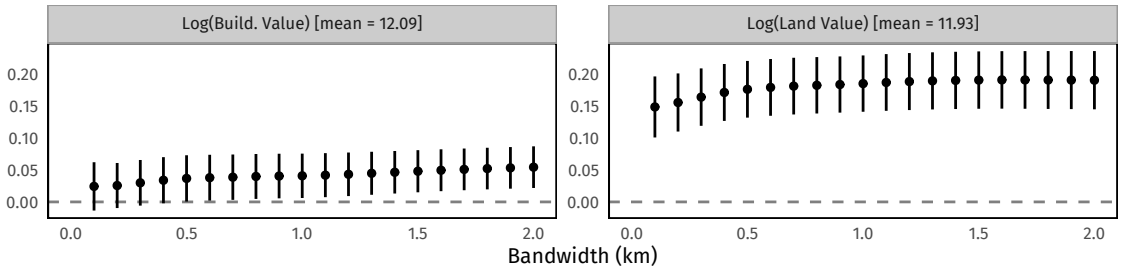
However, these regressions do not identify the impact of LUR separately for land and building prices. Thus, I leverage the fact that the tax assessment values are broken down

Figure 1.7: House Price Outcomes: Spatial RDD of Respective Outcome on NALPZ

(a) Logarithm of Total Assessed Value and of Last Sale Price (2018 prices)



(b) Assessed Value of Building per m^2 and of Land per m^2



Notes: Plots the estimated β coefficient from Equation 1.11. Respective outcome is regressed on the Natural Language Processing Zoning Stringency Index (NALPZ), border segment fixed effects, and border segment-specific distance controls. Panel (a) additionally controls for building size and lot size. Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level. Mean outcome refers to the average outcome value with a bandwidth of 100m.

into building and land components. I then use these outcomes, normalized by the size of the building and lot respectively, in place of the total house prices. The results for these regressions are given in Figure 1.7b. Unlike with the total house prices, here the evidence for stringent LUR being capitalized into house values is weaker. The results for both the building and lot values are not significantly different from zero across all the bandwidths.

Together, these results provide evidence for houses in neighbouring towns being substitutes for one another. Thus restrictive building policies in one town do not necessarily lead to higher prices overall if there is sufficient housing in neighbouring towns. In other words, population mobility may arbitrage away price differentials *across* municipalities within a region, but increase the price level of the region as a whole.

1.7 Balancing, Specification, and Robustness Checks

Having established a strong relationship between land use regulations and lower density development, primarily via larger lot sizes, I now turn to investigating whether these results could be driven by other factors. First, I present a pair of balancing checks, where

I test whether the Natural Language Processing Zoning Stringency Index can predict neighbourhood demographic characteristics in 1970, before LUR were commonplace, and whether the NALPZ is related to unobserved neighbourhood amenities. I recover these unobserved neighbourhood amenities by calibrating an urban spatial general equilibrium model. Second, I conduct a series of specification and robustness checks to test whether different sources of variation may partially explain the results.

Neighbourhood Demographic Characteristics Pre-Massachusetts Zoning Act

One is typically concerned that the forcing variable in an RDD design is manipulated by the units of observation. As I am considering tax parcels of land, and the forcing variable is the distance to the town border, this is not problematic in my setting. However, there is still the concern that towns which implemented stronger/weaker regulations varied significantly, even within a neighbourhood that belongs to two or more towns.

Since the tax assessors database is constantly updated, I am unable to obtain data before and after the introduction of municipal-level LUR. Therefore, to investigate whether there were any pre-existing differences before the Massachusetts Zoning Act, I compare the demographic composition of US Census Blocks directly to either side of the borders (*ie* blocks touching the border). I test whether *future* regulatory intensity predicts these demographic outcomes in the past. I use data from the 1970, 1990, 2000 and 2010 censuses and run a modified version of Equation 1.8:

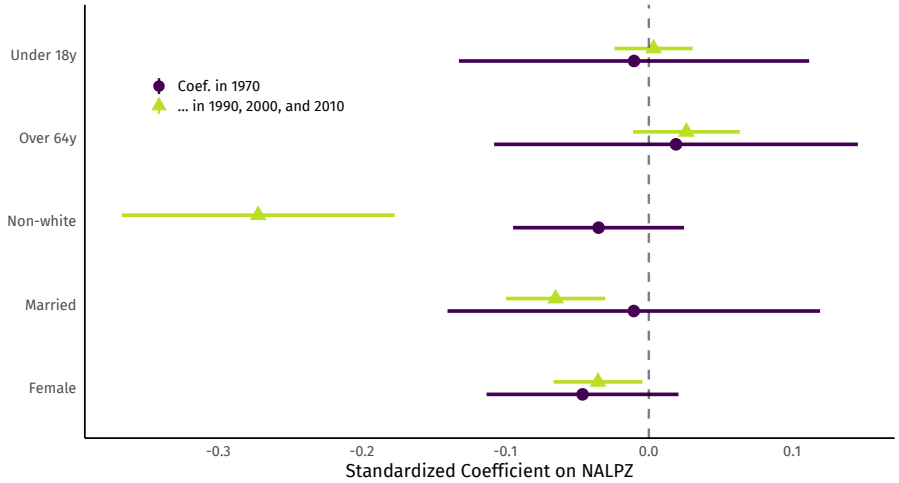
$$y_{bsmt} = \beta_1 \text{NALPZ}_m + \beta_2 \text{NALPZ}_m \times \{year_t > 1970\} + \pi_{st} + \alpha_s \text{wdist}_{bsm} + u_{bsmt} \quad (1.12)$$

where the outcomes are now indexed by time, π_{st} are year-specific segment fixed effects, and I no longer need to control for geographic distance (as the distance to the border is zero by construction). However, I do control for how far the average cell of a block group is away from the comparison border. This is the wdist_{bsm} term. It refers to the block grid-cell average distance to the border. I allow its effect to vary at the border-segment level (α_s). This controls for segment-specific gradients in demographics characteristics. For example, larger block tend to cover space further away from a municipal border (thus having higher wdist_{bsm} values), making them less representative of the area immediately surrounding municipal borders. The unit of observation, b, is now the census block. The coefficient of interest is β_1: if there are no pre-existing differences in demographics and housing, it should be zero.

β_2 may be non-zero if the implementation of LUR changed the demographic composition through modified residential development and residential sorting. This is an interesting outcome in and of itself. If, for example, existing residents lobbied for more restrictive land use policies to stem the flow of in-migration to the municipality, barring individuals from certain demographics disproportionately from moving to the town, that would be captured by β_2.

The coefficients β_1 and β_2 are plotted in Figure 1.8. The purple circles represent β_1 and the green triangles β_2. For comparability across the various outcomes, I report standardized coefficients. As can be seen in the figure, the estimated β_1 coefficients are not statistically different from zero. None of the estimated effects have standardized coefficient estimates over 0.05 in absolute terms. In addition the 95% confidence intervals reject any absolute effect size above 0.15 standard deviations.

Figure 1.8: Demographic, Housing Characteristics and Land Use Regulation



Notes: Plots the estimated β_1 and β_2 values from estimating Equation 1.12. Respective demographic or housing characteristic is regressed on the NALPZ interacted with 1970 and post-1970 dummies, and segment by year fixed effects. β_1 captures the ability of the NALPZ to predict census block demographic characteristics in 1970. β_2 captures the ability of the NALPZ to predict census block demographic characteristics after 1970 (1990, 2000, 2010). Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level.

Looking at the same demographic composition after the introduction of the Massachusetts Zoning Act, I find that the share of non-white residents is significantly lower in highly regulated towns. Taken literally, it implies that a one standard deviation increase in NALPZ corresponds to roughly a 0.24 standard deviation decrease in the share of the population that is non-white. This provides evidence for the theory that restrictive zoning policies have resulted in residential sorting, either deliberately or as a side effect.

Amenities

Given the observable data from the census pre-Massachusetts Zoning Act, I am able to test for demographic differences across town borders that may confound my estimates. However, there may still be unobservable differences in local amenities that may induce sorting by preferences for housing with specific characteristics, or for low-density neighbourhoods. Furthermore, as I am looking at small geographic regions, simply controlling for *some* observable amenities (*eg* parks, waterfront locations, access to playgrounds, etc.) does not varying the empirical estimates much.

To address this concern, I estimate local amenities as implied by the canonical quantitative spatial equilibrium model of a city (Ahlfeldt et al., 2015) at the census block group level (a small collection of city blocks). To calibrate the model, I use the parameter values from several recent papers that have structurally estimated the model: Ahlfeldt et al. (2015), Tsivanidis (2018), and Heblich et al. (2020). I then test whether these amenity values vary in adjacent census block groups to either side of municipal borders.

Model Intuition

For the purposes of estimating implied amenities, it is sufficient to discuss and parameterize the worker demand system. Details on the model are given in Appendix Section A5. Intuitively, workers trade-off wages, commuting costs, housing costs, and local amenities when choosing residence (i) and workplace (j) locations.[28] Workers are heterogeneous with regards to residence-workplace location pairs.

Solving the model results in the following expression for residential amenities:

$$\frac{B_i^*}{\widetilde{B_i^*}} = \left(\frac{H_i}{\widetilde{H_i}}\right)^{1/\epsilon} \left(\frac{q_i}{\widetilde{q_i}}\right)^{1-\beta} \left(\frac{\text{CMA}_i}{\widetilde{\text{CMA}_i}}\right)^{-1/\epsilon} \tag{1.13}$$

where B_i^* are residential block-specific amenities, H_i refers to block resident population, q_i is the price of housing, and $\text{CMA}_i = \sum_{j=1}^{S} \left(w_j / e^{\kappa \tau_{ij}}\right)^{\epsilon}$ is a measure Commuting Market Access, which captures how closely in terms of commuting time (τ_{ij}) a residential block is located is to well paying jobs (w_j).[29] $\widetilde{X} = \left(\prod_i^N X_i\right)^{1/N}$ denotes the geographic mean of the respective variables, which is included to remove terms that are invariant across residential locations. I use $B_i^*/\widetilde{B_i^*}$ as my measure of local amenities. As this term has no natural scale, I standardize it to ease interpretation of the results.

[28] Here I use the term "block" to refer to a location, as it corresponds nicely with the US Census geographic terminology as well as my empirical geographic unit.

[29] More details in Section A5

Intuitively, if a residential block has a high population in spite of high housing costs and poor access to well-paying employment opportunities, there must be high local amenities to compensate the residents.

In addition to the observable data $\{H_i, q_i, \tau_{ij}, w_j\}$, I require estimates of the models parameters $\{\epsilon, \beta, \kappa\}$ to back out implied amenities. I consider various parameter combinations from recent contributions to the literature to obtain several different measures of amenities. The exact parameter values are given in the Appendix, Table A.4.

Estimating Amenities

To estimate the amenity values I use data from the US Census Bureau at the block group level and the Open Source Routing Machine as described in Section 1.4.

I restrict my sample to census block groups that are adjacent to *one* municipal border (*ie* I drop block groups that border several towns or are in the interior). This is to ensure that the census block groups I am comparing across borders are as similar to one another as possible.

The estimation equation is similar to the one used to test for demographic balance in census blocks in 1970 (Eq. 1.12), but without the time dimension (I am only considering data from 2010):

$$\mathrm{y}_{bsm} = \beta_1 \textsc{nalpz}_m + \pi_s + \alpha_s \mathrm{wdist}_{bsm} + u_{bsm} \tag{1.14}$$

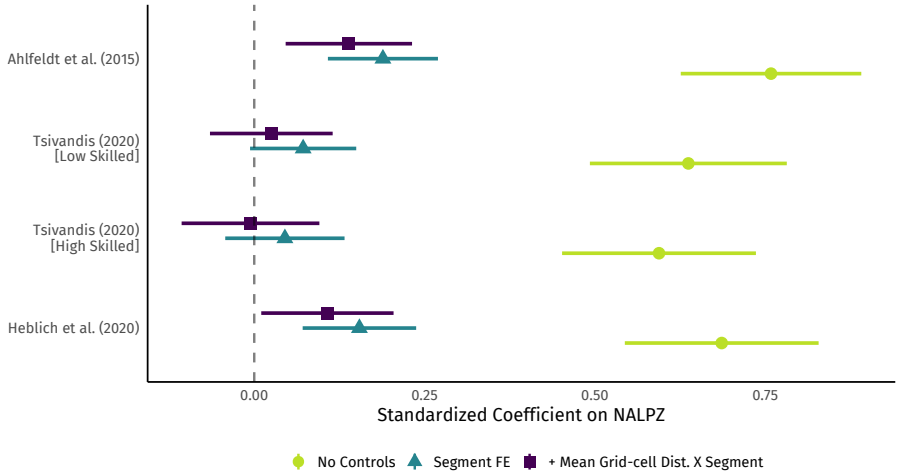
where $\textsc{nalpz}_m$ and π_s are defined the same as previously. y_{bsm} refers to the standardized measure of residential amenities ($B_i^* / \widetilde{B_i^*}$) as implied by the model under four different parameter combinations (see Table A.4). The wdist_{bsm} term is the same as used in Equation 1.12, but is calculated at the block group level. This controls for segment-specific gradients in amenities.

Amenities and Land Use Regulation

I first show results only including $\textsc{nalpz}_m$, to get an idea of the raw relationship between residential amenities and land use regulation. Then, I subsequently include segment fixed effects, and finally the segment-specific grid-cell average distances.

These are shown in Figure 1.9. The raw coefficients (green circles) show that there is a very strong relationship between land use regulations and implied amenities on average. This suggests that households are willing to pay higher housing costs and live further away from well-paying employment opportunities to live in communities that are more stringently regulated.

Figure 1.9: Local Amenities and Land Use Regulation



Notes: Plots the estimated β value from estimating Equation 1.14 under different sets of controls. Unobserved amenity values, calculated from an Ahlfeldt et al. (2015) spatial model under different parameter combinations, are regressed on the NALPZ and the noted controls. Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level.

However, this is comparing block groups in Massachusetts that are located far from one another. As highlighted in Figure 1.3, a large cluster of highly-regulated towns is located in the North-West section of Massachusetts, isolated from the major employment centres of Boston, Springfield, and Worcester, whereas the towns closer to these centres are more relaxed on average. There are significant differences between the households, and therefore the sort of housing demanded, that choose to live in these different regions.

Once I include segment fixed effects, and more so by additionally including segment-specific mean grid cell distance linear gradients, the estimated relationship between LUR and implied amenities drops significantly. In two of the four sets of parameter combinations (values for both high- and low-skilled workers from Tsivanidis (2018)), I cannot reject the null hypothesis that there is no relationship between LUR and amenities. The standardized effect sizes for the other two parameter combinations are never larger than 0.15. This provides evidence that the main empirical strategy allows me to compare units (houses, parcels, buildings) that are in similar neighbourhoods and have similar levels of local amenities.

Specification and Robustness Checks

To validate my main findings, I also consider alternative specifications and controls to explore the robustness of my results. The first change made is to rerun the baseline regression without the triangular weights (*ie* the weight given to units closer to and farther from municipal borders is the same). Next, I model the function of geography as

a quadratic in each term (but still allowing the effect of the running variable to differ by border-segment, also as a quadratic).

Next, I control for the presence of school districts. As school district borders generally align with municipal borders, it is not directly possible to control for the districts for every observation (parcel). To overcome this, I control for measures of school quality. These school quality measures are described in Section 1.4. I also estimate a model controlling for school district fixed effects, which identifies the effect of LUR in smaller municipalities that share a school district. This specification is quite demanding as many municipalities have their own school district, so the effective sample size reduces significantly.

To the extent that property taxes systematically relate to LUR as well as impact housing development, I control for the residential property tax rate in a further robustness specification. I then include all of school district quality and rank measures together with the property tax rate control.

There is also a concern that new development is mostly infill development: *ie* new houses are built on the sites of demolished previous houses and that this is more important in determining lot sizes than LUR.[30] Thought of another way, there may be path dependence when redeveloping land for new residential use. To address this issue, I re-estimate my primary specification on the subsample of houses that were built on previously undeveloped land, where a developer would only be restricted by LUR and geography. Specifically, I look at development taking place *after* 1971 on land that was previously vegetation (*eg* forest, bush) or agricultural.

The results of these robustness checks for the housing market, attributes, and price outcomes are shown in Figures 1.10a, 1.10b, and 1.11, respectively. The baseline specification, along with all the robustness checks, are with a bandwidth of 100m.

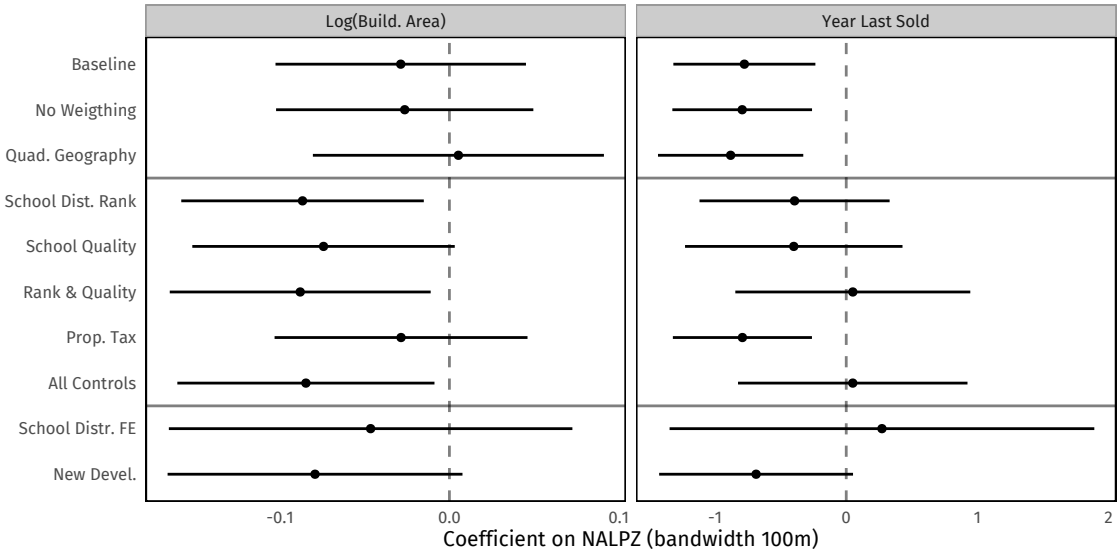
Looking at the housing market outcomes, the estimates for the age of the house are quite stable across the various specifications. The results change the most when conditioning on only new development post-1971. The small, but statistically significant, results confirm previous evidence that LUR have a modest but persistent effect on the rate of new development. The results for year last sold also do not vary much. The baseline estimates were already small, and become insignificant especially with the school district controls.

Turing to the housing attributes outcomes, lot and building size, again there is little change in the estimated coefficient on NALPZ. The effect of stringent LUR on lot sizes remains large and statistically significant across the estimated equations, except when

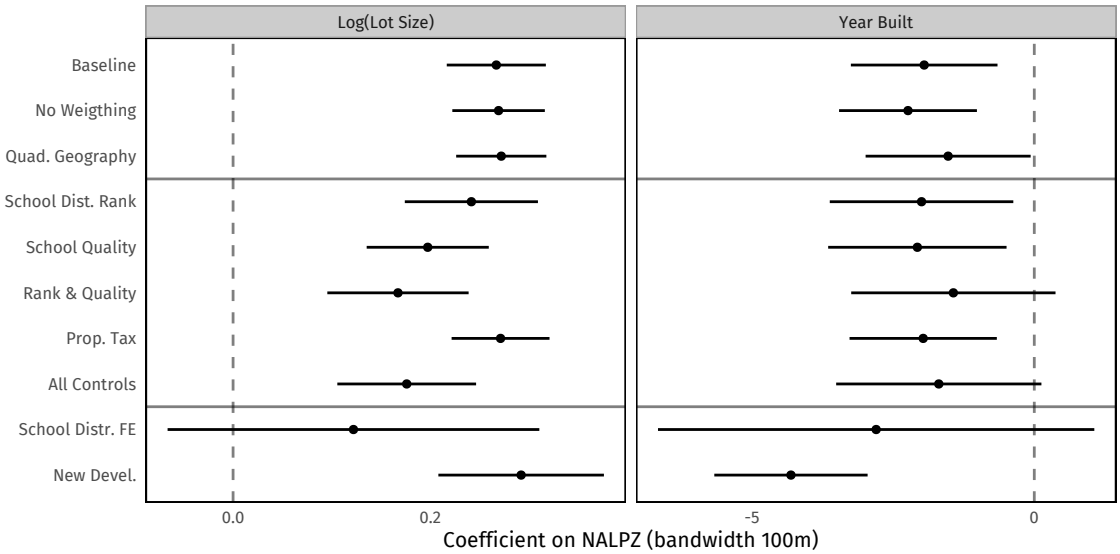
[30] Of course, this process would have to vary with the overall restrictiveness of LUR as well to be an issue.

Figure 1.10: Robustness Checks Baseline

(a) Housing Market Outcomes

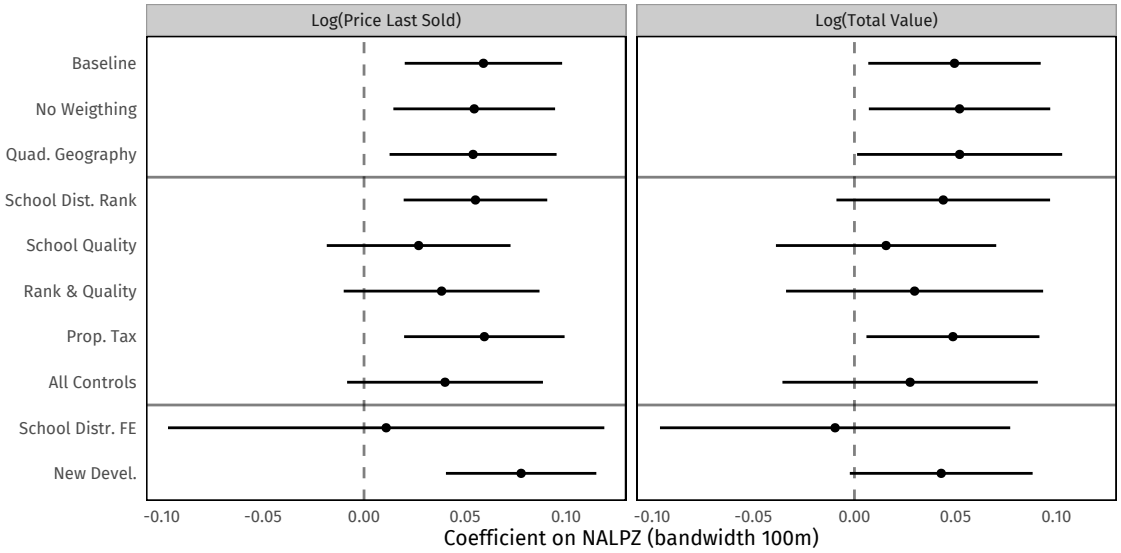


(b) Housing Attributes Outcomes



Notes: Plots the estimated β value from estimating different specifications of Equation 1.8. "No Weighting" removes triangular weights and assigns uniform weights. "Quad. Geography" adds a square term of the town border segment-specific distance controls, as well as quadratic controls for longitude, latitude, distance to coast, and distance to Boston. "School Dist. Rank" adds a quadratic for school district rank taken from Niche.com. "School Quality" adds per pupil expenditure and graduation rate controls from ClearGov.com. "Prop. Tax" controls for municipal property tax rates. "All Controls" includes all school district controls and the residential property tax rate. "School Distr. FE" controls for school district fixed effects. "New Devel." estimates the model on the subset of houses that were developed after 1971. Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level.

Figure 1.11: Robustness Checks: House Prices



Notes: Plots the estimated β value from estimating different specifications of Equation 1.11. "No Weighting" removes triangular weights and assigns uniform weights. "Quad. Geography" adds a square term of the town border segment-specific distance controls, as well as quadratic controls for longitude, latitude, distance to coast, and distance to Boston. "School Dist. Rank" adds a quadratic for school district rank taken from Niche.com. "School Quality" adds per pupil expenditure and graduation rate controls from ClearGov.com. "Prop. Tax" controls for municipal property tax rates. "All Controls" includes all school district controls and the residential property tax rate. "School Distr. FE" controls for school district fixed effects. "New Devel." estimates the model on the subset of houses that were developed after 1971. Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level.

including school district fixed effects, which results in imprecise estimates. Nonetheless, most of the 95% confidence interval lies above zero. The effects on building size remain close to zero. Overall, these results confirm the role LUR play in increasing the amount of land used per house.

Finally, the robustness results for the housing price outcomes are shown. Here, the estimated relationship between housing prices and LUR shrinks when school quality is controlled for. This suggests that the price differentials across municipalities with different regulatory environments is due to access to better schools, rather than LUR. This all supports the view that houses that are nearby but in different towns (with similar quality schools) are close substitutes.

1.8 Conclusion

Land use regulations are ubiquitous across the US, but their causes and impacts are not fully understood. I create a new regulation index, called the Natural Language Processing Zoning Stringency Index (NALPZ), by applying a machine learning algorithm, a Latent Dirichlet Allocation model, to over 40,000 pages of zoning bylaw documents.

This method builds off previous work to greatly increase spatial coverage. This lets me address the question of *how* stringent land use regulations are manifested in housing development.

By exploiting the variation in the regulatory environment at municipal borders in Massachusetts, I confirm previous studies showing that stringent land use regulations reduce housing density. Moreover, I extent these findings by showing that lot sizes are most responsive to LUR, being considerably larger in more regulated jurisdictions. This provides evidence for parcel-specific land use regulations—regulations that encourage higher land usage per house—being most responsible for restricting housing supply. These include regulations such as minimum lot sizes, setback requirements, strict floor-area-ratios.

Furthermore, the results suggest that spatially close houses in differing towns are highly substitutable. This provides evidence for restrictions in one locality leading to price increases in neighbouring towns as well, if overall housing supply is not responsive enough, as previous work has shown.

My findings suggest that land use regulations encourage less dense development on larger parcels of land, without the compensation of allocating more land to residential use overall, effectively limiting the supply of housing available. For policy makers looking to increase available housing in high-growth regions, scrutinizing these local constraints to development is a promising avenue.

Appendix A1 Natural Language Processing Details

Term Frequency-Inverse Document Frequency Weighting

tf-idf weighting assigns more weight to tokens that appear more often in fewer documents, under the presumption that these tokens are better able to discriminate between different documents. Conversely, tokens that appear seldomly in almost all documents do not tell us much. Formally, it is the product between a term frequency (tf) part, and an inverse document frequency (idf) part. There are different ways to measure both of them, and those used in this paper are described below.

The formula for the term frequency part is given as:

$$\text{tf}_{vd} = x_{vd} / \sum_{v \in \mathcal{V}} x_{vd} \tag{A.1}$$

where the term frequency for token v in document d depends on the count of that token in the document (x_{vd}) divided by the total length of the document for all tokens in $\mathcal{V}$.

The expression for the inverse document frequency is as follows:

$$\text{idf}_v = \log(D/df_v) \tag{A.2}$$

where D refers to the number of documents in the corpus. This term is term specific, but is the same for every document.

The product of these two terms is the tf-idf weighted token count:

$$\text{tf-idf}_{vd} = \text{tf}_{vd} \times \text{idf}_v \tag{A.3}$$

Dictionary Methods

Table A.1: Harvard IV-4 Category Dictionary Examples

Active (2045)	**Legal (192)**	**Place (318)**	**Region (61)**
damage	truant	range	municipality
investor	counsel	province	rural
aid	law	ridge	slum
argue	disputable	skyline	spot
protect	eye	arena	city

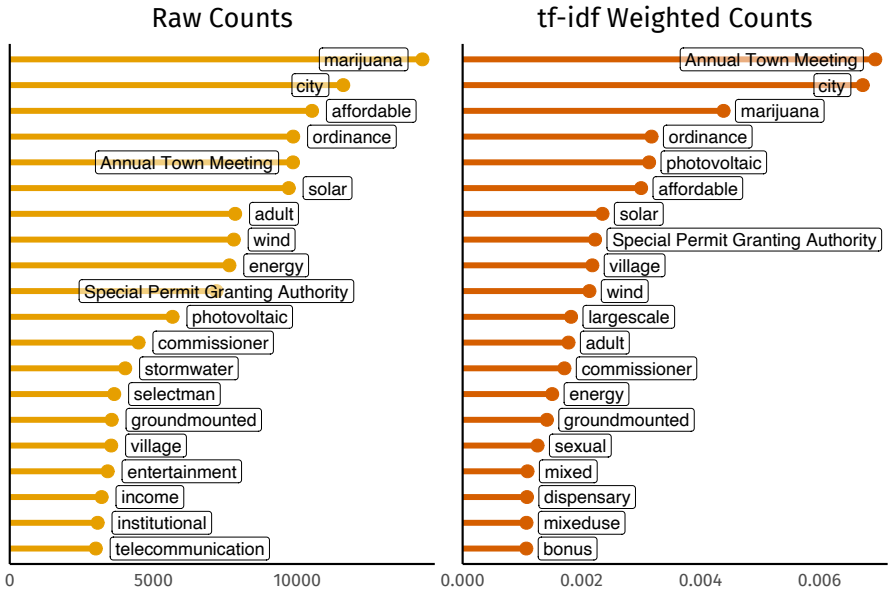
Aquatic (20)	**Land (63)**	**Building (46)**	**Object (104)**
wave	mainland	construction	constitution
breaker	cave	shutter	board
creek	scene	bedroom	paint
channel	border	bathroom	stamp
water	island	chamber	notice

Nature (61)		
maple	grass	manure
calcium	flower	

Notes: Each sub-table is a dictionary category used when investigating dictionary NLP methods. The number in parentheses indicates the total number of words belonging to that topic. Five words, chosen at random, are listed under the heading.

Figures

Figure A.1: Top Word Counts from Municipal Bylaws



Notes: TF-IDF weights are calculated corpus-wide rather than per document.

Figure A.2: Histogram of Raw and tf-idf Weighted Token Counts

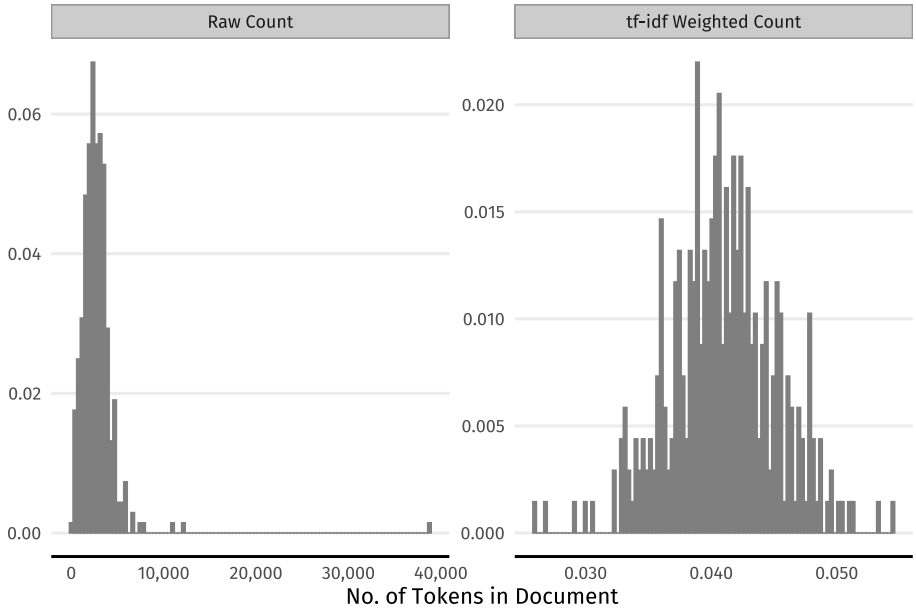
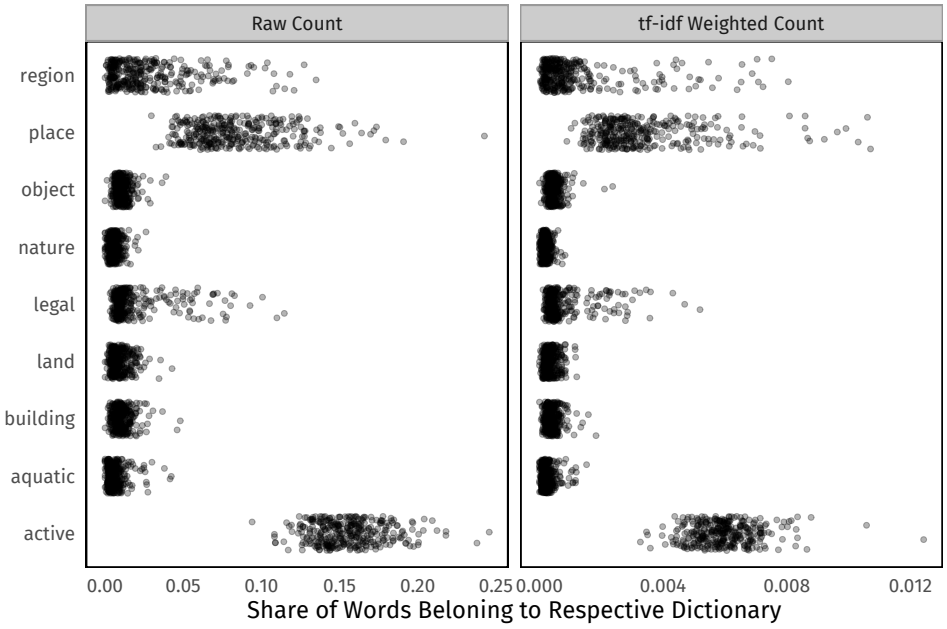
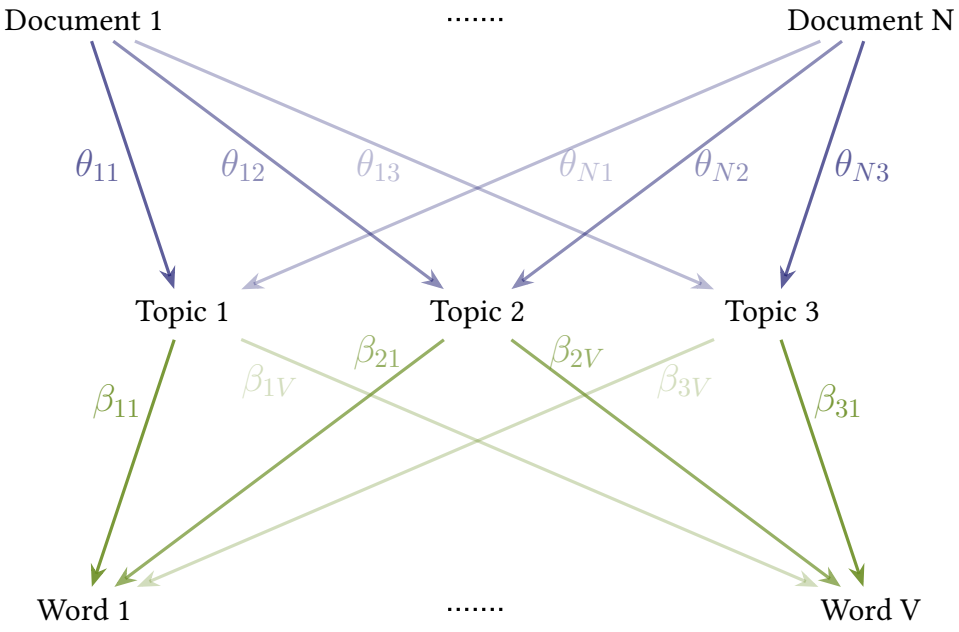


Figure A.3: Distributions of Raw and tf-idf Weighted Dictionary Scores



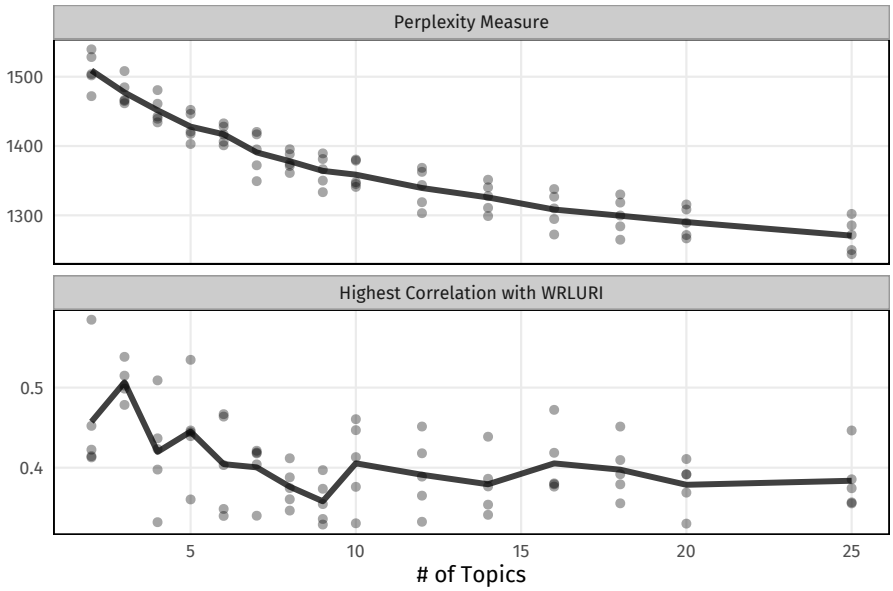
Notes: Raw counts are divided by their document length for comparability. The measures can be interpreted as the share of document tokens belonging to the respective category.

Figure A.4: LDA Data Generating Process



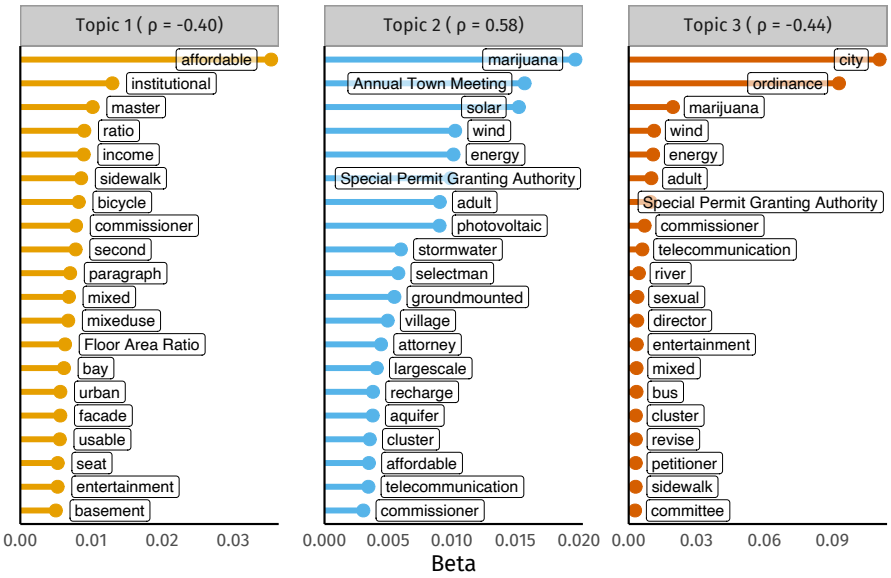
Inspired By: Bandiera et al. (2020).
Notes: Example assumes three latent topics. θ are the parameters of the true posterior distribution, while γ, as used in the text, are the parameters from the approximating distribution.

Figure A.5: LDA: Cross-Validation for Number of Latent Topics



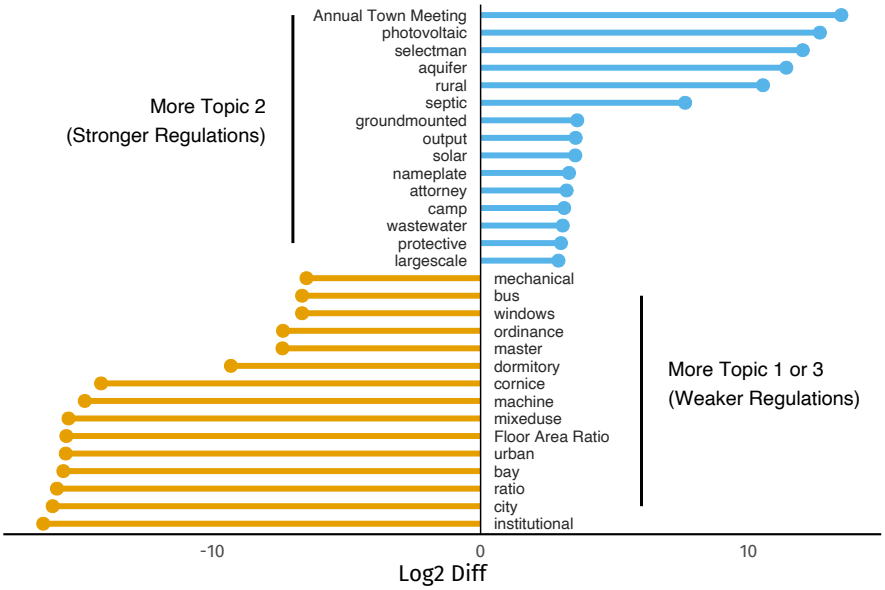
Notes: Lower is better in the first panel and higher is better in the second. Each point is a fold.

Figure A.6: LDA: Top Words per Latent Topic



Notes: Each value corresponds to the β value from the estimated LDA model (*ie* the estimated probability of a word being drawn from that specific topic).

Figure A.7: LDA: Words that Most Discriminate Between Topic 2 and Topics 1 or 3



Notes: Tokens with a β value less than 0.03 are excluded. Each point represents the base two logarithm of the ratio of β_{2v} and $\beta_{1v} + \beta_{3v}$.

Appendix A2 Characteristics of Highly Regulated Towns

With a measure of land use regulatory intensity in hand, that covers almost the entirety of one state, I now turn to discussing the geographic patterns of the Natural Language Processing Zoning Stringency Index, as well as town-level predictors of stringent LUR.

The near-universal coverage, at small geographic units, of the regulation measure enables me to present, to the best of my knowledge, new stylized facts about the geographic distribution of LUR. The first is that LUR are highly correlated over space. This has already been noted for larger geographic regions when looking at the Wharton Residential Land Use Regulation Index (*eg* cities in California and New England are highly regulated while those in the sunbelt are not). But even *within* these regions, there is a high degree of geographic clustering.

Next, I show that though more regulated municipalities allocate less overall land to development, of the land developed is given to residential purposes and less towards commercial or industrial uses.

Geographic Clustering of Land Use Regulation

To get an idea of the geographic clustering of land use regulatory intensity, I plot the relationship between a town's own NALPZ and the average of their direct neighbours. This is shown in Figure A.8. The blue line indicates the linear fit between the two, while the dashed grey line indicates 45°. It shows a clear pattern between the regulatory environment of neighbouring towns. On average, towns with more strict zoning regulations have neighbours with similarly strict land use policies. This aligns well with the geographic distribution of NALPZ shown previously in Figure 1.3.

This fact, to the best of my knowledge, has not been shown at this geographic detail. This has already been noted at the Metropolitan Statistical Area level using the WRLURI measure, but not *within* these units. Figure A.8 uses NALPZ for the entire sample in Massachusetts, but the relationship remains when considering the subsample of towns that surround Boston (specified as those that are part of the Pioneer/Rappaport Housing Regulation Database) and those further from the state's economy centre.

Though there is a high degree of geographic clustering of land use stringency, there also exists a great deal of heterogeneity with respect to the differential zoning policies between neighbouring towns. Figure A.9 highlights this distribution. The average differential of NALPZ between any two neighbouring town pairs is 0.74 of a standard

deviation (median is 0.58 of a s.d.). This is the key variation that my empirical strategy exploits to estimate the impact of more restrictive zoning.

Predictors of Restrictive Land Use Regulation

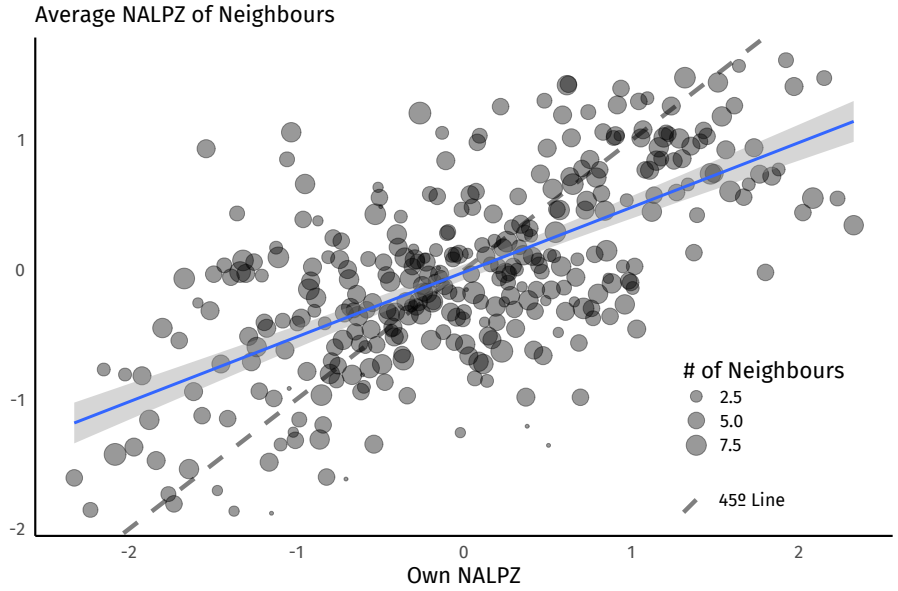
I now explore what town-level characteristics in 1970, pre Massachusetts Zoning Act, best predict future NALPZ levels. Specifically, I look at 1970 town-level census data and 1971 land use data from MassGIS to investigate the sorts of towns that implement stringent LUR.

The two strongest predictors of NALPZ are shown in Figure A.10: the percentage of land covered in forests (in the first panel) and the natural logarithm of housing units per square kilometre (in the second). Both of these variables correlate very strongly with NALPZ, with coefficients of correlation of -0.81 and 0.74 respectively. These results align with those found by Glaeser and Ward (2009), who consider minimum lot sizes (one aspect of zoning regulations). Remarkably, these patterns persist well into the future; corresponding relationships for 2010 are shown in Figure A.11. This fact may point to restrictive land use policies being used to preserve the contemporaneous city shape, makeup, and characteristics, once individual towns were given the legislative ability to implement their own zoning policies.

Two other interesting features of Massachusetts towns in 1970 that relate to regulatory intensity are shown in Figures A.12 and A.13. In the first panel of Figure A.12, the fraction of land being used for residential purposes is plotted against NALPZ, while in the second, the fraction of *developed* land being used for residential purposes is plotted instead. This highlights that, on average, towns that are more strictly regulated have less land overall for residential use, but of the land they already developed, slightly more is residential. This reversal pattern when considering all land compared with only developed land is not apparent when considering the share of land allocated to commercial or industrial uses as shown in Figure A.13. Unlike with residential land coverage, the implication does not vary if one considers all land, or only already-developed land. Towns with stricter zoning regulations allocate less absolute and relative land to industrial purposes.

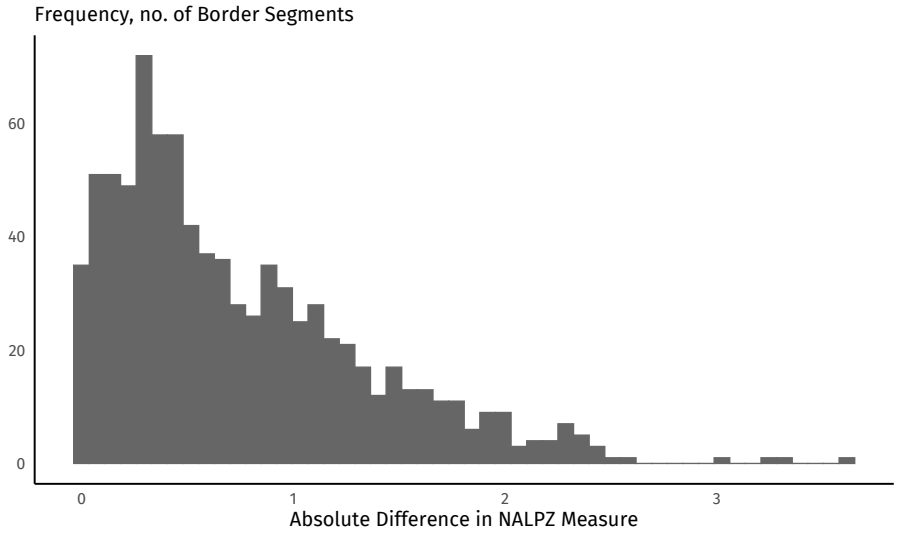
The trends shown in the last two figures are also virtually unchanged when one looks at the data in 2010, as shown in Figures A.14 and A.15.

Figure A.8: Relationship Between Own NALPZ and Average of Neighbouring Towns



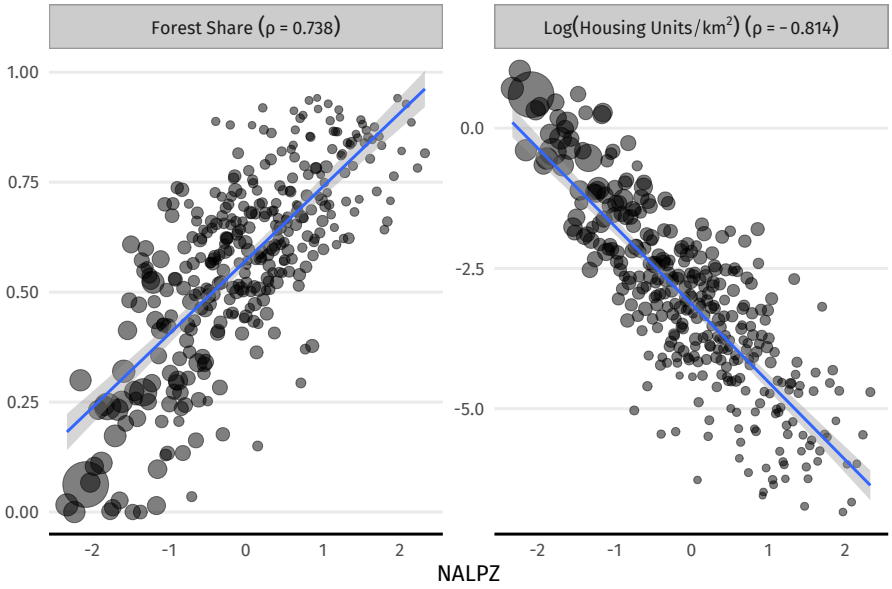
Notes: Each observation is a Massachusetts town. Size of dot corresponds to number of neighbours.

Figure A.9: Distribution of Absolute Differential of NALPZ at Borders



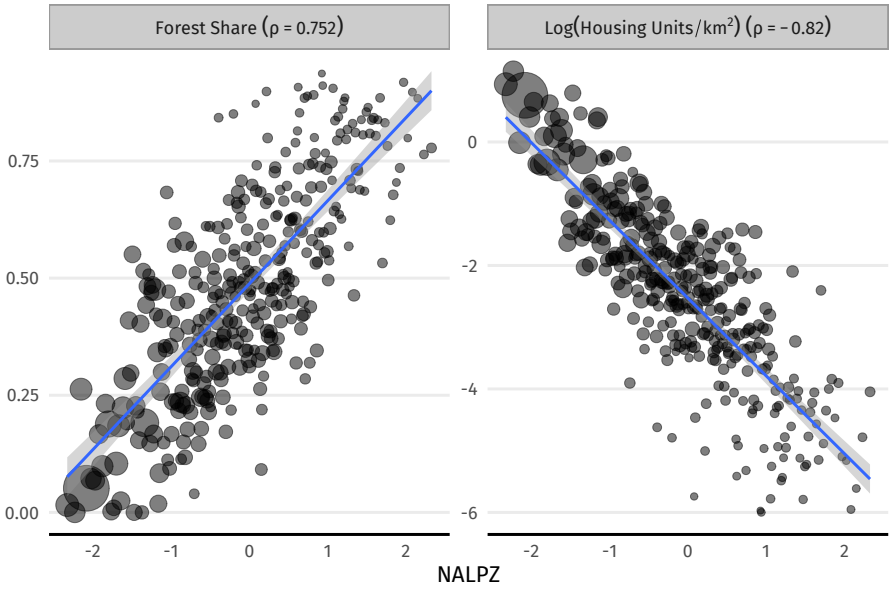
Notes: Each observation is a shared border between two Massachusetts Towns.

Figure A.10: 1970 Town Predictors of NALPZ



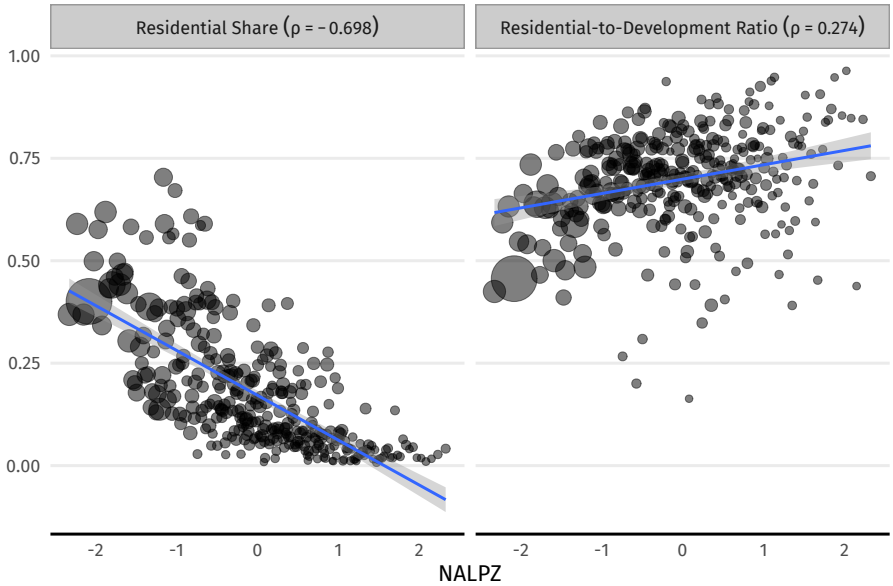
Notes: Each observation is a Massachusetts town in 1970. Size of dot corresponds to population.

Figure A.11: 2010 Town Predictors of NALPZ



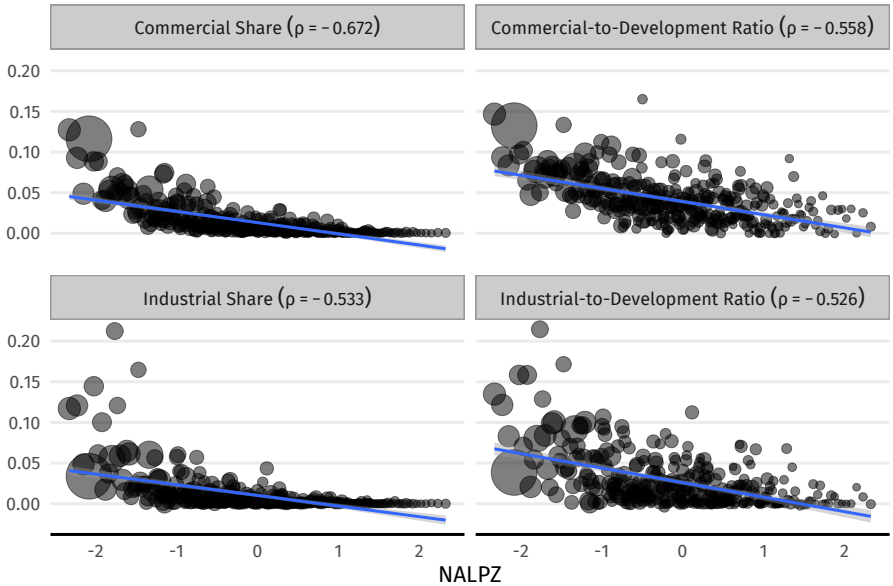
Notes: Each observation is a Massachusetts town in 2010. Size of dot corresponds to population.

Figure A.12: 1970 Town Residential Development and NALPZ



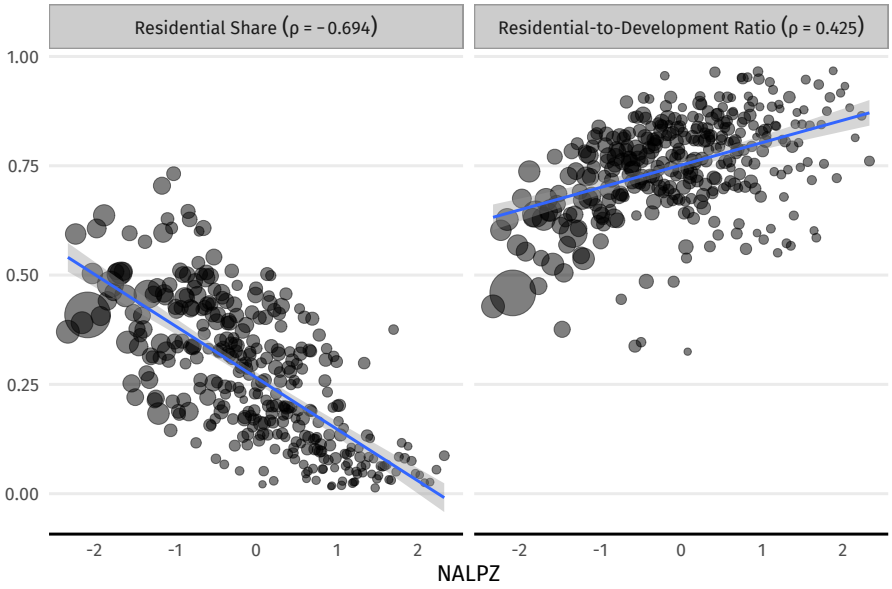
Notes: Each observation is a Massachusetts town in 1970. Size of dot corresponds to population.

Figure A.13: 1970 Town Industry Development and NALPZ



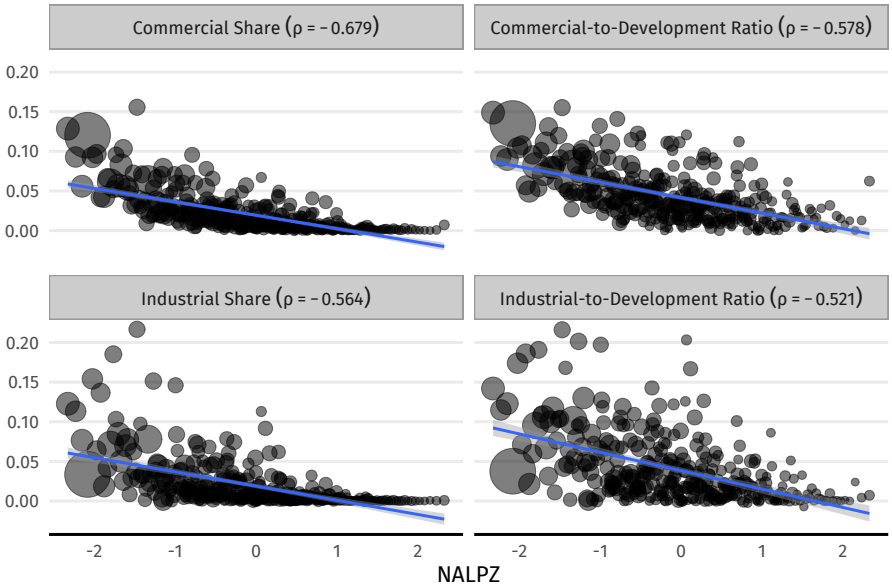
Notes: Each observation is a Massachusetts town in 1970. Size of dot corresponds to population.

Figure A.14: 2010 Town Residential Development and NALPZ



Notes: Each observation is a Massachusetts town in 2010. Size of dot corresponds to population.

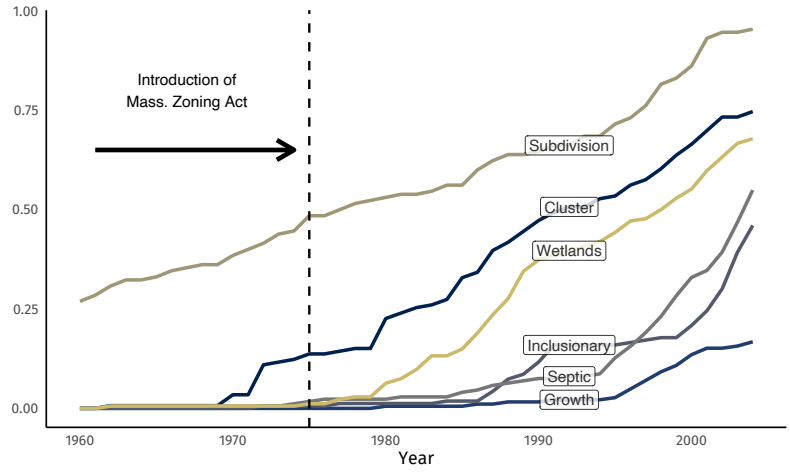
Figure A.15: 2010 Town Industry Development and NALPZ



Notes: Each observation is a Massachusetts town in 2010. Size of dot corresponds to population.

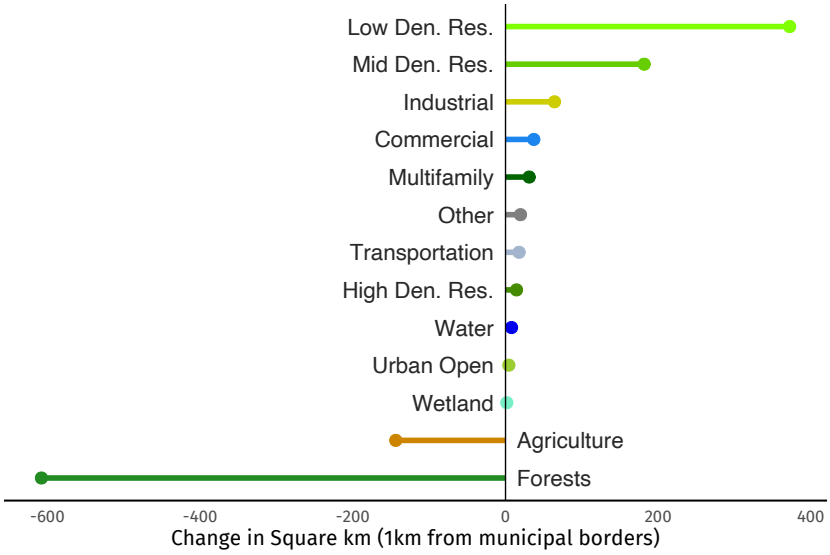
Appendix A3 Additional Figures

Figure A.16: Share of Towns Adopting Land Use Regulations by Year, Cumulative



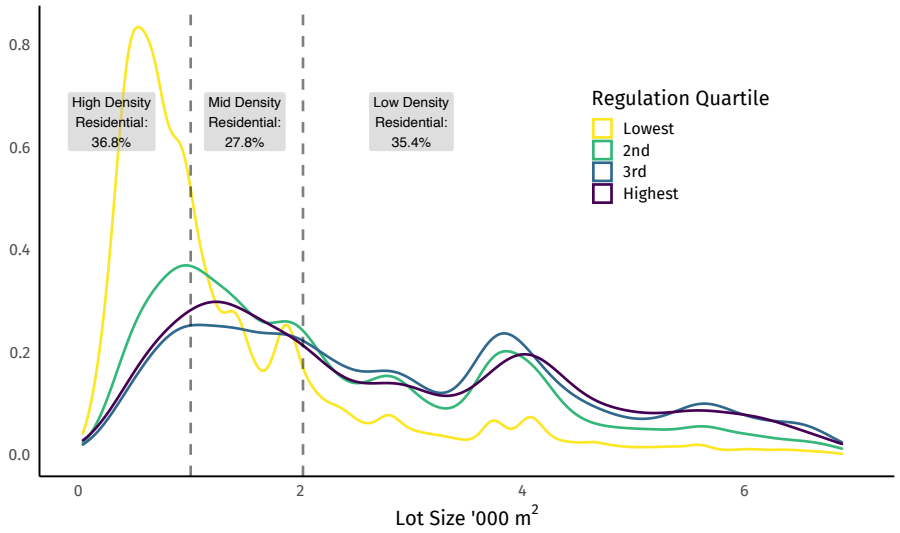
Notes: Some LUR enable development and others inhibit it. Data comes from the Pioneer Institute/Rappaport Institute (PIRI, 2005) Housing Regulation Database for Massachusetts.

Figure A.17: Land Use Conversion from 1971 to 1999



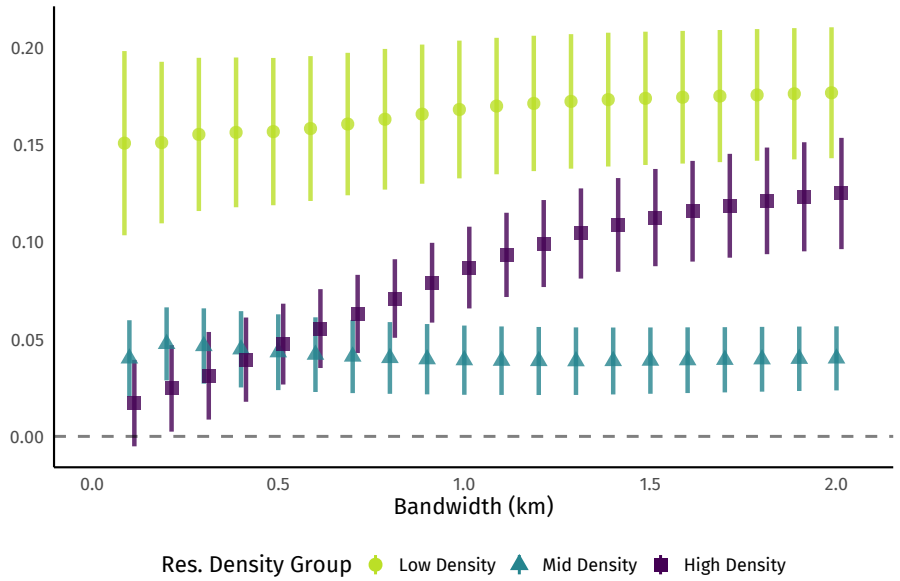
Notes: Each bar indicates the change in area in km^2 for the respective land use category from 1971 to 1999. Data from MassGIS

Figure A.18: Distribution of Lot Sizes by Regulation Quartile



Notes: Each line represents a density polygon plotted for each quartile of the NALPZ. The dashed lines separate the density classes according to the land use categories from MassGIS.

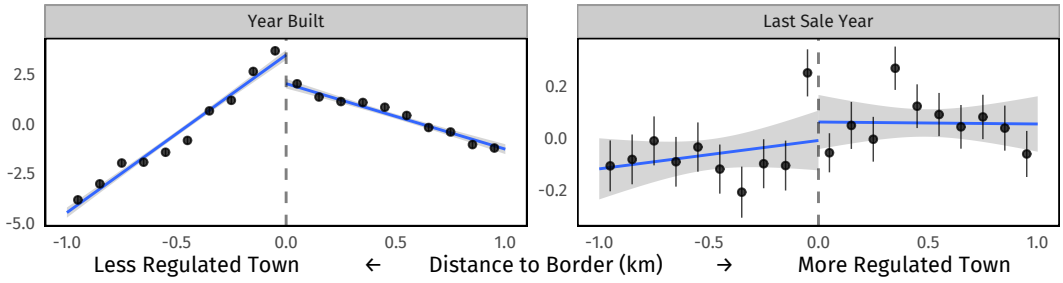
Figure A.19: Spatial RDD by Residential Density Grouping: Logarithm of Lot Size on NALPZ



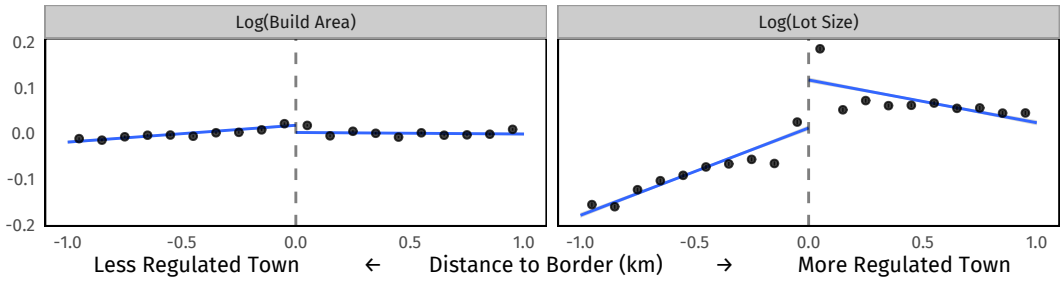
Notes: Plots the estimated β value from estimating Equation 1.8 separately for each group of parcels according their density grouping. Respective outcome is regressed on the Natural Language Processing Zoning Stringency Index (NALPZ), border segment fixed effects, and border segment-specific distance controls. Density groupings are defined by MassGIS. Point estimates and 95% confidence intervals are given. Standard errors clustered at the town level.

Figure A.20: Residualized Outcomes by Less/More Regulated Town for Every Town Border

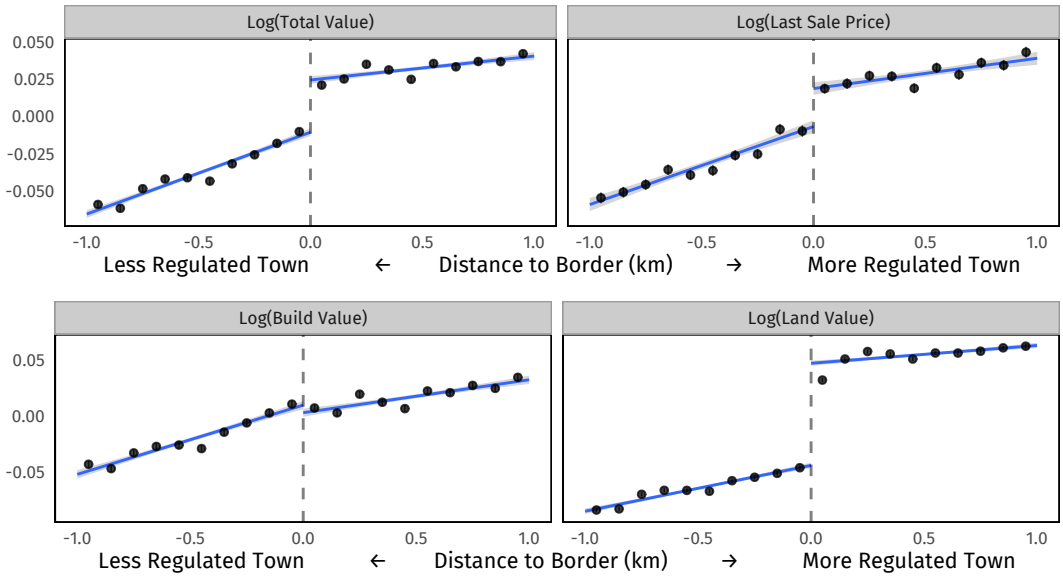
(a) Housing Market Outcomes



(b) Housing Attributes Outcomes



(c) House Price Outcomes



Notes: Each panel plots the respective outcome residualized by border-segment fixed effects. Residuals are binned by every 100m. Mean and the respective standard error plotted for each bin. Towns may have housing units in both the "less regulated" and "more regulated" categories, due to having multiple neighbours, but as each unit is matched to a unique border an observation can only belong to one category.

Appendix A4 Spatial RDD Regression Tables

Table A.2: Spatial RDD: Main Results

	Log(Lot Size)	Year Built	Log(Building Size)	Year Last Sold	Build. Height	Log(Build. Height)
	(1)	(2)	(3)	(4)	(5)	(6)
0.1km	0.267***	−1.950***	−0.029	−0.776***	0.179	0.014
	[0.217,0.317]	[-3.248,-0.651]	[-0.103,0.045]	[-1.317,-0.235]	[-0.084,0.442]	[-0.011,0.038]
	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	42,428/408/226	42,428/408/226
0.2km	0.267***	−2.504***	−0.025	−0.806***	0.284***	0.022**
	[0.223,0.311]	[-3.711,-1.297]	[-0.102,0.051]	[-1.320,-0.292]	[0.078,0.490]	[0.003,0.041]
	105,726/632/290	105,726/632/290	105,726/632/290	105,726/632/290	94,422/537/250	94,422/537/250
0.3km	0.270***	−2.514***	−0.019	−0.788***	0.360***	0.028***
	[0.227,0.312]	[-3.777,-1.251]	[-0.095,0.058]	[-1.280,-0.296]	[0.181,0.539]	[0.011,0.044]
	158,384/676/293	158,384/676/293	158,384/676/293	158,384/676/293	149,965/598/259	149,965/598/259
0.4km	0.272***	−2.607***	−0.014	−0.823***	0.369***	0.029***
	[0.231,0.313]	[-3.900,-1.314]	[-0.090,0.062]	[-1.310,-0.335]	[0.204,0.534]	[0.014,0.044]
	212,819/708/295	212,819/708/295	212,819/708/295	212,819/708/295	208,260/629/264	208,260/629/264
0.5km	0.273***	−2.576***	−0.009	−0.851***	0.363***	0.029***
	[0.232,0.314]	[-3.875,-1.277]	[-0.084,0.067]	[-1.330,-0.373]	[0.196,0.529]	[0.013,0.044]
	266,765/722/297	266,765/722/297	266,765/722/297	266,765/722/297	266,388/647/268	266,388/647/268
0.6km	0.270***	−2.528***	−0.005	−0.845***	0.365***	0.030***
	[0.230,0.311]	[-3.799,-1.257]	[-0.080,0.071]	[-1.317,-0.373]	[0.196,0.534]	[0.014,0.045]
	319,649/737/297	319,649/737/297	319,649/737/297	319,649/737/297	323,052/661/269	323,052/661/269
0.7km	0.268***	−2.437***	−0.001	−0.826***	0.358***	0.030***
	[0.228,0.308]	[-3.681,-1.192]	[-0.077,0.075]	[-1.298,-0.354]	[0.191,0.526]	[0.015,0.045]
	372,480/751/298	372,480/751/298	372,480/751/298	372,480/751/298	379,757/680/270	379,757/680/270
0.8km	0.266***	−2.310***	0.000	−0.788***	0.345***	0.029***
	[0.227,0.306]	[-3.535,-1.085]	[-0.075,0.076]	[-1.263,-0.314]	[0.181,0.510]	[0.015,0.044]
	424,921/759/299	424,921/759/299	424,921/759/299	424,921/759/299	436,222/692/270	436,222/692/270

continues

continued

	Log(Lot Size)	Year Built	Log(Building Size)	Year Last Sold	Build. Height	Log(Build. Height)
	(1)	(2)	(3)	(4)	(5)	(6)
0.9km	0.264***	−2.153***	0.002	−0.745***	0.342***	0.029***
	[0.225,0.303]	[-3.370,-0.936]	[-0.074,0.077]	[-1.221,-0.270]	[0.180,0.504]	[0.015,0.043]
	476,181/767/299	476,181/767/299	476,181/767/299	476,181/767/299	491,889/701/270	491,889/701/270
1km	0.263***	−2.097***	0.004	−0.708***	0.336***	0.029***
	[0.224,0.302]	[-3.309,-0.885]	[-0.072,0.079]	[-1.184,-0.232]	[0.175,0.497]	[0.015,0.042]
	525,329/772/299	525,329/772/299	525,329/772/299	525,329/772/299	545,086/706/270	545,086/706/270
1.1km	0.262***	−2.111***	0.006	−0.674***	0.331***	0.029***
	[0.223,0.302]	[-3.320,-0.902]	[-0.070,0.082]	[-1.149,-0.200]	[0.170,0.492]	[0.015,0.042]
	572,867/778/299	572,867/778/299	572,867/778/299	572,867/778/299	597,265/714/270	597,265/714/270
1.2km	0.262***	−2.127***	0.008	−0.651***	0.323***	0.028***
	[0.223,0.301]	[-3.336,-0.918]	[-0.068,0.084]	[-1.124,-0.179]	[0.164,0.482]	[0.015,0.042]
	618,422/783/300	618,422/783/300	618,422/783/300	618,422/783/300	647,031/719/270	647,031/719/270
1.3km	0.262***	−2.124***	0.010	−0.635***	0.313***	0.028***
	[0.223,0.301]	[-3.337,-0.910]	[-0.066,0.086]	[-1.104,-0.165]	[0.156,0.470]	[0.014,0.041]
	661,056/786/300	661,056/786/300	661,056/786/300	661,056/786/300	693,683/723/271	693,683/723/271
1.4km	0.262***	−2.139***	0.011	−0.623***	0.307***	0.027***
	[0.223,0.301]	[-3.364,-0.915]	[-0.065,0.087]	[-1.090,-0.157]	[0.152,0.461]	[0.014,0.040]
	702,007/789/301	702,007/789/301	702,007/789/301	702,007/789/301	738,117/724/272	738,117/724/272
1.5km	0.262***	−2.163***	0.012	−0.617***	0.302***	0.027***
	[0.223,0.301]	[-3.399,-0.927]	[-0.063,0.088]	[-1.082,-0.153]	[0.151,0.454]	[0.014,0.040]
	740,910/793/301	740,910/793/301	740,910/793/301	740,910/793/301	781,053/726/272	781,053/726/272
1.6km	0.263***	−2.161***	0.013	−0.613***	0.301***	0.027***
	[0.223,0.302]	[-3.408,-0.914]	[-0.062,0.089]	[-1.076,-0.150]	[0.151,0.451]	[0.014,0.040]
	777,388/795/301	777,388/795/301	777,388/795/301	777,388/795/301	821,162/728/272	821,162/728/272

continues

continued

	Log(Lot Size)	Year Built	Log(Building Size)	Year Last Sold	Build. Height	Log(Build. Height)
	(1)	(2)	(3)	(4)	(5)	(6)
1.7km	0.263***	−2.123***	0.014	−0.611***	0.299***	0.027***
	[0.224,0.303]	[-3.376,-0.869]	[-0.061,0.089]	[-1.073,-0.149]	[0.150,0.448]	[0.014,0.040]
	811,506/799/301	811,506/799/301	811,506/799/301	811,506/799/301	858,107/734/272	858,107/734/272
1.8km	0.264***	−2.075***	0.014	−0.612***	0.295***	0.027***
	[0.224,0.305]	[-3.334,-0.816]	[-0.061,0.089]	[-1.073,-0.152]	[0.147,0.443]	[0.014,0.040]
	843,251/800/301	843,251/800/301	843,251/800/301	843,251/800/301	892,419/736/272	892,419/736/272
1.9km	0.265***	−2.037***	0.014	−0.613***	0.291***	0.026***
	[0.225,0.306]	[-3.303,-0.770]	[-0.060,0.088]	[-1.073,-0.153]	[0.143,0.439]	[0.014,0.039]
	872,953/801/301	872,953/801/301	872,953/801/301	872,953/801/301	924,831/738/272	924,831/738/272
2km	0.267***	−1.993***	0.014	−0.614***	0.288***	0.026***
	[0.225,0.308]	[-3.267,-0.720]	[-0.060,0.088]	[-1.074,-0.154]	[0.140,0.436]	[0.014,0.039]
	900,484/801/301	900,484/801/301	900,484/801/301	900,484/801/301	955,040/738/272	955,040/738/272

Notes: Each cell is a separate regression of the outcome variable (column name) on the NALPZ variable. The respective bandwidth is indicated in each row. 95% confidence intervals indicated in brackets. Last row in each cell indicates the sample size (number of single-family tax parcels), the number of town borders (segments), and number of towns included in the regression, respectively.

Table A.3: Spatial RDD: Robustness and Specification Checks

	Log(Total Value)	Log(Lot Size)	Year Built	Log(Building Size)	Year Last Sold	Log(Last Sale Price)
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline	0.049**	0.267***	−1.950***	−0.029	−0.776***	0.059***
	[0.007,0.092]	[0.217,0.317]	[-3.248,-0.651]	[-0.103,0.045]	[-1.317,-0.235]	[0.020,0.098]
	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	32,857/476/252
No Weigthing	0.052**	0.269***	−2.236***	−0.026	−0.793***	0.055***
	[0.007,0.097]	[0.222,0.316]	[-3.457,-1.016]	[-0.102,0.050]	[-1.325,-0.261]	[0.015,0.095]
	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	32,857/476/252
Quad. Geography	0.052**	0.272***	−1.526**	0.005	−0.881***	0.054**
	[0.001,0.103]	[0.226,0.318]	[-2.987,-0.064]	[-0.081,0.091]	[-1.435,-0.327]	[0.013,0.095]
	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	32,857/476/252
School Dist. Rank	0.044	0.242***	−1.996**	−0.087**	−0.394	0.055***
	[-0.009,0.097]	[0.174,0.309]	[-3.621,-0.371]	[-0.158,-0.015]	[-1.118,0.331]	[0.020,0.091]
	47,395/479/214	47,395/479/214	47,395/479/214	47,395/479/214	47,395/479/214	28,099/419/204
School Quality	0.016	0.197***	−2.070**	−0.074*	−0.400	0.027
	[-0.039,0.070]	[0.136,0.259]	[-3.650,-0.491]	[-0.152,0.003]	[-1.228,0.429]	[-0.018,0.072]
	55,239/544/269	55,239/544/269	55,239/544/269	55,239/544/269	55,239/544/269	32,300/473/247
Rank & Quality	0.030	0.167***	−1.433	−0.088**	0.050	0.038
	[-0.034,0.093]	[0.096,0.239]	[-3.242,0.376]	[-0.165,-0.011]	[-0.845,0.945]	[-0.010,0.087]
	46,692/477/211	46,692/477/211	46,692/477/211	46,692/477/211	46,692/477/211	27,640/418/201
Prop. Tax	0.049**	0.271***	−1.967***	−0.029	−0.791***	0.060***
	[0.006,0.092]	[0.222,0.321]	[-3.271,-0.663]	[-0.103,0.046]	[-1.321,-0.261]	[0.020,0.099]
	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	32,857/476/252

continues

continued

	Log(Total Value)	Log(Lot Size)	Year Built	Log(Building Size)	Year Last Sold	Log(Last Sale Price)
	(1)	(2)	(3)	(4)	(5)	(6)
All Controls	0.028	0.176***	−1.691*	−0.085**	0.050	0.040
	[-0.036,0.091]	[0.106,0.247]	[-3.508,0.126]	[-0.161,-0.009]	[-0.825,0.924]	[-0.008,0.089]
	46,692/477/211	46,692/477/211	46,692/477/211	46,692/477/211	46,692/477/211	27,640/418/201
School Distr. FE	−0.010	0.122	−2.800	−0.047	0.272	0.011
	[-0.096,0.077]	[-0.066,0.311]	[-6.663,1.063]	[-0.166,0.073]	[-1.346,1.891]	[-0.097,0.119]
	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	56,196/552/278	32,857/476/252
New Devel.	0.043*	0.292***	−4.309***	−0.079*	−0.687*	0.078***
	[-0.002,0.088]	[0.208,0.376]	[-5.666,-2.951]	[-0.166,0.008]	[-1.426,0.052]	[0.041,0.115]
	21,093/440/247	21,093/440/247	21,093/440/247	21,093/440/247	21,093/440/247	12,264/329/208

Notes: Each cell is a separate regression of the outcome variable (column name) on the NALPZ variable (except for the last two rows, where the outcome is regressed on the variable in the first column). The type of robustness check is named in the respective row of the first column. 95% confidence intervals indicated in brackets. Last row in each cell indicates the sample size (number of single-family tax parcels), the number of town borders (segments), and number of towns included in the regression, respectively.

Appendix A5 Spatial General Equilibrium Model and Amenities

Model Setup

A city[31] embedded in a larger economy is assumed to comprise of a set of discrete locations ($\mathbb{S} = \{1, \ldots, S\}$), which differ in terms of housing supply, local amenities, and access to workplaces. A Worker (o) chooses a residence-workplace pair (i, j) that maximizes their utility. They derive utility from consuming a freely-traded numeraire good (c_o), housing (h_i), residential amenities (B_i), and dis-utility from commuting ($e^{\kappa\tau_{ij}}$) that depends on travel time (τ_{ij}). Workers are heterogeneous with respect to resident-workplace pairs, which is captured by z_{ijo}. Utility takes the Cobb-Douglas form:

$$U_{ijo} = \frac{B_i z_{ijo}}{e^{\kappa\tau_{ij}}} \left(\frac{c_o}{\beta}\right)^{\beta} \left(\frac{h_i}{1-\beta}\right)^{1-\beta} \tag{A.4}$$

where β governs the share of income on the numeraire good. As is standard, the idiosyncratic shocks, z_{ijo}, are assumed to be Fréchet distributed:

$$F(z_{ijo}) = e^{-T_i z_{ijo}^{-\epsilon}} \tag{A.5}$$

where $T_i > 0$ determines the average utility derived from living in block i, and $\epsilon > 1$ governs the dispersion of the shock. Maximizing utility given workplace wage (w_j), housing costs (q_i), results in the following form for indirect utility:

$$V_{ijo} = \frac{z_{ijo} B_i w_j q_i^{1-\beta}}{e^{\kappa\tau_{ij}}} \tag{A.6}$$

The properties of the Fréchet distribution imply that the probability that a worker lives in block i is given as:

$$\pi_i = \underbrace{\frac{\sum_{j=1}^{S} T_i (e^{\kappa\tau_{ij}} q_i^{1-\beta})^{-\epsilon} (B_i w_j)^{\epsilon}}{\sum_{r=1}^{S}\sum_{s=1}^{S} T_r (e^{\kappa\tau_{rs}} q_r^{1-\beta})^{-\epsilon} (B_r w_s)^{\epsilon}}}_{\Phi} = \frac{T_i q_i^{(1-\beta)-\epsilon} B_i^{\epsilon} \, \text{CMA}_i^{\epsilon}}{\Phi} \tag{A.7}$$

where

$$\text{CMA}_i = \sum_{j=1}^{S} (w_j / e^{\kappa\tau_{ij}})^{\epsilon} \tag{A.8}$$

[31] Here I consider the "city" to be Massachusetts. As there is only one Combined Statistical Area as defined by the US Census Bureau, this assumption is not unwarranted.

denotes the Commuting Market Access of block i. Intuitively, this term is higher when a block is located close to well paying jobs.

The population is assumed to have full mobility, implying that residents will move until expected utility is equalized across residence-workplace pairs, as well as to the reservation level of utility in the larger economy ($\overline{U}$):

$$\overline{U} = \mathbb{E}[U] = \underbrace{\Gamma\left(\frac{\epsilon - 1}{\epsilon}\right)}_{\gamma} \underbrace{\left[\sum_{r=1}^{S}\sum_{s=1}^{S} T_r (e^{\kappa\tau_{rs}} q_r^{1-\beta})^{-\epsilon} (B_r w_s)^{\epsilon}\right]^{1/\epsilon}}_{\Phi} \tag{A.9}$$

where $\gamma = \Gamma(\frac{\epsilon-1}{\epsilon})$ is the Gamma function. Given the residential choice probabilities (Eq. A.7) and population mobility (Eq. A.9) we arrive at the following relationship:

$$\frac{B_i T_i^{1/\epsilon}}{\bar{U}/\gamma} = \left(\frac{H_i}{H}\right)^{1/\epsilon} \frac{q_i^{1-\beta}}{\text{CMA}_i} \tag{A.10}$$

where H_i is the population of block i and H is the population of the city. We can remove the block-invariant components by dividing the equation by its geometric mean. This makes it possible to write block-level amenities in terms of observable characteristics and model parameters:

$$\frac{B_i^*}{\widetilde{B_i^*}} = \left(\frac{H_i}{\widetilde{H_i}}\right)^{1/\epsilon} \left(\frac{q_i}{\widetilde{q_i}}\right)^{1-\beta} \left(\frac{\text{CMA}_i}{\widetilde{\text{CMA}_i}}\right)^{-1/\epsilon} \tag{A.11}$$

where $B_i^* = B_i T_i^{1/\epsilon}$ is the composite residential amenity term and $\widetilde{X} = \left(\prod_i^N X_i\right)^{1/N}$ denotes the geometric mean of the respective variable over all residential locations. I take parameter values of ϵ, β, and κ from recent literature (described more in the main text). Then I calculate the normalized, composite residential amenity term. I use this in my regressions to test differences in amenity values across municipal borders.

Parameter Values

The parameter values used to estimate the implied amenity values are given in the following table:

Table A.4: Model Calibration: Parameters

Source	ϵ	$\nu = \epsilon\kappa$	$1 - \beta$	κ
Ahlfeldt et al. (2015)	6.6190	0.0951	0.25	0.01537
Tsivanidis (2018) [Low-skilled]	2.840	0.0336	0.24	0.012
Tsivanidis (2018) [High-skilled]	2.054	0.0242	0.24	0.012
Heblich et al. (2020)	5.25	0.05203	0.25	0.0099

Chapter 2

On the Measurement and Causes of Land Use Regulation*

Abstract:

Land use regulations are ubiquitous and are often the most significant impediment to new development. But our understanding of these regulations is hampered by lack of data and the heterogeneity in the forms they take. Using zoning bylaw text and data from Massachusetts, this paper contributes by providing descriptive facts on the measurement and causes of land use regulations. First, I investigate the performance of a set of easy-to-implement machine learning algorithms in predicting survey-based measures of regulation stringency. Second, using several variable selection procedures I test the town-level attributes that best predict zoning. Third, I explore how the demographic composition of towns has varied over 40 decades from 1970–2010 with respect to the regulation level, starting before locally implemented zoning was possible. I find that the latent Dirichlet allocation model, a latent-mixture model works best in deriving an index of regulation from zoning text; current day zoning regulation is best predicted by historical land use patterns; some demographic characteristics have changed substantially since the implementation of zoning, noticeably the share of the population that is non-white.

*The "companion" paper referred to throughout this chapter is Chapter 1 in this dissertation.

2.1 Introduction

The importance of land use regulations has been known for some time (*eg* Frieden, 1979). They became widespread at the local level in the US during the 1970s, along with the construction of the interstate highway system and expansion of suburban building. As land use regulation is often the biggest constraint to new development, it is a somewhat surprising that little is known on why some places are highly-regulated and others less so. Moreover, as there is no federal standard, and generally no state-level standard, the rules are highly heterogeneous, making them difficult to contrast and compare.

When studying land use regulation, there are generally two primary goals. The first is to determine the causes of regulation: why are some towns highly regulated and others not, even within the same metropolitan area? The second is to measure the consequences: how do restrictive land use regulations impact the housing, development, neighbourhood sorting, etc. Shared between these two goals is a need to accurately measure land use restrictions. This paper is focused on the determinants of regulation, as well as on how to better measure the level of restrictiveness.

At the same time, the tools available to the researcher to study these questions have improved. Machine learning and natural language processing methods exist that can find insights from non-standard data sources, such as from geospatial and legal documents. These methods can be combined with proven econometric techniques and economic models to further our understanding of land use regulations.

This paper has three primary goals. The first is to test the ability of a set of widely available machine learning methods to measure the restrictiveness of a locality's land use regulations using only the text from the zoning bylaws as predictors. These methods may prove useful for expanding our knowledge of regulation in jurisdictions where we lack survey responses. Second, using census and land use data from the 1970s, before widespread zoning restrictions, this paper explores the predictive power of commonly cited causes of regulation. As I have at my disposal several indices of regulation (including one derived from a machine learning algorithm), I am able to test whether certain local factors are generally good indicators of highly regulated towns, or if it depends significantly on the index being considered. Third, I investigate how the characteristics of weakly and strongly regulated towns has developed over time.

Among the machine learning methods considered, I find that the latent Dirichlet allocation model, an unsupervised mixture model, derives the strongest predictor, outperforming supervised methods.[1] Turning to the predictors of regulation, the paper finds

[1]Supervised models incorporate the outcome in the estimation procedure while unsupervised models use only the explanatory variables. Supervised models are generally concerned with prediction while

that historical land use, especially the amount of undeveloped land, is the best predictor of current day regulation. Finally, this paper shows that there has been a large divergence between towns in the bottom quintile of regulation and the rest in terms of the share of the population that is foreign, in poverty, and non-white.

The use of machine learning on with text, and using text as data more generally, is relatively new to economics (Gentzkow et al., 2019). I have yet to find any applications to land use or zoning specifically. Given the availability of accessible regulations online, a naturally arising question is if these methods can be used on such texts to expand our understanding of land use regulations. To explore this question I consider the ability of an array of machine learning (ML) procedures to predict the level of land use regulation as measured by existing survey data. The ML models can be classified into three broad categories: i) penalized linear regressions (*eg* lasso), ii) decomposition methods (*eg* principal components analysis), and iii) decision trees (*eg* random forest). I consider these methods because they come with robust implementations in various standard statistical programming languages, making them relatively straightforward for researchers to implement.

As outcome measures of regulation, I use three different indices. The first is the standard in the literature, the Wharton Residential Land Use Regulation Index (Gyourko et al., 2008). The second I derive via principal components analysis from a database on housing regulations in Massachusetts compiled by researchers from the Pioneer and Rappaport Institutes (PIRI, 2005). The third is an index I have created in a companion paper, using a machine learning method called latent Dirichlet allocation with data coming from the zoning bylaws. The index is called the Natural Language Processing Zoning Index (NALPZ).[2] Of the 351 towns in Massachusetts, the first index covers 79, the second 187, and the third 341. Because land use regulations are high dimensional and heterogenous, these indices attempt to approximate the regulatory burden with a unidimensional measure through different methods. Comparing and contrasting results across the indices helps to determine whether particular methods are better suited for a certain index or if they work more generally.

To investigate what community characteristics are best predictors of land use regulation, I test the predictive power of town-level features prior to the 1975 Massachusetts Zoning Act. The Act formally delegated the authority to implement zoning policies to the local (town) level. Prior to this local jurisdictions required approval from the state before changes to local zoning could be implemented. I use demographic measures from

unsupervised methods search for hidden patterns or clusters.

[2]Both the method and the data are briefly described in later in this paper. More detailed explanations can be found in the companion paper.

the 1970 US census as well as land use coverage data from aerial photographs taken in 1971. I use several standard feature selection methods from the statistical learning toolkit to uncover the variables most able to discern between strongly and weakly regulated places.

The last part of the paper describes trends in demographic characteristics over a period of 40 years, from 1970 to 2010. To ease graphical interpretation, I divide the sample of towns in Massachusetts covered by the most extensive index, the NALPZ, into quintiles and explore how features of these groups vary over the decennial censuses.

I find that of the ML methods considered, the latent Dirichlet allocation model is best at generating an index of regulation based on its correlation with the two survey-based regulation outcomes. Given that it an unsupervised method (*ie* does not require a response variable), it is rather surprising that this method performs better than supervised models designed to predict a given outcome. This is most likely due to the supervised models overfitting the relatively small sample sizes covered by the Wharton and Pioneer/Rappaport indices.

Turning to predictors of regulation, I find that historical development and density are better than demographic characteristics (*eg* share of college graduates) and other land use categories (*eg* water coverage) at explaining the amount of regulation, regardless of the index considered or the selection method used. This supports previous work (*eg* Glaeser and Ward, 2009) that found that historical density and forest cover are strong predictors of land use regulation. I build on this line of work by considering a wider array of demographic and land use variables, as well as using techniques designed to find the best predictors among a set of candidates, as opposed to inferring from the significance of some regressors in an OLS framework.

Lastly, I show that a gap in the share of foreign-born, population in poverty, and non-white residents has widened over time between towns in the bottom quintile of land use regulation compared with the other quintiles, with the bottom quintile becoming more foreign-born, poorer, and non-white. This has happened while overall population and density has remained remarkably stable among the quintile groups. This suggests that sorting between towns of varying levels of zoning has occurred, whether zoning was the root cause or not. Future research on zoning will need to account for these facts when theorizing on the causes of land use regulation.

There are several theories for why some jurisdictions are more or less regulated.[3] Early models of zoning specified land use regulation as a function of land prices (Wallace, 1988; McMillen and McDonald, 1991). In this framework, the level of zoning and land values are determined simultaneously. Parcels of land that are zoned for higher density,

[3]See Gyourko and Molloy (2015) for a more thorough summary of the literature.

for example, are done so because that is the use that will maximize the value of that particular tract of land.

Fischel (2001) argues that local residents play an even bigger role in determining the amount of regulation. He theorizes that residents lobby for policies meant to hamper new development to preserve the value of their land (and by extension home), as it is often their largest asset. This is done as homeowners do not have other alternatives to insure against the value of their homes. By extension, homeowners should be more in favour of stringent land use regulations than renters, who do not stand to benefit from increases property values (see Hilber and Robert-Nicoud, 2013; Ortalo-Magné and Prat, 2014). However, empirical evidence for this theory is rather limited (Glaeser and Ward, 2009).

Another potential cause is the supply of land available for development, with both geography and pre-existing land use and density patterns influencing this supply. Saiz (2010) documents a strong relationship between geographical constraints to development at the metropolitan area level (as measured by land lost to water and areas with steep slopes) and land use regulation, suggesting that a dearth of buildable land may lead local residents and politicians to playing a larger role in the development process. Glaeser and Ward (2009) find that historical density and forest cover are strong predictors of the current level of regulation. Specifically, towns in Massachusetts with low historical density and more forest coverage are more regulated today. This suggests a counter hypothesis: that land use regulation tends to conform to pre-existing patterns of land use and development. A key difference between the two studies is the unit of analysis. Saiz (2010) considers the metropolitan area while Glaeser and Ward (2009) use a town. This suggests that conclusions on the causes of regulation may differ on account of the analysis being done between or within regions.

Other work has highlighted the role of strategic interaction between communities in determining zoning regulations. This comes in two primary forms. The first concerns the externalities associated with one town's zoning regulations on neighbouring towns. And example of this is given theoretically by Helsley and Strange (1995) who show that when one town restricts development that population may be displaced into nearby communities, diverting traffic and congestion. Therefore one locality's regulations will depend in part on the regulations of neighbouring towns. The second form concerns sorting into neighbourhoods with respect to tastes for consumption of public goods or into neighbourhoods containing households with similar demographics. This can result in neighbourhoods being stratified along income and demographic lines. Calabrese et al. (2007) show that towns may specify a minimum housing quality and property tax to ensure local public goods are sufficiently financed. This leads to income-based stratifi-

cation. Intentional or not, this neighbourhood sorting mechanism often divides among ethnic as well as socio-economic lines, which is why it is often referred to as "exclusionary" zoning. However, as these two demographic characteristics are highly correlated it is difficult to disentangle whether the fiscal or exclusionary motive is most responsible.

2.2 Data

The data come from three primary sources. I use US census data for demographic information at the town-level in Massachusetts. I focus on the year 1970, the census immediately prior to the implementation of the Massachusetts Zoning Act (1975), to describe towns before they were delegated the authority to implement their own zoning regulations, and at subsequent decades to chart the evolution of the demographic composition over time. I complement this with data on land use across Massachusetts from 1971 (also pre-Massachusetts Zoning Act). This data classifies every parcel of land in the state according to its primary usage. The third source is the zoning bylaws from the individual towns.

US Census Data

Data on town-level demographics is based on the US Census. The data itself comes from IPUMS National Historical Geographic Information System. The primary unit of analysis is the county subdivision, which corresponds directly to a town in Massachusetts. When investigating the best predictors of regulation I focus on the year 1970, for reasons mentioned above. The density variables (for population and housing units) also incorporate the town size and the amount of developed land (from the land use data). Summary statistics on the census data are in the second group of Table 2.1.

Land Use Data

Land use data come from shapefiles provided by MassGIS. The data consists of polygons covering the entire state indicating the primary use of each tract of land. The year of data collection was 1971, four years prior to the Massachusetts Zoning Act. The land use categories were derived from aerial photographs taken from aeroplane and latter digitized. The category "undeveloped" includes pure undeveloped land (forests and open space) as well as agricultural land. "Non-developed" further includes water and wetland areas. Table 2.1 displays the summary statistics for the land use variables under the third group.

Massachusetts Bureau of Geographic Information

The geographic variables mainly come from Massachusetts Bureau of Geographic Information (MassGIS), the governmental portal for open-access data for the state. Distance to Boston is calculated from town shapefiles. Distance to the coast additionally uses shapefiles outlining the Atlantic coast. Share of aquifer coverage also comes from polygons outlining their extent from MassGIS. These variables are summarized in the last group of Table 2.1.

Wharton Residential Land Use Regulation Index

The first external index of zoning regulations is the Wharton Residential Land Use Regulation Index (WRLURI) from Gyourko et al. (2008). This index is based on a survey of 2,649 local jurisdictions across the US (79 in Massachusetts). The authors created the index by applying factor analysis on the responses to the survey. The survey asked questions on the involvement of local and state actors in the development process, supply and density restrictions, and on the project approval process.

Housing Regulation Database of Massachusetts Municipalities

The second source of information on the land use regulatory environment in Massachusetts comes from the Housing Regulation Database of Massachusetts Municipalities, a joint project between the Pioneer and Rappaport Institutes (PIRI, 2005). Based on data gathered from municipal bylaws and telephone surveys, they compiled a dataset describing the regulatory environment for 187 towns within a 150 mile radius of Boston.

As the dataset does not come with a measure of the overall regulation level, I derive one through Principal Components Analysis on the coded responses. These are the variables included in the PCA, and by extension in the index[4]:

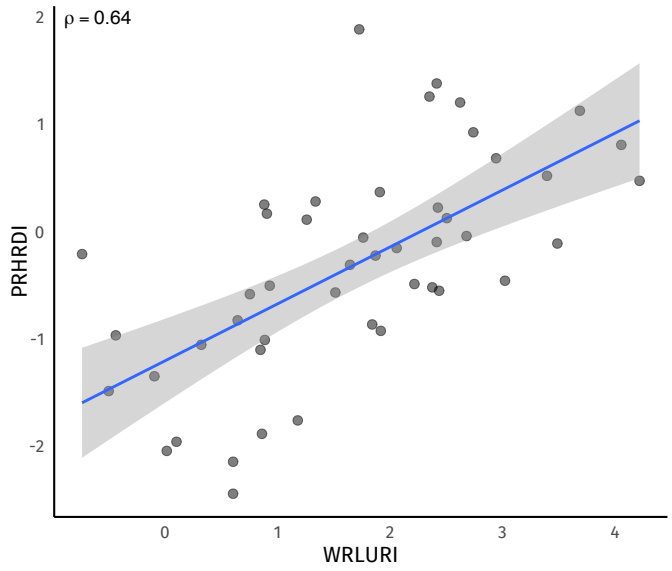
`ZONEDIST`	No. of zoning districts
`OVERLAY`	No. of overlay district
`RESDIST`	No. of residential districts
`ZONEWEB`	Indicator if zoning regulations on town's website
`MFALLOW`	Indicator if multifamily housing allowed
`CLUSTER`	Indicator if cluster zoning allowed
`INCLUDE`	Indicator if provisions for inclusionary zoning
`GROWRATE`	Indicator if targeted growth rates (in permits issued)

[4]See the codebook for the full description of the variables and the data.

MLAEXCLUD	Indicator if undevelopable land excluded from minimum lot size calculations
MLACBA	Indicator if portion of minimum lot size must be contiguous
SHAPRULE	Indicator if constraint to shape of lots
FRONT150	Indicator if frontage requirements larger than 150 feet (in residential areas)
MAXFRONT	Longest single-family frontage requirement in town
FRONTOUT	Indicator if portion of front lot may be excluded from frontage requirement

The variables are all standardized to have mean zero and standard deviation one before conducting PCA. The first component is then extracted and again standardized. I will refer to this as the Pioneer/Rappaport Housing Regulation Database Index (PRHRDI). The relationship between PRHRDI and WRLURI is shown in Figure 2.1.

Figure 2.1: Relationship Between Wharton and Pioneer/Rappaport Regulation Indices



Notes: Each point represents a town in Massachusetts that has data on both the Wharton Residential Land Use Regulation Index and an index derived via principal components analysis on the Housing Regulation Database of Massachusetts Municipalities ($n = 48$).

Natural Language Processing Zoning Index

The last index used as an outcome is the Natural Language Processing Zoning Index (NALPZ). The the creation of the index is described in detail in the companion paper. To summarize, the index is derived by estimating a latent Dirichlet allocation model

(see methods section) on the text from the zoning bylaws. This method allows for the creation of an index for every town with bylaw text available. This is the case for 341 towns in Massachusetts, out of a total of 351.

As the index is created from latent variables that have no inherent meaning, the WRLURI and PRHRDI discussed above are used to select the latent variable best explaining the level of regulation. Details on this point are also found in the companion paper.

Zoning Bylaws

The zoning bylaws come in variety of different formats (PDF, Word documents, webpages) and first needs to be transformed into a machine readable format. This is especially problematic when the bylaws are scanned PDFs and first need to be processed with optical character recognition software. The general processing strategy is explained in the companion paper. The result of the preprocessing is a document-term matrix (DTM) where the columns refer to unique tokens (generally words) and the rows documents (here the zoning bylaws for a town). The entries consist of either raw or normalized counts.

How these counts are normalized plays as big a rule in the predictive power of an estimator as does choosing an estimator itself. To get a better idea of how the text is processed, I will demonstrate with two example sentences on landscaping taken directly from the zoning bylaws of Lancaster.

i. `District boundary planting is required on any premises along the full length of any boundary abutting or extending into a Residential District`

ii. `Street planting is required for nonresidential premises abutting an arterial street, as designated on the Zoning Map`

Once these sentences are pre-processed,[5] the basic DTM will take the following form:

	abut	abutting	along	arterial	boundary	designate	district	...	nonresidential	planting	premise	require	residential	street	zoning
i.	0	1	1	0	2	0	2	...	0	1	1	1	1	0	0
ii.	1	0	0	1	0	1	0	...	1	1	1	1	0	2	1

[5] This primarily consists of removing stopwords (generally conjunctions) and punctuation, and reducing a word to its stem ("required" becomes "require", "abutting" becomes "abut")

Table 2.1: Massachusetts Towns Summary Statistics

	N	Mean	SD	Min	Max
Regulation Indices					
NALPZ[1]	341	0.00	0.94	−2.33	2.33
PRHRDI[2]	187	0.00	1.00	−2.45	2.45
WRLURI[3]	79	1.57	1.31	−0.74	4.80
Demographic Variables (1970 US Census)					
Population ('000)	351	16.21	40.03	0.05	641.07
Housing Units ('000)	351	5.39	14.23	0.06	232.45
% Rural	351	57.01	41.66	0.00	100.00
% Female	351	50.91	2.35	31.72	58.41
% Under 18	351	35.27	4.91	17.10	46.82
% Over 64	351	10.42	3.94	1.88	25.85
% Non-white	351	1.42	4.24	0.00	68.64
% Married	351	79.48	4.62	41.72	94.23
% Foreign	350	5.47	3.02	0.00	18.02
% College	350	7.22	5.02	0.00	29.88
Labour Force Participation Rate	350	72.89	4.98	52.49	85.46
% Poverty	350	6.99	3.95	0.00	29.91
% Vacant	351	13.03	17.52	0.74	81.30
% Owner Occupied	350	65.11	17.08	15.52	94.44
Home Value ($'000)	351	114.71	50.09	14.22	319.50
Population Density ('000/km^2)	351	0.45	0.94	0.00	8.29
Pop. Den. of Developed Land ('000/km^2)	351	1.06	1.04	0.03	8.34
Housing Unit Density ('000/km^2)	351	0.15	0.32	0.00	2.78
HU Den. of Developed Land ('000/km^2)	351	0.35	0.35	0.04	2.79
Land Use Variables					
% Water	351	3.00	3.30	−0.07	24.33
% Wetland	351	3.29	3.87	0.00	33.24
% Undeveloped	351	68.47	22.87	0.10	97.65
% Non-developed[4]	351	75.25	22.19	0.58	99.05
% Forest	351	57.74	21.63	0.00	94.16
% Agriculture	351	7.82	6.76	0.00	51.11
% Residential	351	16.87	14.77	0.74	70.37
% Commercial	351	1.29	1.92	0.00	12.80
% Industry	351	1.00	2.29	−0.06	21.24
% Transportation	351	1.24	1.90	−0.05	13.89
% Other Urban	351	4.97	5.39	0.04	32.21
Residential-to-All Developed Land	351	69.98	12.29	16.33	96.35
Commercial-to-All Developed Land	351	3.86	2.73	−0.09	16.50
Industry-to-All Developed Land	351	2.57	3.18	−0.38	21.46
Geographic Variables					
Town Area (km^2)	351	59.65	33.81	2.73	265.75
% Aquifer	351	10.99	16.34	0.00	97.36
Distance to Boston	351	76.21	51.02	0.00	197.61
Distance to Coast	351	52.51	55.29	0.07	180.85

[1] Natural Language Processing Zoning Stringency Index

[2] Pioneer/Rappaport Housing Regulation Database Index

[3] Wharton Residential Land Use Regulation Index

[4] Non-developed includes all undeveloped land plus water and wetlands.

With all the pages of the bylaws included, the number tokens runs into the thousands.[6] For this reason, standard regression techniques normally do not work as the number of tokens (variables) far exceeds the number of documents (observations).

Once the basic DTM has been derived, there are several normalization procedures that can be applied. Here is an overview of the most relevant.

Term Frequency The first aspect to consider is how the counts are handled. These counts are referred to as the term frequency. Often some tokens appear significantly more than others and therefore play an outsized role any predictions. To dampen this effect, the counts can be scaled down by either dichotomizing them or applying a logarithmic transformation. In the first case all counts that are zero remain zero, and all non-zero counts are set to one. In the second, the term frequency becomes $\widetilde{\text{tf}}_{vd} = 1+\log(\text{tf}_{vd})$, where the subscripts index the token (v) and the document (d).

Inverse Document Frequency Once it has been decided how to render the term frequencies, further scaling can be applied by multiplying the tf terms by the inverse document frequency. This is formally defined as:

$$\text{idf}_v = \log\left(\frac{1+D}{1+\text{df}_v}\right) + 1$$

where D indicates the total number of documents and df_v refers to the number of documents token v appears in ($\text{df}_v = \sum_d \mathbb{1}(v \in \mathcal{V}_d)$). Then, the weighted matrix elements are calculated by multiplying the tf with the idf:

$$\text{tf-idf}_{vd} = \text{tf}_{vd} \times \text{idf}_v$$

The purpose of this normalization is to give more weight to tokens that appear in fewer documents, thus potentially having more predictive power than tokens that appear in all documents.

[6] It is worth briefly discussing why "abut" and "abutting" are rendered as two different tokens in the table above. To stem a word you must first designate its part of speech (*eg* verb, noun). In Python the most common framework for this is `nltk`. This is not as trivial a task for a computer as it is for a human. The more advanced taggers use a combination of word endings (*eg* "-ing") and a statistical model of the tags of the preceding words to determine a word's part of speech. There exists no perfect algorithm for this process. In this case "abutting" was tagged (incorrectly) as a noun in the first sentence and a verb in the second.

n-grams Instead of generating a DTM with single-worded tokens, we can also group adjacent words together as one token. Continuing with the two example sentences above, we can generate a new DTM with two-worded tokens, resulting in the following matrix:

	(abut, arterial)	(abutting, extend)	(along, full)	(arterial, street)	(boundary, abutting)	(boundary, planting)	...	(require, nonresidential)	(require, premise)	(residential, district)	(street, designate)	(street, planting)	(zoning, map)
i.	0	1	1	0	1	1	$\cdots$	0	1	1	0	0	0
ii.	1	0	0	1	0	0	$\cdots$	1	0	0	1	1	1

Now neighbouring words are combined into unique tokens. For example, "district boundary planting" becomes "(district, boundary)" and "(boundary, planting)". This helps alleviate some concerns with the standard method which ignores all context (words that provide context such as adjectives are overlooked). However, this drastically increases the size of the matrix (along the "term" direction).

Document Normalization In the above example the sentences are approximately the same length. This is not the case for the zoning bylaws. Generally speaking, larger jurisdictions' bylaws have more pages. To mitigate longer documents having more *absolute* terms, we can normalize each document's vector of tf or tf-idf counts (the rows in the DTM). This results in only the *relative* occurrence of the terms distinguishing between documents.

There are two normalization methods considered here. The first divides each element in the DTM by the L_1 norm of the respective row, otherwise known as the Taxicab norm: $\|\boldsymbol{x}\| = \sum_i |x_i|$. More common, however, is dividing each element by its row's L_2 norm, or the Euclidean norm: $\|\boldsymbol{x}\|_2 = \sqrt{\sum_i x_i^2}$.

Max. Document Frequency Often there are domain-specific terms that appear across all documents that do not confer much information. For example, in the legislative case, the word "section" appears in virtually all bylaws, but would not provide much additional information. One way to filter out these cases is by setting the maximum document frequency. This sets a threshold for the number of documents containing a specific term, above which the term is removed entirely from the matrix. If the threshold is 95%, for example, and and the word "section" appears in 99% of the documents, then

it would be removed.

Min. Document Frequency The minimum document frequency is the inverse of the maximum document frequency. This specifies the threshold of a term's appearance across documents below which it is removed. This generally throws out highly specific words, such as proper nouns. For example, the zoning bylaws of Belchertown, Massachusetts will contain the word "Belchertown" often. This word will most likely not appear in other bylaws (with the notable exception of neighbours). These words are generally too specific to be useful in prediction exercises.

2.3 Methods

ML Methods

As there is little guidance from the literature, several different machine learning[7] methods are evaluated. The goal for any estimator is to best predict the level of regulation in a sample on which the model was not trained (out-of-sample prediction). 5-fold cross validation[8] is used both to evaluate the different models and select the best tuning parameters where necessary. With unsupervised models (*ie* the outcome variable is not part of the estimation procedure) the entire sample is used to evaluate fit as there is no risk of overfitting. The model is still estimated with $1/5$ of the sample left out.

Penalized Linear Models

The first class of machine learning methods consist of simple extensions to OLS. They have the immediate advantage that they are estimable when the number of variables (here: unique tokens) is larger than the number of observations. This is achieved by penalizing the incorporation of (and magnitude of) additional parameters. This is why they are often referred to as "shrinkage methods".

Lasso Lasso is the first penalized linear model considered. The objective function being minimized is similar to the one under OLS with a penalization term included. Formally, the coefficients are estimated as follows:

[7] The term "machine learning" is used quite broadly. Specifically, models that can deal with cases where the number of variables are larger than the number of observations are considered.

[8] k-fold cross validation is when the sample is first split into k subsamples. Then the model is subsequently estimated with one of the k subsamples left out. The out-of-sample prediction accuracy is calculated on the left out subsample (usually the mean squared error or R^2). The scores are averaged across the k-folds to arrive at the overall score for the estimator.

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.1}$$

where λ is the parameter that controls the amount of shrinkage. Due to the non-linear nature of the problem (caused by the absolute value penalizations), this estimator results in some coefficients being exactly zero. The non-zero coefficients will be shrunk towards zero.

The tuning parameter, λ, governs the degree of regularization (*ie* the penalty associated with the number and magnitude of the coefficients). This parameter is chosen based on 5-fold cross validation.

Ridge The second linear model is the ridge regression. It is quite similar to lasso, only instead of the L_1 penalization, it is replaced with a L_2 penalty. Concretely, the objective function is:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.2}$$

The problem is now linear and an explicit solution exists as the penalization term is the square of the estimated coefficients rather than their absolute values. All the coefficients are now shrunk towards zero, but not set exactly to zero.

On account of how ridge is estimated, it is efficient to simultaneously conduct leave-one-out cross validation to tune the parameter of λ. This is similar to k-fold cross validation, only instead of leaving out a fraction $n \times 1/5$ of the sample for each fold, one observation is removed at a time and the model is estimated n times. The out-of-sample prediction error is averaged across all n models to choose the optimum λ.

Elastic Net The last penalized linear model is the elastic net. It combines lasso and ridge by setting the penalization term to a convex combination of the terms from the other two methods. The penalization term is now:

$$\lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right) \tag{2.3}$$

with $\alpha \in [0, 1]$. It is easy to see that when $\alpha = 0$ we are back at lasso and with $\alpha = 1$ at ridge. Both the α and λ parameters are selected via 5-fold cross validation.

Decomposition Methods

The next class of ML techniques consists of decomposition methods, so called because they normally deal with matrix factorization. The aim is to reduce the number of dimension of a matrix down to the most relevant (with the exact definition of relevance depending on method being considered). These are generally unsupervised ML methods (in that they do not aim directly to predict an outcome). The PCA regression and partial least squares methods are the exceptions, and may be considered semi-supervised techniques.

Principal Components Analysis (PCA) The first decomposition method is PCA. It summarizes a matrix (X) with a specified number of components (m) that explain the most variation in the original data. Concretely, it finds m latent variables that take the following form:

$$Z_{im} = \boldsymbol{x}_i' \boldsymbol{\phi}_m \tag{2.4}$$

with the largest variation subject to $\sum_{j=1}^{p} \phi_{jm}^2 = 1$. The components after the first have the additional constraint that they must be orthogonal to the components that came before it. Of the m components, the one with the strongest correlation to the outcome averaged across the five folds is chosen as the predictor.

PCA Regression The PCA regression method is an extension to the baseline PCA to make it semi-supervised. Once the m principal components are extracted, they are then used in a standard OLS regression. Concretely, the regression takes the form of:

$$y_i = \sum_{j}^{m} \gamma_j Z_{ij} + u_i \tag{2.5}$$

Different to the standard PCA from above, all m components are included in the regression and therefore contribute to predicting the outcome. The exact number of components is selected through cross validation.

It is worth considering why the standard PCA would ever be preferable to PCA regression, given the extra regression step that will result in in-sample predictions being weakly better. The primary reason is overfitting: though the in-sample R^2 will necessarily be higher, it need not be for out-of-sample predictions.

Partial Least Squares (PLS) PLS is closely related to PCA. Whereas PCA seeks to maximize the variance of the estimated components, PLS simultaneously looks for solutions that have high variance and high correlation (with the outcome variable). Because PLS shares objects from OLS (maximize correlation) and with PCA (maximize variation), it can be thought of as a half-way point between the two methods (hence, the label of semi-supervised). In practice, the objective of maximizing variation dominates and the results are closer to PCA. Cross validation is used to choose the number of components.

As PLS is not a true unsupervised learning technique, it is often referred to as a cross decomposition technique. With this in mind, the results using PLS will be evaluated inline with supervised methods.

Singular Value Decomposition (SVD) Virtually identical to PCA,[9] with the exception that the input matrix does not need to be centred before computation (*ie* the variables do not need to be demeaned). The number of singular values chosen corresponds directly to the number of components to be extracted. This number is decided through cross validation.

Latent Dirichlet Allocation (LDA) The LDA technique is a multinomial mixed-membership model which assumes that the distribution of token counts arises from a specified number of latent "topics". A more comprehensive description of the model can be found in the companion paper. Though placed under the category of "decomposition methods", LDA has a statistical foundation and the process that is assumed to generate the distribution of tokens is modelled explicitly. There are three main tuning parameters for the model that need to be selected through cross validation. The first is the number of latent topics. The other two are parameters that govern the priors on the document-topic and topic-word distributions. The higher these two parameters are, the more concentrated the probability mass of a document (topic) is on specific topics (words).

[9] SVD can be considered more of a matrix algebra technique than a machine learning method. In fact, most computational implementations of PCA use SVD to derive the components.

Unlike with PCA, increasing the number of latent variables affects all estimated topics (*eg* the model estimated with two latent topics may be very different than one estimated with three). For each fold of every model specification, the out-of-sample prediction score is based on the latent variable that most strongly correlates to the response variable.

Decision Trees

The tree based methods begin by partitioning the predictor variable space into subspaces and assigning the mean of the observed outcome as the predicted value. The goal of these partitions is to minimize the residual sum of squares. Formally:

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{2.6}$$

where J is the total number of regions, R_j is the j^{th} subspace of the predictors, and $\hat{y}_{R_j}$ is the mean of the outcome variable in that subspace. Because it is computationally infeasible to consider every partition of the predictor space, most methods work sequentially. At each potential split, every predictor (p) is considered individually and the best split found. Of these p potential splits, the one that minimizes Equation 2.6 is chosen.

The standard method suffers from high variance and overfitting, resulting in poor out-of-sample predictions. To alleviate these concerns, the methods considered here are extensions of the standard decision tree. They built several trees and aggregate them under different assumptions to smooth out the variability from the individual trees.

Bagging The first extension to the standard decision tree is bagging. It is based on bootstrapping. Given B bootstrapped samples (with replacement), the bagging method averages the prediction values across the B estimated trees. If $\hat{f}(\boldsymbol{x}_i^b)$ is the decision tree from bootstrap b, then the bagging estimator is:

$$\hat{f}_{\text{bagging}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}(\boldsymbol{x}_i^b) \tag{2.7}$$

B is a tuning parameter that needs to be chosen via cross validation.

Random Forest The random forest method extends the bagging estimator by decorrelating the individual trees. One concern with bagging is that the first partitions in the predictor space will be dominated by the best predictors. This results in the individual trees, though from different samples, being very similar. Random forest overcomes this

issue by randomly selecting a fraction of all p estimators at each split for consideration. This ensures that some of the B trees will consider different, potentially relevant, variables at each split.

Just as with bagging, the random forest estimator requires B to be specified by the researcher. This is again decided through cross validation. The number of variables considered at each split, m, must also be set. I set this equal to $m \approx \sqrt{p}$, following standard practice (Friedman et al., 2001).

Boosting Boosting is an iterative method that fits subsequent decision trees[10] to the residuals from the previous iteration. This "slow learning" method is meant to focus on the unexplained variation at each loop. The residuals for iteration b can be expressed as follows:

$$r_{ib} = y_i - \lambda \sum_{j=1}^{b} \hat{f}_b(\boldsymbol{x}_i) \tag{2.8}$$

which are then used with the independent variables $\boldsymbol{x}_i$ in the next decision tree to estimate the next predictor function, $\hat{f}_{b+1}(\boldsymbol{x}_i)$. λ determines the rate of learning. Both B and λ are chosen from cross validation.

Variable Selection

Predicting the level of regulation from zoning bylaws is useful for towns where surveys have not yet been conducted. But it still leaves the question of what town characteristics best describe strongly and weakly regulated towns unanswered. Several variable selection methods are employed to highlight the features of a town that best distinguish between the level of regulation among towns.

Univariate F-Test

The first method to test the relevance of specific variables is also the most straightforward. It conducts a series of pairwise univariate regressions between the set of town variables and the chosen regulation index and calculates the F-statistic for the regressor. The relevance of the regressors for prediction are determined by the magnitude of the F-statistics.

[10] Though the method is not exclusive to decision trees and can be extended to most other classes of supervised learning models.

Forward Selection

Forward selection is an iterative procedure that sequentially selects the best predictor among a set of candidates through K-fold cross validation given a scoring function. The procedure is implemented here both with OLS and lasso as estimators, though it is compatible with other methods. The general procedure is:

1. Denote the set of selected variables as $\mathcal{P} = \{\}$ and set of potential variables as $\mathcal{C} = \{x_j\}_1^p$
2. (If lasso: select λ tuning parameter through 5-fold cross validation with all variables included)
3. Cycle through the set of candidate variables. For each $x_j \in \mathcal{C}$:
    a) Fit the model (OLS or lasso) through 5-fold cross validation on all the regressors in $\mathcal{P}$ and x_j. Calculate the out-of-sample (OOS) R^2 for each fold.
    b) Average the OOS score across folds
4. Select regressor with highest average score. Add to set of selected variables $\mathcal{P}$ and remove from set of candidate variables $\mathcal{C}$.
5. Repeat steps 3–4 until desired number of variables has been selected

Backward Selection

Backward selection operates under a similar logic to forward selection, only it starts with a full model (*ie* all variables included) and sequentially removes the variables with the lowest OOS predictive power. The procedure is as follows:

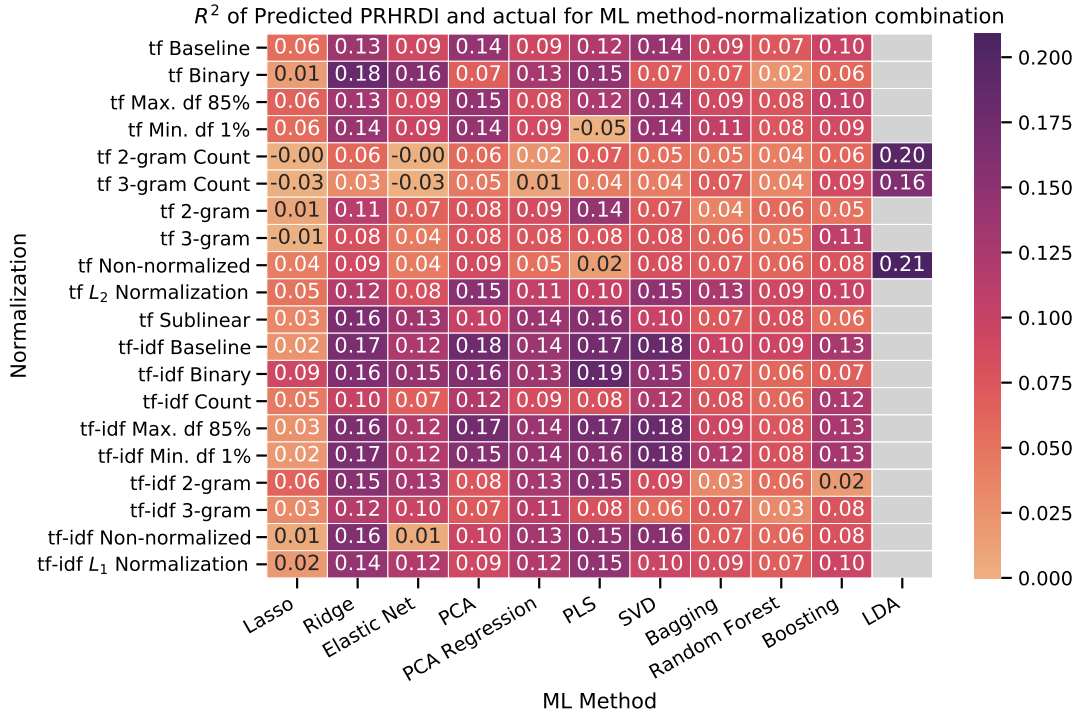
1. Denote the set of all variables as $\mathcal{C} = \{x_j\}_1^p$
2. (If lasso: select λ tuning parameter through 5-fold cross validation with all variables included)
3. Cycle through the set of variables. For each $x_j \in \mathcal{C}$:
    a) Fit the model (OLS or lasso) through 5-fold cross validation on all the regressors in $\mathcal{C}$ *without* x_j. Calculate the out-of-sample (OOS) R^2 for each fold.
    b) Average the OOS score across folds
4. Remove the regressor with the lowest average score from the set of variables $\mathcal{C}$.
5. Repeat steps 3–4 until desired number of variables has been selected

2.4 Results

ML Methods for Measuring Regulation

The first question that is addressed is this: are machine learning methods suitable to measure the level of regulation at the town level? Using different normalization procedures for the document-term matrix (discussed in Section 2.2), and a selection of different, relatively easy-to-implement ML methods (highlighted in Section 2.3), I first explore what combination best predicts out-of-sample regulation using the Pioneer/Rappaport Housing Regulation Database Index and the Wharton Residential Land Use Regulation Index as responses.[11] The results from this exercise are shown in Figures 2.2 and 2.3.

Figure 2.2: Predicting PRHRDI from Text: ML and Normalization Methods



Notes: tf Baseline: raw counts, unigrams, maximum document frequency 90%, minimum document frequency 3%, L_1 normalized. tf-idf Baseline: logarithmic counts, unigrams, maximum document frequency 90%, minimum document frequency 3%, L_2 normalized. Name of normalization indicates change from baseline. See Section 2.2 for details on document normalization.

Figure 2.2 shows the results with PRHRDI as the response variable. The vertical axis indicates the DTM normalization method, and the horizontal axis specifies the ML technique. The cell elements refer to the average R^2 value from 5-fold cross validation. When the ML method requires tuning parameters, these too are selected through 5-fold cross

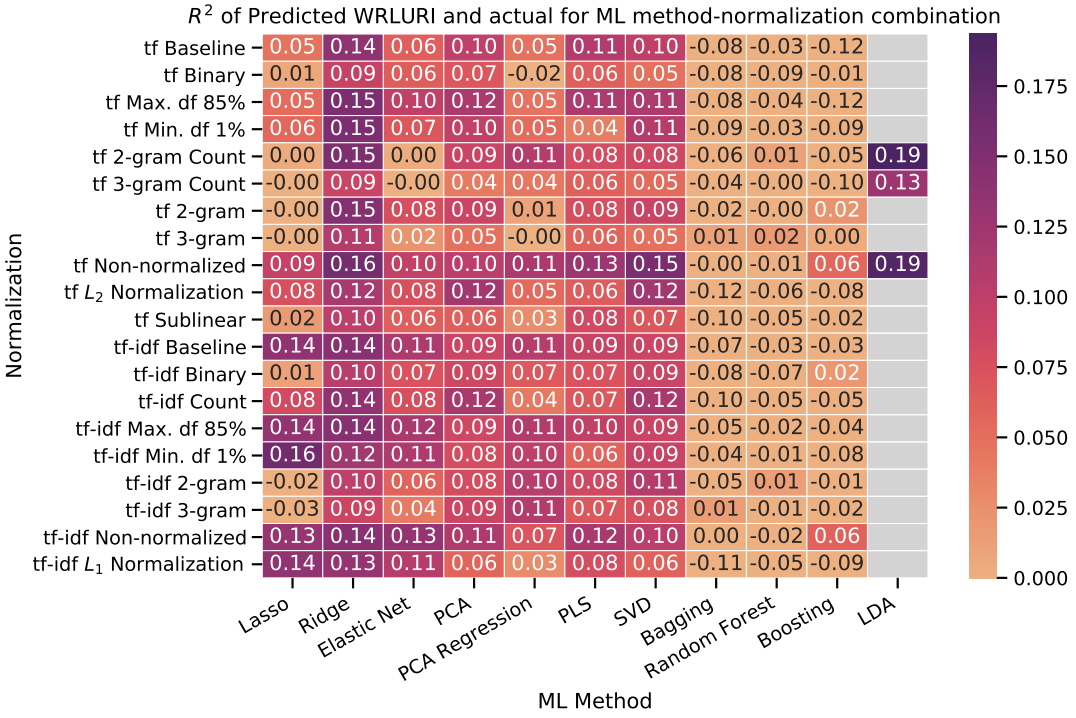
[11] The NALPZ is not considered as it is derived from the same text used to estimate these ML models.

validation. The results displayed always use the best parameter combination. Results for the latent Dirichlet allocation model are not shown for every normalization method as it requires a DTM with only count elements.

Clear patterns emerge. First, the matrix decomposition methods result in relatively better predictions. The LDA models seems particularly good at deriving a predictive measure of regulation, with singular value decomposition not much worse. Second, the decision tree-based methods have somewhat middling performance. Third, of the penalized linear models ridge regression is the most promising. Lasso performs relatively poorly, and elastic net is somewhere in between.

Turning to the normalization methods, DTMs subject to the inverse document frequency tend to perform better. Taking a logarithmic transformation of the term frequency (the baseline) also performs better than under the count or binary case. The clear exception to all of this is the basic count DTMs with the LDA model. The unigram version (the baseline) performs slightly better in this case than when using bigrams or 3-grams.

Figure 2.3: Predicting WRLURI from Text: ML and Normalization Methods



Notes: tf Baseline: raw counts, unigrams, maximum document frequency 90%, minimum document frequency 3%, L_1 normalized. tf-idf Baseline: logarithmic counts, unigrams, maximum document frequency 90%, minimum document frequency 3%, L_2 normalized. Name of normalization indicates change from baseline. See Section 2.2 for details on document normalization.

Figure 2.3 shows the corresponding results with the WRLURI as the response variable.

The first thing to note is that overall the WRLURI is more difficult to predict, no doubt in part due to the lower sample size (79 vs. 187). Second, the decision tree models are poor predictors. The OOS R^2 is often negative. This is probably owed to overfitting on the small-sized samples. Third, the LDA model again provides the best predictor of regulation. Ridge regression is very consistent across DTM normalization methods and also a decent prediction technique.

For researchers looking to derive measures of regulation from bylaws (at least for zoning) two primary conclusions can be drawn from these results. First, the latent Dirichlet allocation model does a decent job of uncovering a latent measure of regulation. The advantage of this method is that it does not require responses to calibrate the model. This is useful in situations where current measures or indices of regulation do not exist. The disadvantage of this model is that the researcher must select the latent category that best discriminates between strongly and weakly regulated localities. Second, when existing measures of regulation exist and one wants to extrapolate to other samples, ridge regression will preform well across a wide range of document-term matrix formats.

Best Predictors of Regulation

Having established what techniques work for measuring regulation, I now turn to the variable selection results. Given the variable selection procedures described in Section 2.3, I report the ten most highly predictive variables for each index. I consider both of the external indices of regulation, the PRHRDI and WRLURI, as well as the index derived through natural language processing on the zoning bylaws, the NALPZ. These are shown in Table 2.2.

Each column specifies a different variable selection procedure. The three different panels correspond to the three different indices of regulation. The rows convey the ranking of the most predictive variables of the respective index. To ease interpretation, the variables are colour-coded with accordance to their grouping in Table 2.1. Orange highlights variables that belong to land use variables and blue demographic variables. Looking across the different specifications, the land use variables are more often the best predictors of regulation. This is especially true of variables that relate in one way or another to the supply of developable land: the four variables that are found to be the most predictive (being ranked first in at least one specification) are the share of undeveloped or non-developed land, the fraction of forest cover, and the share of households considered to be rurally located. This is quite remarkable given that the indices are derived in different ways and that several variable selection procedures are considered.

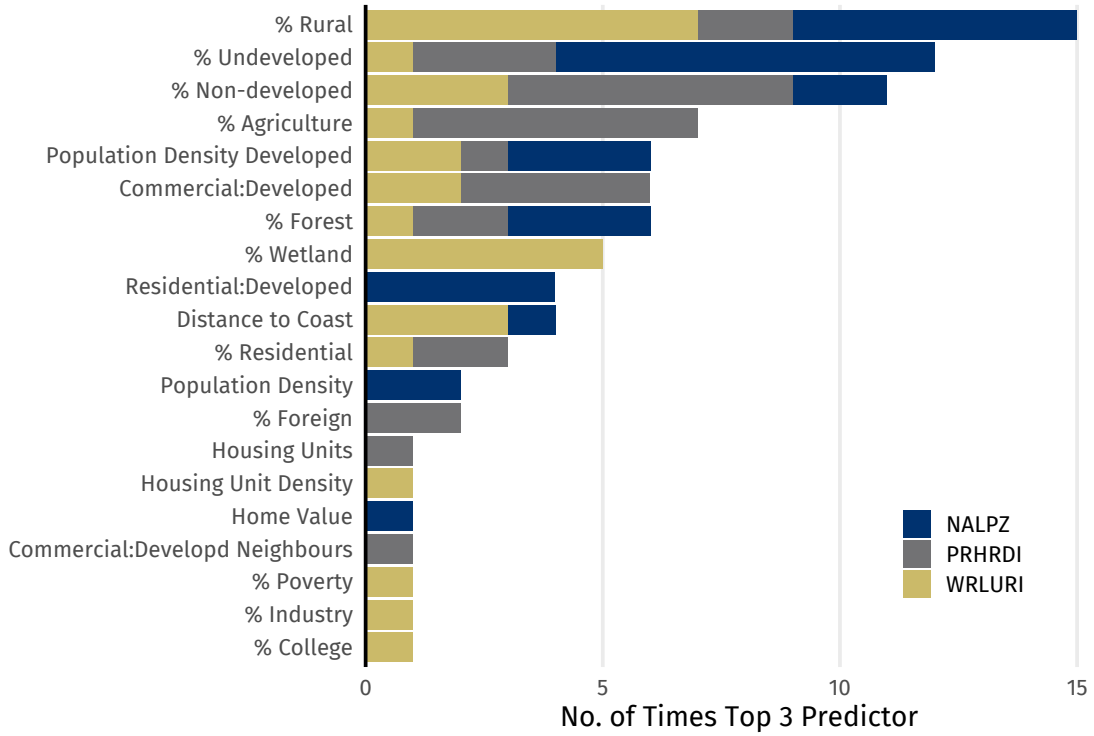
Table 2.2: What Town-level Characteristics Predict Regulation?: Variable Selection Procedures

		OLS		LASSO	
	Univariate F-Test	Forward Selection	Backward Selection	Forward Selection	Backward Selection
		Panel A: NALPZ			
1	% Undevel.	% Undevel.	% Undevel.	% Undevel.	% Rural
2	% Non-devel.	% Rural	Pop. Den. Devel.	% Rural	Pop. Den. Devel.
3	% Forest	Residential:devel.	Pop. Den.	Residential:devel.	Home Value NBR
4	% Rural	Dist. to Coast	Home Value NBR	Dist. to Coast	Residential:devel.
5	% Residential	Pop. Den. Devel.	Residential:devel.	Pop. Den. Devel.	Dist. to Coast
6	% Other Urban	Home Value NBR	% Wetland	Home Value NBR	Industry:devel.
7	Pop. Den. Devel.	Pop. Den.	% Water	% Forest	% Owner Occ.
8	% Forest NBR	% Owner Occ. NBR	% Non-devel.	Industry:devel.	% Rural NBR
9	% Undevel. NBR	Town Area	% Vacant NBR	% Owner Occ.	% Aquifer
10	% Commercial	% Non-devel.	Town Area	Pop. Den.	Population
		Panel B: PRHRDI			
1	% Non-devel.	% Non-devel.	% Forest	% Non-devel.	% Rural
2	% Undevel.	Comm.:devel.	Agriculture	Comm.:devel.	Comm.:devel. NBR
3	% Residential	Agriculture	Pop. Den. Devel.	Agriculture	No. of HUS
4	% Forest	Comm.:devel. NBR	Pop. Den.	Comm.:devel. NBR	Agriculture
5	% Commercial	% Rural	% Commercial NBR	% Rural	% College NBR
6	% Rural	% Foreign	% Residential NBR	% Foreign	% Foreign
7	Pop. Den. Devel.	% College NBR	Home Value NBR	% College NBR	Comm.:devel.
8	HU Den. Devel.	% Residential NBR	% Rural NBR	No. of HUS	% Over 64 NBR
9	% Other Urban	% Rural NBR	% Over 64 NBR	% Residential	% Married
10	% Residential NBR	% Over 64 NBR	% Commercial	% Over 64 NBR	HU Den. NBR
		Panel C: WRLURI			
1	% Rural	% Rural	% Forest	% Rural	% Non-devel.
2	Comm.:devel.	% Wetland	Dist. to Coast	% Wetland	Dist. to Coast
3	% Non-devel.	% Poverty	Agriculture	Comm.:devel.	% Rural
4	Pop. Den. Devel.	Dist. to Coast	% Undevel.	% Water	% Wetland
5	% Other Urban	% Non-devel.	% Wetland	% Foreign	Population
6	Population	% College	HU Den.	% Non-white	Industry:devel. NBR
7	HU Den. Devel.	HU Den.	HU Den. Devel.	LFPR	% Undevel.
8	% Commercial	Pop. Den. Devel.	% Water	No. of HUS	% Foreign
9	No. of HUS NBR	% Forest	% Foreign	Town Area	% Water
10	% Forest	% Commercial NBR	No. of HUS	Agriculture	% College

Notes: Each column specifies a different variable selection procedure (as described in Section 2.3). Each panel refers to a different regulation index. The rows indicate the rank of the selected variables in terms of predictive power. Orange coloured words are land use variables and blue coloured words are demographic variables. NBR: neighbours. LFPR: labour force participation rate.

Zooming in more on the best predictors, Figure 2.4 shows the number of times that a variable is one of the top three most predictive across all specifications. For example, the variable "share undeveloped" is a top three predictor three times for the WRLURI, six times for the PRHRDI, and two times for the NALPZ, for a total of twelve times. These results highlight exactly how much more predictive land use characteristics are than demographics. The share of the population that is foreign born, for example, is one of the best demographic predictors among all demographic variables, but only makes two top-three appearances and only for one index.

Figure 2.4: Variables Most Often Selected in Variable Selection Procedures



Notes: The length of each bar indicates the number of times the respective variable was one of the best three predictors for each variable selection-regulation index combination (column-panel combination from Table 2.2).

These results speak most in favour of historical patterns of land use and density being the primary determinant of regulation today. Other prominent theories, particularly home owners looking to maximize the value of their properties, or strategic sorting based on demographic characteristics, cannot be explained by these facts. This does not mean either of these two channels does not play a role, only that their effects are overshadowed by the supply of buildable land and pre-existing land use patterns.

Differential Trends in Demographics with Respect to Land Use Regulation

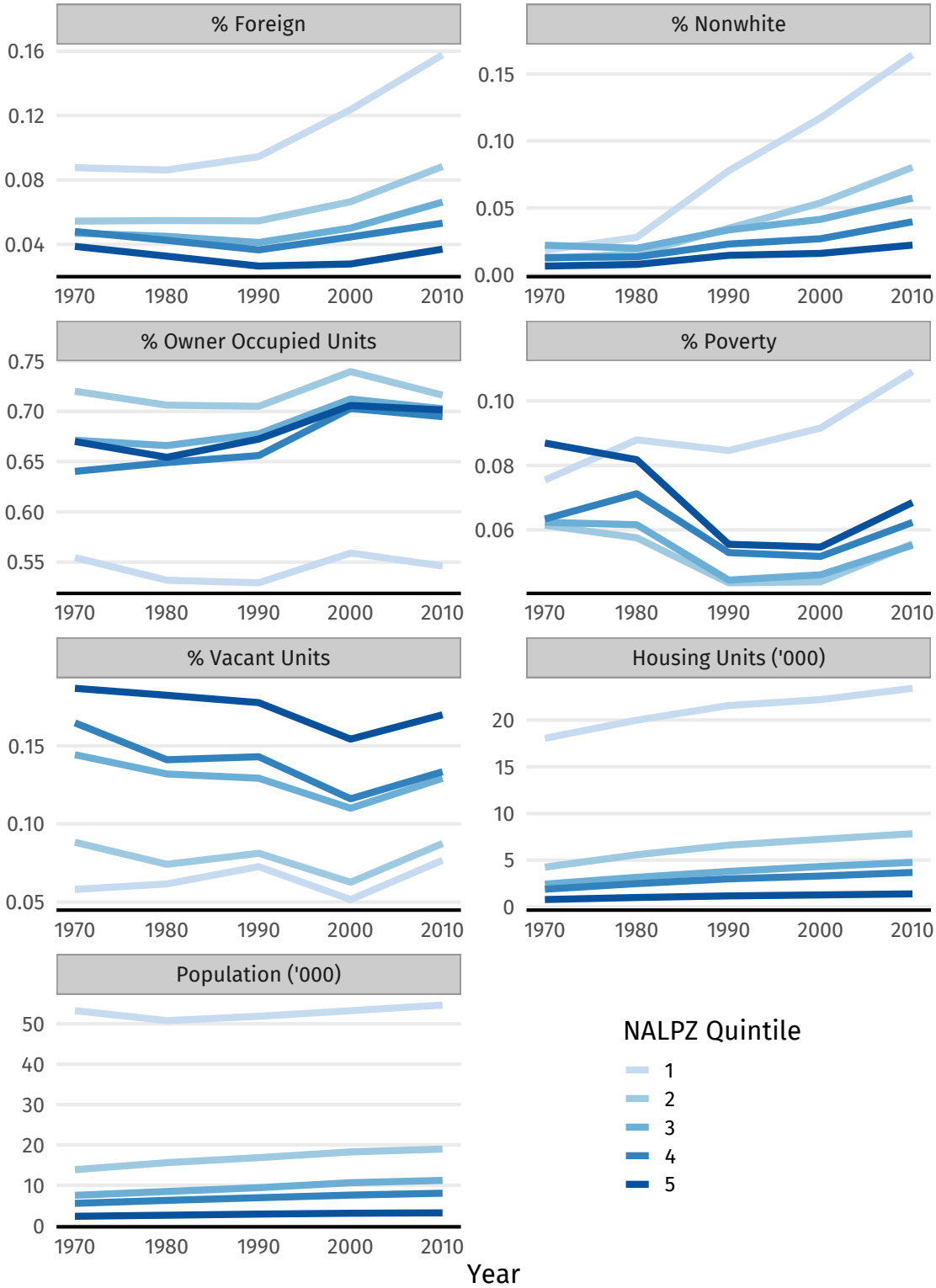
Most analyses of land use regulations have looked the the consequences at a given point in time: how do the attributes of localities with varying degrees of regulation look? I extend this by investigating how the demographic composition of towns in Massachusetts has changed over time differentially with respect to the degree of land use regulation. The main results of this exercise are shown in Figure 2.5.

The figure plots the average value of the respective demographic variable for towns within quintiles of land use regulation as measured by the NALPZ. Darker lines refer to towns in the top quintile (most regulated) and light lines to those in the bottom quintile (least regulated). It is apparent that there are important differences between towns with respect to land use regulations, both initially, before these regulations were widely adopted, and in the trends over time. Of standard demographic measures, the share foreign, in poverty, and non-white show the most heterogeneity with respect to regulation (other demographic variables such as age do not, and are not reported here). For example, in 1970 the difference between towns regarding the fraction of non-white residents was negligible, and low everywhere in the state. However, after 1980, this share increased significantly for towns in the bottom quintile of regulation and only very slowly for the other quintiles. There is a similar pattern with the share of foreign residents, though there were also initial differences. These patterns are no doubt driven in part on account of less regulated towns being closer to Boston, which attracts more immigrants. However, the overall population and number of housing units has remained remarkably stable over the same time frame. This suggests that part of the difference is due to sorting of residents between towns of varying levels of regulation.

Another important difference between the regulation quintiles is regarding the share of population considered in poverty. While in most towns this share has changed very little or decreased slightly, it has increased by over three percentage points for towns in the lowest quintile of regulation. This highlights an important fact: the least regulated quintile of towns, due to regulation or not, is very different from the other $^{4}/_{5}^{\text{ths}}$ of towns.

Two other noticeable differences between the regulation quintiles concerns housing. Housing units in towns in the bottom regulation quintile are much less likely to be owner occupied. In these towns only about 55% of units are owner occupied, a fraction that has been virtually unchanged over 40 years, while for other towns it averages between 65–70%. Another housing characteristic that varies between the quintiles is the share of housing units that are vacant. There is a clear gradient between the quintiles, with more regulated towns containing more vacant units. The gradient was apparent both before

Figure 2.5: Demographic Trends Among Quintiles of Regulation



Notes: Each subfigure plots the average value of the respective variable at the town level divided into quintiles based on the Natural Language Processing Regulation Index. Data from decennial US Census, 1970–2010.

and after the Massachusetts Zoning Act, with little variation over the decades. This suggests two possible explanations. The first is that towns with a large stock of unoccupied housing tend to regulate more strongly to deter more development to encourage use of the existing housing. The second is that investors or developers with unsold units lobby against further development to maintain the value of their already-built properties.

2.5 Conclusion

In spite of the variety, ubiquity, and impact of land use regulations, the causes and consequences of them are still poorly understood. As regulation is often the biggest constraint to new development, it is vital that we improve our knowledge of *why* these regulations exist in the first place, and *how* exactly they impact the housing market and beyond.

This paper contributes to this challenge by answering two related questions: i) what town attributes best predict strong regulation, and ii) how have differently regulated towns varied over time. In addition this paper tests various machine learning methods in their ability to predict an index of regulation restrictiveness using only the text from zoning bylaws.

I find that the latent Dirichlet allocation model, an unsupervised mixture model, performs the best in creating a predictor of regulation as benchmarked by survey-based indices. Among supervised methods, ridge regression is also a good and robust choice. Various feature selection procedures show that historical development (specifically the lack thereof) is the best predictor of the current level of regulation. Furthermore, investigating demographic trends among towns of differing levels of regulation shows that there were large baseline differences in the vacancy rate and share of housing units owner occupied before localized land use regulation. However, some town-level demographics characteristics, the fraction of non-white residents in particular, has diverged significantly since the use of widespread zoning.

Future research will need to incorporate these facts into a better understanding of the origins of land use regulations. Specifically: why do historical land use patterns best predict current zoning while certain demographic variables simultaneously vary substantially with the degree of land use regulation? Another promising avenue is to dissect the unidimensional measure of land use regulation used here to investigate whether certain facets of zoning are more relevant for different outcomes. The subindices that comprise the WRLURI (Gyourko et al., 2008) would be a good place to start.

Chapter 3

Identifying and Teaching High-Growth Entrepreneurship

Experimental Evidence from Academies for University Students in Uganda*

Abstract: We disentangle the extent to which entrepreneurial success can be attributed to skill formation and to selection. To study skill formation of nascent entrepreneurs among Ugandan university students, we randomly accept applications to a business training program fostering an entrepreneurial mindset. We measure labor market outcomes, business creation and success, and cognitive and non-cognitive skills as key outcomes up to three years after program participation. To better understand individual motivation for entrepreneurship, we experimentally vary marketing messages to all interested students prior to their application decision, emphasizing either entrepreneurial profit or entrepreneurial freedom. Lastly, we describe endogenous self-selection through non-experimental comparisons of key outcomes among applicants and eligible students from the same population who were aware of the entrepreneurship training program but did not express interest.

*This chapter is based on joint work with Vojtěch Bartoš, Kristina Czura, Michael Kaiser, and Timm Opitz.

3.1 Introduction

Entrepreneurship is key for economic development (Schumpeter, 1911). While most individuals in low-income countries are self-employed (e.g., 78.1 percent of the working population in Uganda was self-employed in 2019), these are mainly small-scale businesses that are only remotely related to the Schumpeterian entrepreneurship that drives economic growth (Porta and Shleifer, 2008; Hsieh and Olken, 2014). They typically lack capital and entrepreneurial ability, preventing them from reaping the full benefits of high-return investment opportunities (De Mel et al., 2012; Beaman et al., 2014; Bruhn et al., 2018). While relieving credit constraints shows some improvement in terms of business profits, it does not result in sustained business growth (Banerjee et al., 2015). Interventions aimed at improving business practices and managerial capital have not been shown to result in sustained increases in profits or employment (McKenzie and Woodruff, 2014). More promising approaches focus on the role of the psychology of entrepreneurship. Campos et al. (2017) show that training programs focusing on soft skill concepts, such as *personal initiative* and the *entrepreneurial mindset*, outperform programs teaching accounting, finance and marketing skills.[1]

Most business training studies target existing businesses—with the notable exception of Klinger and Schündeln, 2011, Blattman et al., 2014, and Premand et al., 2016—but neglect the importance of selection into entrepreneurship. Levine and Rubinstein (2017) and Levine and Rubinstein (2018) provide evidence that successful entrepreneurs in the USA are positively selected on human capital. Moreover, evidence from high-income countries shows that cognitive and non-cognitive traits predict entrepreneurial success (Andersen et al., 2014; Koudstaal et al., 2016; Levine and Rubinstein, 2017). Yet little is known on whether non-cognitive traits are shaped by entrepreneurial activity, or whether people select into entrepreneurship based on these traits. This distinction is important for policy. If relevant non-cognitive traits are malleable, this would favour programs aimed at developing an entrepreneurial mindset. If they are not, interventions designed to identify high-potential entrepreneurs would be more promising.

We seek to disentangle the extent entrepreneurial success can be attributed to skill formation and to selection. First, we causally identify the effects of a business training program, which develops an entrepreneurial mindset, on business creation and business performance. In our field experiment, training is randomly offered to university students in Uganda who had expressed interest in entrepreneurship, a suitable sample positively selected on human capital. Second, we study how selection into the entrepreneurship

[1]Entrepreneurial mindset is one's ability to spot and benefit from opportunities that are encountered in daily life. Personal initiative captures one's desire to proactively tackle problems (Frese et al., 2007).

training program varies by motives and personality traits. Using panel-data drawn from the same population, we document how students interested in entrepreneurship differ from those that are not with respect to socio-economic, cognitive and non-cognitive factors, as well as labor market outcomes, including self-employment. Third, we causally identify what motives draw students to entrepreneurship training.

We partner with a Ugandan organization, StartHub Africa, that provides extra-curricular entrepreneurship training academies at local leading universities. We track three semesters of training academies (henceforth "waves") conducted at eight to ten universities with a combined enrollment of around 2,000 students in our study sample.[2] Each wave consists of a marketing campaign, an application phase, and an entrepreneurship training academy. A wave begins with an untargeted marketing campaign to raise general awareness of the program. Then, to be eligible for the program, students must attend an information session that consists of short presentations that summarize the training program. This is also where the application forms are distributed.

Our experimental design relies on two sources of exogenous variation. First, we randomly vary the motivational message for becoming an entrepreneur that is marketed in the information session video presentations: financial gains or creative freedom. This allows us to causally identify the motivations of applicants. Second, among those who applied, we randomly offer admission to the program to identify the effect of being offered admission on business creation, survival and performance. We complement these analyses by documenting patterns of entrepreneurial self-selection by comparing applicants to those who were aware of the training program but did not express interest along several repeated measures of socio-economic indicators, personality traits and preferences.[3] The data collection effort includes surveys at different points in the self-selection and application process, as well as surveys administered both before and after the entrepreneurship training academies (Figure 3.1).

This study relates to four strands of literature. First, we contribute to the literature on entrepreneurship and business training in low-income countries by studying a unique sample of highly-educated, high-potential individuals (see Levine and Rubinstein (2017) and Levine and Rubinstein (2018)) who aspire to be entrepreneurs. Despite extensive research on business training interventions, there is a paucity of evidence on the effects of training on high-skilled youths. Interventions in low-income countries typically provide middle-aged, incumbent micro-entrepreneurs with education on busi-

[2]Two waves have been conducted to date. We plan to include one more wave. We will discuss the feasibility of this extension and base our power calculations both on the status quo and the planned implementation.

[3]We elicit data on the Big-5 personality traits, grit, personal initiative and aspirations. Further, we gather measurements of time and risk preference as well as individuals' degree of loss aversion.

ness skills and managerial capital, which have not been found to result in sustained increases in revenue, profits or employment (Bruhn and Zia, 2013; Hsieh and Olken, 2014; McKenzie and Woodruff, 2014; McKenzie, 2017; Bruhn et al., 2018; Rigol et al., 2018). This population, however, may lack the necessary skills for becoming successful entrepreneurs (Bjorvatn and Tungodden, 2010; Hurst and Pugsley, 2011; Levine and Rubinstein, 2018; Carlson and Rink, 2019) or may be unwilling or unable to change the way they run their businesses (Burmeister and Schade, 2007). With respect to our target population, the most closely related study is Premand et al. (2016) who analyze the inception of an official entrepreneurship track at universities in Tunisia. They document modest increases of one to four percent in self-employment rates but no effect on overall employment.[4] Our setting differs from theirs in that we study an extra-curricular program that is more likely to only attract the genuine subpopulation of those interested in pursuing entrepreneurship.

Second, we contribute to the literature on the entrepreneurial mindset. The entrepreneurship training program we study is based on a curriculum that aims to foster an entrepreneurial mindset and personal initiative. Campos et al. (2017) show that this type of training results in larger increases of profits than a traditional business training program. Ubfal et al. (2019) find transient, short-term effects of this type of training on micro-entrepreneurs in Jamaica. We complement this burgeoning literature by offering further evidence on the merits of non-traditional training programs and enhance it by focusing on nascent entrepreneurs who have been found to benefit from traditional training programs (see Klinger and Schündeln, 2011).

Third, we contribute to the literature on selection into entrepreneurship and predictors of entrepreneurial success. Levine and Rubinstein (2017) show that successful entrepreneurs select along both cognitive and non-cognitive dimensions. Evidence from high-income countries suggests that cognitive and non-cognitive traits are important predictors of entrepreneurial success (Andersen et al., 2014; Koudstaal et al., 2016; Levine and Rubinstein, 2017). For example, entrepreneurs are generally more risk-tolerant (Bouchouicha and Vieider, 2019) and display more overconfidence (Åstebro et al., 2007; Herz et al., 2014). Evidence is scarce on whether non-cognitive traits are shaped by entrepreneurial activity or whether people select into entrepreneurship based on these traits. On one hand, an established view suggests that preferences are relatively stable (Schildberg-Hörisch, 2018). There is however recent evidence that personality traits, such as grit, may be malleable –– at least among young adolescents (Alan

[4]This speaks to substitution from wage employment to self-employment, and does not imply overall employment effects. Alaref et al. (2020) present results from a medium term follow-up and show that any effects were short lived: four years after the program, there are no differences in self-employment and wage employment rates between the treatment and control groups.

et al., 2019). We extend this literature by documenting personality traits, preferences, and beliefs before individuals select into entrepreneurship, how these differ by interest in entrepreneurship, and by identifying how entrepreneurship training affects these characteristics.

Fourth, we speak to the motivations of becoming an entrepreneur, and whether selection patterns differ by motivation. A sparse literature using observational data from the USA stresses that non-pecuniary benefits, such as being one's own boss or having flexible working hours, play a first-order role for business creation decisions and that these independence-oriented workers are willing to forgo higher earnings from wage-employment (Hamilton, 2000; Hurst and Pugsley, 2011; Hurst and Pugsley, 2015). Guzman et al. (2020) and Ganguli et al. (2018) confirm the importance of motives and differential responses to monetary and non-pecuniary motives resulting in selection patterns into entrepreneurship competitions in randomized field experiments in the USA and the UK, respectively.[5] We complement this recent literature by identifying the differential selection decisions made by high-skilled youth in a low-income country using random variation in the salience of different motives for entrepreneurship.

3.2 Research design

Background

StartHub Africa (SHA) conducts the academy at local universities during the academic semester. There is one academy per university which has a target class size of 40 students that spans nine weeks with one three-hour session each week. The academy covers all stages of training for nascent entrepreneurs: developing a business idea, creating a prototype, and implementing the idea. In the curriculum developed by SHA, management skills, such as cost accounting, and basic principles of finance and marketing are included, but emphasis is placed on developing participants' personal initiative to foster their entrepreneurial mindset. In this respect the training program is similar to the program studied by Campos et al. (2017). Lecturers are encouraged to create an interactive atmosphere, and the standardized materials SHA provides to the instructors require active input from the participants. Finally, the curriculum contains a number of practical exercises outside of the classroom. For instance, students are taught basic principles of market research, then brainstorm product ideas and spend the rest of the session venturing out on campus to assess people's reaction to their product ideas. The training

[5] Ashraf et al. (2020) vary the salience of career incentives in a recruitment drive for public health workers in Zambia, and also show that the salience of motives affects selection patterns, and later, performance on the job.

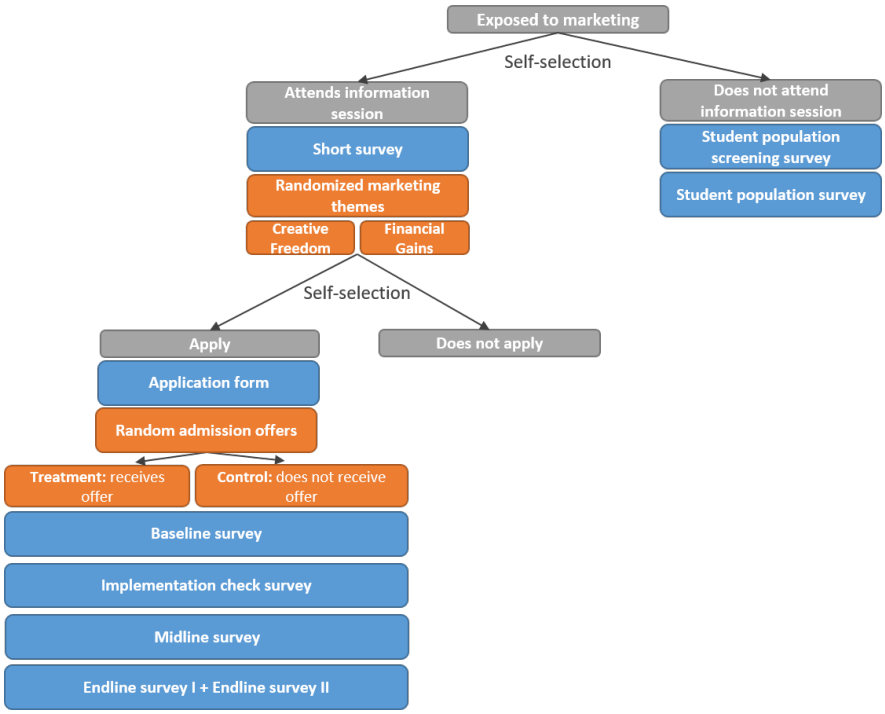
is taught by university lecturers or respected entrepreneurs from the local community that have been extensively trained and are continuously supported by SHA.

The academy is preceded by a marketing and application phase which spans the first three weeks of the semester. During the marketing phase, SHA creates awareness of the program using posters and flyers across campus, and in short pitches in classrooms and at campus events. Students are informed that attending an information session is a prerequisite for applying. Six to twelve of these 30-minute sessions are held per day over two or three days in a central location at each university. The information sessions provide detailed information on the academy's content, the expectations of the participants, in particular the time commitment necessary to complete the academy, success stories from previous participants, and the possibility to ask questions to SHA staff. To harmonize the information sessions as much as possible, the same SHA staff hold the information sessions throughout each day. Moreover, the presentations are video-based and contain the same structure: motivation for the academy, details, deliverables and requirements of the academy, and success stories from alumni. After the information session, students could pick up an application form in person, fill it out (in 10 to 15 minutes) and return it either to the team conducting information session, or to a well-know place on campus indicated on the application form. Application forms were only available to participants of the information sessions.

Experimental design

We exploit two sources of exogenous, experimental variation. First, in the *entrepreneurship training experiment*, admission to the academy is randomly offered to a subset of applicants. We use this variation to estimate the causal effect of being offered admission to the academy on entrepreneurial activity and economic outcomes. We also investigate effects on cognitive and non-cognitive skills. Second, to understand in more detail the characteristics and motivations of these young entrepreneurs, we add a second layer of exogenous variation: In the *selection experiment*, we randomly vary whether marketing for the academy emphasizes financial independence or creative freedom as motivation for becoming an entrepreneur. This variation allows us to identify how motivation impacts the application decision and to study heterogeneous effects based on individual characteristics. Figure 3.1 presents the experimental design. Finally, using a sample drawn from the same population, we document endogenous self-selection by comparing eligible students who did not express interest in the academy to applicants. We also investigate how key outcomes from the *entrepreneurship training experiment* evolve differentially over time between students who did not express interest to those who applied

Figure 3.1: Experimental design and data collection



Notes. Different phases of the experimental design and self-selection decisions is marked in grey, exogenous experimental variation is marked in orange, and data collection is marked in blue.

but did not receive training.

We first discuss the *selection experiment* and the complementary observational examination of self-selection, and then the *entrepreneurship training experiment* because this follows the chronological journey of a student from hearing about the training to submitting an application and possibly being offered admission. Nonetheless, the main research question draws on hypotheses about the *entrepreneurship training experiment*. The sample selection procedure will be detailed in Section 3.3.

Understanding selection and motives

The first layer of experimental variation is induced by randomly exposing clusters of students to different marketing messages during the information sessions. In the *selection experiment*, the content of two motivational video presentations is randomly varied between emphasizing i) that entrepreneurship offers the possibility of achieving financial independence, and ii) that entrepreneurship offers the freedom to be creative. For this, the respective motives are varied in four of the twelve overall slides reiterating the ben-

efits of becoming an entrepreneur and in the corresponding voice-over of these slides.[6] Everything else is kept constant. Support staff ascertained that no student listened to two information sessions by either staying in the room for the next session or entering early during an ongoing session. This exogenous variation allows us to cleanly identify how the pool of applicants differs across these two messages.

To analyze selection into the academy, we compare students who are interested in entrepreneurship, indicated by applying to the academy, with those who are not interested in entrepreneurship indicated by being aware of the entrepreneurship academy and not attending an information session. We refer to this latter group as the *non-interested subpopulation*. In other words, conditional on having been exposed to the marketing phase, we investigate what drives certain individuals to opt-in to the academy.

Entrepreneurship training experiment

The *entrepreneurship training experiment* allows for causally estimating the effect of the academy on individuals' self-employment probability, as well as on labor market outcomes and personality traits. Having participated in an information session, students decide whether to apply to the academy. A random sample, stratified by year and field of study, is then drawn from the set of all applications and offered admission to the training program — the treatment group. The remainder is placed into the control group.

Hypotheses

Grounded in the results of previous work, there are several hypotheses we seek to test. The first set of hypotheses concerns the effects of entrepreneurship training on economic and business outcomes and inputs. First, as shown by Klinger and Schündeln (2011) for a traditional entrepreneurship training program, we hypothesize that participating in the entrepreneurship academy fosters business creation. Yet, as our entire sample consists of highly-educated students that are all interested in entrepreneurship, we may not find significant differences between treatment and control groups at the extensive margin. Therefore, we further hypothesize that participation in the academy will improve business performance, captured by indicators such as monthly sales and profits, measures of capital and labor input, and measures of general economic self-sufficiency. One particular dimension we are interested in is labor input, and whether treated subjects create jobs through the businesses they create. The hypotheses are summarized in Ta-

[6] In Appendix Section C2 we present in detail how information sessions differed across the two marketing themes.

ble 3.1, Family 1.1. Positive findings for these hypotheses would provide evidence for entrepreneurial activity being teachable.

Second, we seek to identify channels through which the entrepreneurship training effects the primary outcomes of business creation and performance. Campos et al. (2017) find that a personal initiative training program can deliver lasting improvements for small business owners and they identified several channels: application of successful business practices, increased personal initiative, increased capital and labor inputs, substantial innovative activity (e.g., in the form of new products originating from own ideas) and product differentiation. We therefore hypothesize that participation in the academy leads to implementing more successful business practices, improved financial professionalization, marketing activities, product and process innovation, and better access to business networks. The hypotheses are summarized in Table 3.1, Family 1.2, Hypotheses 1 to 6. Finding effects along these dimensions would lend evidence to the most effective channels through which entrepreneurship training impacts the economic outcomes listed in Family 1.1.

Moreover, as laid out before, there is evidence that entrepreneurs are positively selected on cognitive and non-cognitive traits. Little is known, however, about whether non-cognitive traits may be shaped beyond adolescence. We therefore test hypotheses that investigate whether participating in the academy shapes non-cognitive traits. These hypotheses are summarized in Table 3.1, Family 1.2, Hypotheses 7 and 8. These hypotheses allow us to test whether — and to what extend — non-cognitive traits are malleable through participation in entrepreneurship training.

The second set of hypotheses concerns selection into entrepreneurship. First, individuals may have different motives for desiring to be an entrepreneur. Guzman et al., 2020 study entrepreneurs and find that women and individuals located in more altruistic cultures are motivated more by social-impact messages than money, whereas men and those in less altruistic cultures are motivated more by money than potential social-impact. Ganguli et al., 2018 document a crowd-out between extrinsic, cash-based and intrinsic, social motives for social entrepreneurs. While extrinsic motivational messages affect effort in applications for a start-up grant, it reduces the pool of applicants at the same time. Further, business success was less likely: social entrepreneurs motivated by extrinsic messages worked fewer hours per week, created fewer employment opportunities, and profited less from their venture. We therefore test which marketing message attracts more applicants: whether monetary motives or the promise of independent work better draws young, highly-educated individuals to entrepreneurship. We also investigate the types of individuals that are drawn to the different marketing messages. We consider measures of average cognitive ability, over-confidence and entrepreneurial

self-assessment. These hypotheses are summarized in Table 3.1, Family 2.1, Hypotheses 1 to 4. These hypotheses test whether stressing different motivations for becoming an entrepreneur lead to differential application patterns, both in terms of the quantity of applications and the attributes of the applicants themselves. Finding differences between the two messages would also speak to how different motivations to undertake entrepreneurship training are correlated with certain individual characteristics, and how such motivations shape the composition of applicants.

Further, we document selection into entrepreneurship (as proxied by selection into the academy) by comparing those that applied to the academy to those that were exposed to the marketing campaign but did not apply for the program (*non-interested subpopulation*). The outcomes of interest are listed under Hypothesis Families 2.2.1 and 2.2.2, and mirror those in Hypothesis Families 1.1 and 1.2 from the primary outcomes of the *entrepreneurship training experiment*. Comparing baseline characteristics and outcomes between the two groups allows us to identify the dimensions on which individuals select into entrepreneurship. Those and additional measures are investigated at endline to document how the non-interested subpopulation evolves over time compared to those that applied to the training and were not admitted (control group).

The outcome variables and their measurement are detailed in Section 3.3, while the empirical analysis is detailed in Section 3.4. Our results will inform to what extent teaching entrepreneurial skills and selection are important aspects for entrepreneurship. This is interesting from an academic perspective as it addresses fundamental questions on skill formation and its potential repercussions for entrepreneurship. It is also of utmost importance for policy: If entrepreneurial skills can indeed be formed, we offer an evaluation of a cost-effective, relatively easy to implement, and scalable intervention for high-potential, well-educated individuals. We can also document whether the nascent entrepreneurs originate from high-skilled individuals that would otherwise be unemployed or whether they are substituting away from formal-employment. If selection is found as relatively more important for entrepreneurial success, our study would inform policy makers that identifying high-potential entrepreneurs is of first-order importance (see McKenzie (2017) and Rigol et al. (2018) who seek to identify high-potential entrepreneurs, and Shane (2009) who warns about dragging people into risky, non-growth entrepreneurship). Our results would also offer some guidance on the motives that attract these entrepreneurs-to-be.

Table 3.1: Overview of hypothesis families.

Family	H #	Hypotheses title	Index	Data collection				Sample	Exogenous variation
				Baseline	Midline	Endline	Short		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. Entrepreneurship training									
1.1. Economic outcomes (primary)	1	Business creation	✓	✓	✓	✓		Applicants	Random assignment to training (*entrepreneurship training experiment*)
	2	Business success	✓		✓	✓			
	3	Capital and labor input	✓		✓	✓			
	4	Economic self-sufficiency	✓		✓	✓			
1.2. Business and personal input (secondary)	1	Business practices	✓		✓	✓		Applicants	Random assignment to training (*entrepreneurship training experiment)*
	2	Financial professionalization	✓		✓	✓			
	3	Marketing	✓			✓			
	4	Innovation	✓			✓			
	5	Networks	✓	✓	✓	✓			
	6	Entrepreneurial mindset	✓	✓	✓	✓			
	7	Owner's non-cognitive traits	✓	✓	✓	✓			
	8	Preferences	✓	✓	✓	✓			
2. Selection									
2.1 Selection into entrepreneurship (primary)	1	Submitted application					✓	Attendees at Info Sessions	Random assignment to marketing messages (*selection experiment*)
	2	Cognitive ability					✓		
	3	Over-confidence	✓				✓		
	4	Entrepreneurial self-assessment	✓				✓		
2.2.1 Economic outcomes (non-experimental)	1	Business creation	✓	✓		✓		General population (exposed to marketing campaign)	
	2	Business success	✓			✓			
	3	Capital and labor input	✓			✓			
	4	Economic self-sufficiency	✓			✓			
2.2.2 Business and personal input (non-experimental)	1	Business practices	✓			✓		General population (exposed to marketing campaign)	
	2	Financial professionalization	✓			✓			
	3	Marketing	✓			✓			
	4	Innovation	✓			✓			
	5	Networks	✓	✓	✓	✓			
	6	Entrepreneurial mindset	✓	✓	✓	✓			
	7	Owner's non-cognitive traits	✓	✓	✓	✓			
	8	Preferences	✓	✓	✓	✓			

Time frame

The proposed project consists of three waves of entrepreneurship training academies. Each wave consists of the implementation of the entrepreneurship academies, the experimental variation introduced in both the *entrepreneurship training experiment* and the *selection experiment*, and the data collection before and after the intervention. As detailed below, there will be a baseline survey, an implementation check survey (one to two months after the intervention), a midline survey (six months later) and two endline surveys. The Endline Survey I takes place 12 months after the intervention, the Endline Survey II 24 months after the intervention of the last wave.

The first wave started in September 2019, and the second wave started in January 2020. The third wave is scheduled for September 2020. The Endline Survey I will take place in February 2021 (Wave I), July 2021 (Wave II), and February 2022 (Wave III). The Endline Survey II is scheduled for February 2023 for all three waves. We expect to finish the analysis in the summer of 2023. Table 3.2 sets out the detailed time line for all steps in all waves. The implementation of Wave I and Wave II is already in progress, while the data collection for the midline survey (Wave I) and the implementation check survey (Wave II) in 2020 is scheduled. Later data collection and the implementation of Wave III is planned.

Due to the current Covid-19 crisis, we may not be able to implement Wave III as planned in September 2020, but have to postpone it to the spring semester 2021. In this case, all of the following dates will be postponed by around six months. In the worst possible case, we may not be able to implement Wave III at all. Although we deem this highly unlikely, we are conservative in the statistical power calculations below and account for a worst-case scenario with only the two already implemented waves and a base-case scenario with all three planned waves. The Covid-19 crisis will not have any effect on the scheduled data collection as only the endline survey will be conducted as an in-person survey, all other surveys are conducted via telephone.

Treatment assignment and statistical power

Selection experiment

Each information session presenter was provided with a randomly drawn marketing theme — financial independence or creative freedom — for the first session of the day. This was randomly chosen by the research team using a fair coin. The themes for the remaining sessions were then alternated by the presenter.

Table 3.2: Timeline

Stage/Instrument	Sample	Status	Date
Piloting	3 academies, 380 applicants	Completed	March-May 2018
Wave I	10 academies		
Marketing / information sessions / short surveys	$n = 1019$	Completed	Aug.- Sept. 2019
Application data / Baseline survey	$n_{\text{app}} = 713$, $n_{\text{base}} = 672$	Completed	Aug.- Sept. 2019
Enterpreneurship academy	$n = 414$	Completed	Aug. 2019 - Jan. 2020
Implementation check survey	$n = 625$	Completed	Jan. - Feb. 2020
Midline survey		Scheduled	Sep. - Oct. 2020
Endline survey I&II		Planned	Jan. - Feb. 2021 & Jan. - Feb. 2023
Wave II	8 academies		
Marketing / information sessions / short surveys	$n = 760$	Completed	Feb. - March 2020
Application data / Baseline survey	$n_{\text{app}} = 584$, $n_{\text{base}} = 562$	Completed	Feb. - March 2020
Enterpreneurship academy	$n = 313$	In process	Feb. - July 2020
Student screening survey	$n = 926$	Completed	Feb. - March 2020
Student population survey I		In process	May- June 2020
Implementation check		Scheduled	July - Aug. 2020
Midline survey		Planned	Jan. - Feb. 2021
Endline survey I&II		Planned	July- Aug. 2021 & Jan. - Feb. 2023
Student population survey II		Planned	Jan. - Feb. 2023
Wave III	9 academies		
Marketing / information sessions/ short surveys		Planned	Aug.- Sept. 2020
Application data		Planned	Aug.- Sept. 2020
Enterpreneurship academy		Planned	Aug. 2020 - Jan. 2021
Student screening survey		Planned	Sep. - Oct. 2020
Student population survey I		Planned	Oct. - Dec. 2020
Implementation check		Planned	Jan. - Feb. 2021
Midline survey		Planned	July - Aug. 2021
Endline survey I&II		Planned	Jan. - Feb. 2022 & Jan. - Feb. 2023
Student population survey II		Planned	Jan. - Feb. 2023

Notes. Midline survey of Wave I is scheduled for September and October 2020 due to time lags in the disbursement of research funds that have been fully secured in June 2020. Wave III is planned for the fall semester 2020/2021. Due to Covid-19, Wave III may have to be postponed to the spring semester 2021. All following dates will be postponed by around six months in this case. In the worst-case scenario of no possibility to implement Wave III, endline survey II will be conducted in July and August 2022 for Wave I and II.

Entrepreneurship training experiment

The randomization procedure offered admission to the training program to individuals with complete applications. Within *each* training cohort (i.e., university-semester), the target was to offer admission to 40 students, an optimal classroom size determined by SHA.[7] We targeted a control group of equal size; however, the group sizes were constrained by the number of applications received.

Thus the treatment and control group sizes were a function of the number of applicants. Specifically, if there were over 120 applications, we picked 45 students at random and offered admission, assigned 75 to the control group and omitted the remaining students from the study.[8] We anticipated low demand in some training cohorts and chose

[7] SHA allowed for deviations from the optimal size within a range of between 30 to 45 students. In case of excess (insufficient) interest, the classes were larger (smaller).

[8] This is done due to capacity and resource constraints. In practice, it is rare to receive over 120 applications for an academy.

to over-sample the control group when possible; in case of low demand, having a sufficiently sized treatment group took priority. When we received between 85 and 120 applications, 45 students were randomized into the treatment group, and the rest was assigned to the control group. In case of 80 to 85 applications, we assigned 40 students to control and offered treatment to the remaining ones. Finally, if there were less than 80 applications we offered treatment to $n_T = min[\mathsf{n}_{\texttt{applications}}, 40]$, and assigned $\mathsf{n}_{\texttt{applications}} - n_T$ to control.[9]

Having chosen the experimental group sizes, we implemented the following randomization algorithm which stratifies along two dimensions. First, we grouped students according to how many years they had studied their current degree. This is top coded at three years as this is the modal number of years students require to complete a Bachelor degree.[10] The rationale for this is that students who are close to graduation are more likely to move into (self-) employment in the near future. Second, the algorithm ascertains that the share of business students (students who study business, management, finance, marketing or related fields) is balanced between treatment and control *within* each year of study. Students' prior knowledge about business and entrepreneurship concepts may interact with the training content and business students' responses to the training program would systematically differ vis-à-vis non-business students.

We form six cells based on the program of study: business-related (two dimensions: yes or no), and years into the program (three dimensions: one, two or three years). We first use both cells for third-year students, and within each assign an equal number to either treatment or control. This ensures that all applications from third-year students are used.[11] We then applied the same procedure to second-year students. If not all applications from second-year students were necessary to complete target group sizes, we chose a subset at random. Finally, if group sizes were still not exhausted, we included (a random subset of) first-year students.[12] The exact same procedure will be used in Wave III.

Our calculations show both the worst-case scenario, in which we cannot implement

[9]Note that the second term can be zero if less than 40 applications are received.

[10]Most applicants are Bachelor students (≈ 87 percent) and those that are not are almost exclusively enrolled in "certificate" and "diploma" programs, which can either be a preparatory or supplementary degree. These usually take two years and can precede or follow a Bachelor degree.

[11]In theory, it would be possible to receive applications from third-year students in excess of the experimental group sizes. In such cases, we would have randomly picked the respective number. In practice this was never the cases.

[12]As an example, suppose there are 80 third year applicants; 56 in business-related degrees, 24 in non-business related degrees. The procedure allocates 28 of the business students to *each* of treatment and control; similarly, 12 of the non-business students would be in *each* of treatment and control. Overall, there would be 40 students in treatment and 40 in control, but the shares of business and non-business students would be equal across the groups.

the planned third wave at all, and the base-case scenario, in which we proceed with our project as planned or with some delays. To benchmark the statistical power of detecting effects of the training program on business success, we are conservative and present minimum detectable effect sizes based on the actual training cohort sizes from the first two waves of training conducted in the fall of 2019 and the spring of 2020 as the worst-case scenario. We further provide power calculations for various scenarios of attrition and non-compliance given the realized sample size.

During the first two waves we worked with 18 cohorts, meaning 18 university-by-semester blocks. There are 727 and 497 students in the treatment and control groups respectively. This corresponds to an average treatment group size and control group size of 40.4 and 27.6, respectively, and 68 students per cohort in total.

To incorporate myriad factors such as attrition, non-compliance, varying treatment and control group sizes into the power calculations, we perform simulations. We specify a data generating process and set the magnitude of our treatment effect to be equal to a pre-specified percentage of the standard deviation of a generic outcome; this can be interpreted as an effect size in percentage terms. This maps well into our strategy to deal with concerns from testing multiple hypotheses which rests on constructing normalized indices of our outcome variables with a mean of zero and a standard deviation of one.[13]

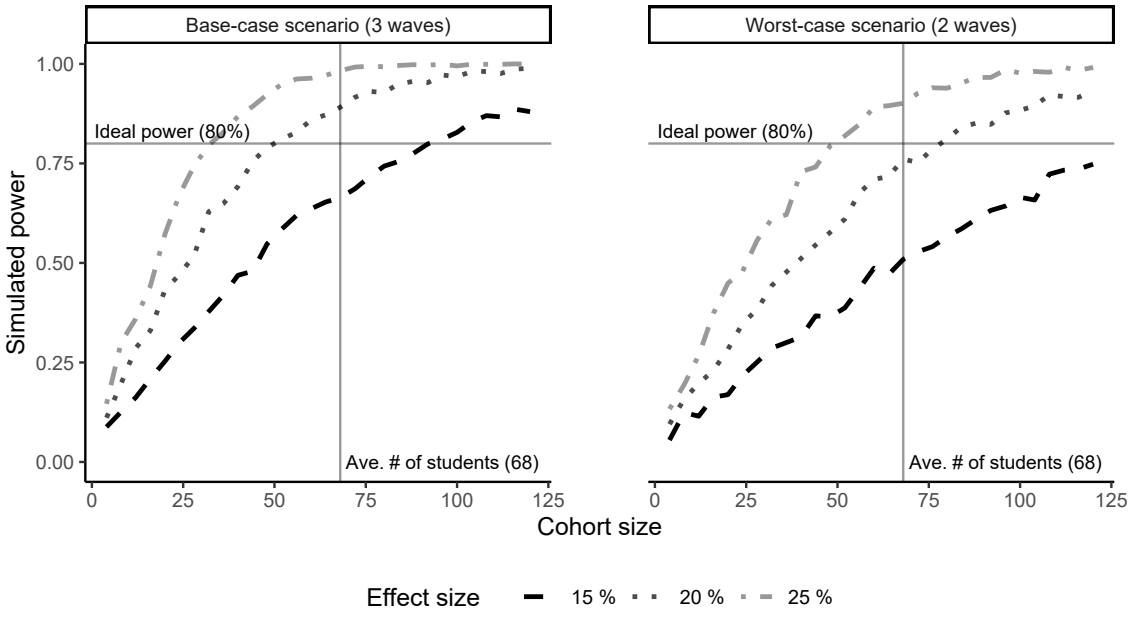
For the simulations, we estimate the primary specification (see Equation (3.1)) in a simulated sample and conduct a two-sided t-test of the null hypothesis of a zero effect of the treatment using standard errors that are robust to heteroskedasticity for inference. For the simulated sample, we set the number of cohorts, rates for attrition, non-compliance, and percent of sample treated as specified in the next paragraph. Then we vary the sample size per cohort starting from four, going until 122 in steps of four.[14] We draw 1,000 simulation samples per sample size considered. Across all simulated samples, we calculate the share of rejected null hypotheses at $\alpha = 0.05$ which is the measure of simulated power.

The simulation results are shown in Figure 3.2. Panel a presents the base-case scenario based on three waves of academies (left panel) and the worst-case scenario based on the two waves of academies that have been implemented already. We set the following parameters for our benchmark simulations: attrition rate of 5 percent, corresponding to twice the actually observed attrition in the implementation check of the first wave in fall 2019; a non-compliance rate of 25 percent as calculated based on the attendance data

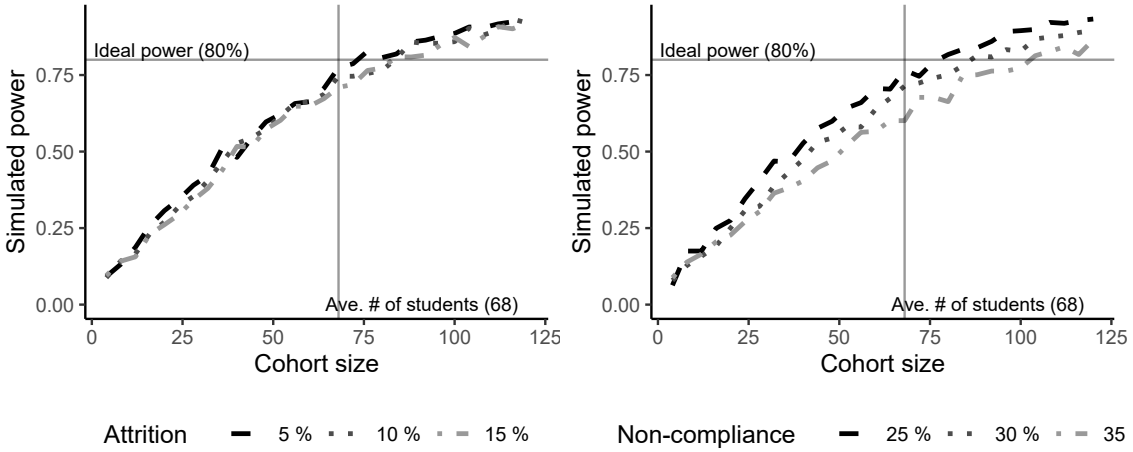
[13] In Section 3.2, we detail the procedure. In short, combining several measures into one index measure reduces the number of hypotheses to be tested. Rather than testing one hypothesis per variable, general conclusions are drawn by testing a hypothesis regarding the index.

[14] The lower sample sizes are not realistic, though they help to visualize the trend in power with respect to cohort size.

Figure 3.2: Statistical power simulations



(a) Effect size in base (3 waves) and worst-case scenario (2 waves)



(b) Worst-case scenario and attrition **(c)** Worst-case scenario and non-compliance

Notes: The simulations in Panel (a) have the following specifications: attrition rate is five percent, non-compliance is 25 percent, within-cohort correlation is 10 percent and the treatment probability is 59 percent. Statistical power to detect an effect of 15 percent, 20 percent or 25 percent for different average cohort sizes is presented. Cohort size is the sum of treatment and control group individuals. The right hand panel reports the worst-case scenario (two waves) while the left hand panel illustrates calculations for the base-case scenario (three waves). The worst-case simulations vary the attrition rate in Panel (b) and the non-compliance rate in Panel (c) for an effect size of 20 percent.

for the first wave in fall 2019; and a correlation within training cohorts of 10 percent, corresponding to a generously upward rounded measure from pilot data. The right half of panel a indicates that the design is sufficiently powerful (76 percent) to detect an effect of 20 percent (or 0.2 of a standard deviation) even in the worst-case scenario which

seems to be a typically observed change (McKenzie and Woodruff, 2014).[15] In the base-case scenario in the left half of panel a, our design would be well-powered to detect an effect size of 20 percent (89 percent power). If the effect size is actually only 15 percent of our standardized variable, the statistical power of our design reduces to 66 percent.

In panel b and c of Figure 3.2, we take the worst-case scenario and calculate the power to detect a 20 percent effect considering even more severe scenarios of attrition and non-compliance, holding the other parameters constant.[16] Panel b reports that attrition rates of 10 percent and 15 percent would only have a marginal effect on the power of the design. Panel c shows that non-compliance rates of 30 percent and 35 percent decrease statistical power to detect an effect of 20 percent to 71 percent and 60 percent, respectively. Overall, our design is well-powered for the base-case scenario with three waves. The conservative, worst-case scenario still yields better power than previous studies despite being below the generally accepted appropriate target of 80 percent power (McKenzie and Woodruff, 2014).

3.3 Data

Data collection and processing

Measuring treatment effects at two levels and describing selection into entrepreneurship requires a multitude of surveys. Figure 3.1 details our data collection efforts, and to which subpopulation surveys are administered. We make all survey instruments available through attachments to the pre-registration in the AEA registry #4502.[17]

Selection into entrepreneurship

The highest level of self-section occurs when individuals select into being interested in entrepreneurship training and attend an information session (see top of the pyramid in Figure 3.1). From this subpopulation, we collect the following data during the in-

[15] Our study is not only well-powered to detect typical effect sizes, it also improves on existing studies. McKenzie and Woodruff (2014) notes that in most studies the power to detect an increase of 25 or even 50 percent in profits or revenues is well below generally accepted levels of power of above 80 percent.

[16] In results not reported, we can also demonstrate that a correlation of 0.15 within training cohorts has only a negligible effect on the minimum detectable effect.

[17] To ascertain data integrity and safety, and to ensure survey respondents' privacy, we collect, manage and store data in the following way: First, the interview data is collected by experienced local enumerators. Prior to each data collection effort, PIs personally conduct extensive multi-day workshops with the enumerators. Data is collected using Kobo toolbox, and its Android-based mobile device app. Data is stored on secure drives provided by the University of Munich digital infrastructure. When data is collected using pen and paper, data is digitized also using Kobo toolbox in a timely manner and physical records are safely kept at the University of Munich to ensure privacy thereafter.

formation session: pen and paper based *short surveys* eliciting contact details, field of study, measures of cognitive ability using four Raven matrices, student's assessment of how many of these they believed they completed correctly and their assessment of their own entrepreneurial potential on a scale from one through ten.[18] To reach the non-interested subpopulation we track those classes where the training academy was advertised using short pitches. We classify all students of such a class as having been exposed to marketing. We return to the same classrooms a few weeks later and distribute *student population screening surveys*. These surveys mimic short surveys conducted during information sessions and also elicit students' awareness of entrepreneurship training programs. This allows us to identify students who were aware of the academy based on whether they have heard about our training program or about any entrepreneurship training program at their university.[19] The pool of students who are aware of a training program but did not apply constitutes the sampling frame for the *student population survey*. We then randomly sample 80 students per university, and survey them at two points in time. First, we conduct a phone survey mimicking the *baseline survey* conducted with academy applicants, which allows us to describe predictors of selection into entrepreneurship (*Student Population Survey I*). Second, we repeat this in *Student Population Survey II* to analyze how the subpopulation of non-interested students evolved over time relative to those who expressed in training but were not admitted—the control group. There is no experimental variation at either stage of this comparison.

Selection experiment

Attending information sessions is a necessary requirement for students to be able to apply to the training program since the exogenous variation of the marketing messages in the *selection experiment* is implemented in the information sessions. At the end of an information session, interested students can pick up a paper-based application form. Thus, application form data is only available for the subset of those interested in the training who actually submit a (complete) application form. Application forms contain contact details, demographic information, questions about motivations for and experience with entrepreneurship. We also include questions on students' expected future wage income, as well as expected earnings from entrepreneurship. With the experimental variation of the marketing messages we identify how selection into applying for entrepreneurship training varies with the stressed motives.

[18] A short and standardized illustration on how Raven matrices work in general and how students ought to indicate their answers on the short surveys was provided.

[19] Most universities do not offer alternative entrepreneurship training programs. Thus, it is reasonable to assume that students who were aware of a general academy were aware of *our* academy despite being unable to exactly recall the name of the program.

Entrepreneurship training experiment

To causally identify the effect of being offered entrepreneurship training, admission to the training program is offered on a random basis among those who apply. We gather pre-treatment data by conducting a *baseline survey* prior to individuals being informed about their admissions decisions. After the entrepreneurship training academy, we conduct an *implementation check survey* (around one to two months after the academy ends) and a *midline survey* (around six month later) with the treatment and control groups. Finally, we carry out two *endline surveys*: Endline Survey I will be conducted 12 months after each cohort is finished with their training; Endline Survey II surveys the entire sample around two years after the last round of academies. This survey will be conducted simultaneously for all cohorts and allows us to look at how medium to long-term effects evolve.

While the baseline, implementation check and midline survey are conducted over the phone, the endline surveys will be conducted in person. As detailed below, the surveys elicit information on socio-economic characteristics and main outcome variables, such as prior and ongoing wage and self-employment, preferences measures (risk and time preferences, degree of loss aversion), and non-cognitive traits (Big-5, grit, aspirations and personal initiative). Financial compensation for participation in the endline surveys helps to minimize attrition.

Key outcomes

We use the collected data to construct outcome measures for our five families of hypotheses, as laid out in Section 3.2. To test hypotheses we follow the approach by Kling et al. (2007) and aggregate variables into indices to test each main hypothesis (see Table 3.1) when possible. This reduces the number of tests conducted within each family. For instance, rather than testing for effects across ten business practices, we define an index using adherence to those ten practices and only conduct one hypothesis test. This hypothesis test in turn is part of a family of hypothesis tests. While we focus on indices of outcome measures here to address multiple hypothesis testing, we will also look at individual outcome variables during the analysis. We will clearly mark which results are accounting for multiple hypothesis testing and which are not.

Testing primary Hypothesis Families 1.1 and 2.1 will allow us to draw general conclusions about the entrepreneurship training experiment. Testing hypotheses within Hypothesis Family 1.2 is informative about the mechanisms through which the training program works. Hypotheses 2.2.1 and 2.2.2 set out to analyze dimensions which

correlate with entrepreneurial aspirations and success by comparing applicants to the non-interested subpopulation.[20]

To create a summary index from several continuous variables we calculate the unweighted average of those variables' z-scores. Z-scores are constructed using the control group mean and dividing by the control group standard deviation. Thus, each component of the index has a mean of zero and a standard deviation of one for the control group. To create an index of a set of binary variables we calculate their mean; that is, the fraction of "successes" across all component variables. If required, variables that are used to construct an index are reversed so that meaning is consistent.[21] In Appendix C1 we describe which variables are used to construct the indices in Table 3.1. The pre-analysis plan details the construction of the specific indices.

Hypothesis Family 1.1 consists of four indices: i) business creation (extensive margin), ii) business success (revenue, profits), iii) labor (employees) and capital (assets, inventory) input, and iv), an index of economic self-sufficiency which aggregates earnings from self-employment, wage employment and other sources.

Hypotheses Family 1.2 consists of six primary indices: i) business practices (we draw on an abbreviated version of the 22-item questionnaire used in McKenzie and Woodruff (2016), and retain ten elements of the original questionnaire (see Appendix C1), ii) financial professionalization (contains among others, knowledge and usage of financing instruments, indicators of business registration and licensing), iii) marketing practices, iv) capacity to innovate, v) business networks, and vi) development of an "entrepreneurial mindset" (a composite index constructed from measures of personal initiative, aspirations and entrepreneurial future and self-efficacy (Frese et al., 2007; Bernard and Taffesse, 2014; Campos et al., 2017; Streicher et al., 2019)). For the last two hypothesis families, non-cognitive traits, such as the Big-5 personality traits or grit (Rammstedt and John, 2007; Duckworth and Quinn, 2009), and time and risk preferences, as well as one's degree of loss aversion (Fehr and Goette, 2007; Falk et al., 2018), we create indices where there is a natural grouping (e.g., risk and subjective risk preferences), and investigate sub-indices in other cases (e.g., Big-5 indices).

Hypotheses Family 2.1 is the essence of the selection study and consists of four hypotheses: i) the relative effectiveness of the two randomly chosen marketing messages in terms of attracting applications, ii) whether applicants differ in their cognitive ability (proxied by performance on Raven matrices), iii) whether applicants exhibit differences

[20] We can compare the non-interested subpopulation to the full set of applicants using baseline data (pre-intervention). Using endline data, we compare the non-interested subpopulation to the control group (post-intervention).

[21] For example, all variables used to create the "Innovation" index are arranged so that a larger number indicates more innovative.

in over-confidence, and iv) whether applicants self-assess their entrepreneurial potential differently. We construct a measure of over-confidence by comparing individuals' observed and subjective (self-reported) performance on the Raven matrices (Moore and Healy, 2008; Åstebro et al., 2014).

There are two families of hypotheses which we use to study correlates of entrepreneurial aspirations and success in the wider population, Families 2.1.1 and 2.1.2. They mirror the hypotheses from Families 1.1 and 1.2 and therefore mimic the baseline and endline. These two families of hypotheses describe patterns through which students select into being interested in entrepreneurship training. We non-experimentally study baseline and endline differences between students who were interested in entrepreneurship training, and students who were not. First, the baseline comparison sheds light on how the subpopulation that applied to the training program differs from the general student population at large. Second, by comparing those that did not express interest (*non-interested subpopulation*) to interested students who were not offered admission to the training (control group) at endline, we can observe how those groups evolved over time.

Variation from intended sample size

The final sample size depends on the number of applicants and their response rate. To ensure that potentially interested students know about the academy and come to *information sessions*, we closely monitor the marketing campaign. To reduce attrition (i.e., non response) of the applicants over time, we conduct multiple rounds of follow-up surveys to establish frequent contact and trust.

Changing phone numbers represent the highest threat to maintaining contact with the surveyees. Therefore, in addition to students' own phone number(s), we inquire into contact details from a next-of-kin, their classroom coordinator and ask for an email address. In subsequent surveys, respondents are asked to verify or update this information. We achieved a response rate of 97.1 percent in the implementation check of the first wave.

We may use social media groups of the academies as an additional source of information in the future.[22] If those groups retain additional information, attrition could be treatment-specific. We try to account for this by documenting whether the data used to contact surveyees would have been available for both treatment and control. Further,

[22]Trainers typically create WhatsApp groups to stay in touch with their class members, share materials, and give updates about scheduling and locations.

when testing whether attrition is treatment-specific we will be conservative and test for attrition using 10 percent as threshold for statistical significance.

Should treatment status predict attrition, we will additionally provide treatment effect bounds using two approaches recently proposed in the literature. First, the procedure proposed by Lee (2009) quantifies the distribution of those who were induced to "staying in the sample" by treatment and estimates the best and worst-case scenarios. Second, we construct treatment effect bounds using the method suggested by Behaghel et al. (2015). This approach uses the number of attempts (e.g., phone calls) made to reach a person as instrument in a Heckman-type selection model.

Randomization balance

At this point, implementation of the intervention of the first two waves is completed. We have data available for all participants of the information session where we implemented the selection experiment using randomly chosen marketing messages across both waves. Additionally, we have collected baseline data from applicants and randomized admission offers across both waves (see Figure 3.1).

In Table 3.3, we conduct balance checks using the baseline data and compare those individuals who were offered admission (treatment) to those who were not (control) in the *entrepreneurship training experiment*. Columns 1 and 3 report the *unconditional* means for the treatment and control group. In column 5, we regress the respective variable on a treatment indicator, controlling for training cohort fixed-effects, and report the estimated treatment effect (*regression-adjusted difference*). Using heteroskedasticity robust standard errors, we then conduct a two-sided t-test of whether the treatment effect is equal to zero, and report the p-value in column 6. Overall, Table 3.3 suggests that randomization was successful; from 49 tests we conduct, only one difference is statistically significant at the five percent level. Specifically, treatment subjects report higher time preference scores which we attribute to random sampling variation.

For the *selection experiment*, we only have the short-surveys of participants at the information sessions as baseline data. The elicited characteristics (gender, field and year of study) were balanced across both randomly assigned marketing themes.

Table 3.3: Balance in entrepreneurship training sample

	Treatment		Control		Reg. Adj.		
	Mean (1)	St. dev. (2)	Mean (3)	St. dev. (4)	Diff. (5)	p-value (6)	N (7)
General							
Profit marketing theme (d)	0.44	0.50	0.41	0.41	0.01	0.83	1215
Male student (d)	0.52	0.50	0.58	0.58	−0.02	0.48	1215
Employment							
Working for a wage during the semester (d)	0.10	0.30	0.07	0.07	0.02	0.21	1214
Employer is company (d)	0.04	0.20	0.03	0.03	0.01	0.33	1214
Compensation per month in UGX (ths)	39.37	184.51	35.40	35.40	6.35	0.65	1214
Hours per week working	3.71	13.88	2.42	2.42	1.01	0.19	1214
Business							
Ever owned a business (d)	0.28	0.45	0.25	0.25	0.01	0.73	1214
Currently owning a business (d)	0.22	0.41	0.19	0.19	0.01	0.62	1214
Founder/Co-founder of business (d)	0.26	0.44	0.24	0.24	0.00	0.98	1214
Number of partners in business [*]	0.38	1.82	0.29	0.29	0.10	0.39	547
Business officially registered (d) [*]	0.02	0.14	0.02	0.02	0.00	0.71	546
Business has local trading license (d) [*]	0.03	0.16	0.05	0.05	−0.02	0.13	536
Length of existence of business	0.57	1.74	0.53	0.53	0.04	0.74	1214
Length of work at business in months	0.56	1.72	0.48	0.48	0.07	0.53	1214
Number of full-time employees	0.71	11.25	0.40	0.40	0.48	0.43	1214
Number of part-time employees	0.20	1.20	0.20	0.20	−0.01	0.93	1214
Hours per week working at business	7.71	18.58	6.12	6.12	0.19	0.85	1209
Profit per month at business in UGX (ths)	174.32	689.43	147.19	147.19	31.29	0.43	1214
Number of additional businesses owned [*]	0.06	0.27	0.04	0.04	0.01	0.52	547
Networks							
Personal contacts for business advice	0.70	0.46	0.73	0.73	−0.01	0.71	1210
Number of contacts in family and friends	2.74	3.30	2.86	2.86	−0.02	0.91	1206
Number of contacts outside family and friends	0.80	1.83	1.03	1.03	−0.23	0.12	1210
Contacts can help discussing business ideas (d)	0.67	0.47	0.72	0.72	−0.02	0.51	1202
Contacts helped discussing business ideas in the past (d)	0.40	0.49	0.43	0.43	−0.02	0.42	1202
Contacts can help collecting payments (d)	0.38	0.49	0.41	0.41	−0.02	0.48	1142
Contacts helped collecting payments in the past (d)	0.12	0.32	0.13	0.13	−0.01	0.55	1142
Contacts can help with sharing tools, inputs, employees (d)	0.37	0.48	0.39	0.39	−0.03	0.37	1126
Contacts helped with sharing tools, inputs, employees in the past (d)	0.12	0.33	0.13	0.13	0.00	0.82	1126
Contacts can help with purchasing inputs, stocks (d)	0.36	0.48	0.38	0.38	−0.02	0.51	1129
Contacts helped with purchasing inputs, stocks in the past (d)	0.13	0.33	0.12	0.12	0.00	0.84	1129
Funding							
Ever took loan to fund business idea (d)	0.08	0.28	0.08	0.08	0.00	0.97	1213
Number of known funding initiatives (out of 7)	1.38	1.16	1.41	1.41	0.01	0.88	1172
Non-Cognitive							
Big-5: extraversion	0.88	1.44	0.84	0.84	0.13	0.13	1210
Big-5: agreeableness	1.56	1.27	1.44	1.44	0.03	0.75	1212
Big-5: conscientiousness	1.99	1.22	1.90	1.90	0.00	0.95	1211
Big-5: neuroticism	−1.15	1.35	−1.13	−1.13	−0.07	0.40	1212
Big-5: openness	7.55	1.27	7.50	7.50	0.04	0.56	1213
Grit score (1-5)	3.57	0.44	3.58	3.58	−0.03	0.32	1203
Personal initiative score (1-5)	4.02	0.41	4.01	4.01	0.00	0.99	1208
Stress score (0-16)	6.20	2.25	6.02	6.02	0.11	0.42	1200
Preferences							
Risk preference: scale (1-5)	4.07	0.79	4.06	4.06	−0.03	0.54	1212
Risk preference: final number (1-32)	15.53	11.79	16.09	16.09	0.46	0.52	1213
Loss aversion: Final number (0-6)	4.68	2.02	4.57	4.57	0.16	0.20	1213
Time preference: scale (1-5)	4.03	0.90	3.96	3.96	0.06	0.28	1212
Time preference: final number (1-32)	11.58	12.36	9.79	9.79	1.42	0.05	1213
Entrepreneurial Self-Assessment							
Confidence in ability to start own company (1-5)	4.24	0.66	4.24	4.24	−0.03	0.46	1213
Confidence in ability to pursue self-employed career (1-5)	4.30	0.59	4.22	4.22	0.05	0.16	1213
Confidence in ability to manage challenges of an entrepreneur (1-5)	4.17	0.61	4.17	4.17	−0.03	0.39	1213
Confidence in ability to work in own business one year from now (1-5)	3.90	0.90	3.86	3.86	0.00	0.96	1200

Notes. Columns 1 and 3 report the unconditional mean, columns 2 and 4 the standard deviation for the treatment, who was randomly offered admission to the training program, and control group, respectively. Column 5 reports the regression adjusted mean $\hat{\beta}_1$ estimated using $y_{i,u} = \beta_0 + \beta_1 treat_{i,u} + \alpha_u + \varepsilon_{i,u}$ where α_u is training-cohort fixed effect. Column 6 displays the p-value from a two-sided t-test of $\mathrm{H}_0 : \beta_1 = 0$ using heteroskedasticity-robust standard errors. The last column shows the number of non-missing observations. (d) denotes an indicator variable. Variables marked with a [*] are those that were only measured in the second wave.

3.4 Analysis

OLS will be used if the outcome measure is continuous. We will report results from both logit and OLS regressions for binary outcomes, with the logit specification being our preferred. Inference about treatment effects will be based on two-sided t-tests obtained from using (cluster-)robust standard errors. We precisely state how standard errors are calculated when discussing the empirical specifications for estimating treatment effects. We separately discuss the empirical specifications for the entrepreneurship training study and the selection study. The p-values that govern our conclusions will take into account multiple hypotheses testing by being adjusted to control for the family-wise error rate (FWER). We detail the procedure in Section 3.4.

Entrepreneurship training experiment

In the *entrepreneurship training experiment*, we identify the Intention-to-Treat (ITT) effect of being offered admission to the entrepreneurship training. We separately estimate the coefficient of interest $\beta_{1,r}$ for short-term ($r = 2$, Endline I) and long-term effects ($r = 3$, Endline II) according to Equation (3.1):

$$y_{i,u,r} = \beta_{0,r} + \beta_{1,r}\texttt{treat}_{i,u} + \alpha_u + \texttt{strata}_{i,u} + \varepsilon_{i,u,r} \tag{3.1}$$

where $y_{i,u,r}$ is outcome (measured by an index) for individual i, training cohort $u \in \{1, ..., K\}$, and survey round r. The indicator variable $\texttt{treat}_{i,u}$ is equal to one if individual (applicant) i in training cohort u was randomly offered admission, and zero otherwise. Since randomization of admission offers was stratified by field of study and year of study, we include an indicator variable for every combination of the two variables.[23] Since the probability of being assigned to treatment differs across training cohorts, and is a function of the number of applicants, we include a training cohort fixed effect α_u.

Equation (3.1) is our preferred specification, and results from it will be reported first in the analysis. Put differently, estimates of β_1 from Equation (3.1) will be used to address the questions and hypotheses posed earlier. The following specifications are intended to provide more precise estimates in order to help us better gauge the magnitude of the estimated effects.

To improve the precision of $\widehat{\beta_1}$ we run a second set of specifications which includes a set of pre-treatment predictors. We follow the recommendation in Duflo et al. (2020) and use a variable selection approach. The double post-lasso estimation proposed by

[23] This results in five indicators included in the regressions, with one reference category omitted. These randomization cells refer to every combination of field of study (business and non-business) and year of study (first, second, and third).

Belloni et al. (2014) selects a low-dimensional set of predictors which are then included in the estimation. The method uses two separate Lasso regressions; one model to predict treatment assignment, another model to predict the outcome, and each model returns a set of variables to be included. Denote the union of this (as of now unknown) set of covariates by $X_{i,u,r=0}$. We further include the baseline value of the dependent variable $y_{i,u,r=0}$ whenever available.

$$y_{i,u,r} = \beta_0 + \beta_{1,r}\texttt{treat}_{i,u} + \beta_2 y_{i,u,r=0} + X'_{i,u,r=0}\gamma + \texttt{strata}_{i,u} + \alpha_u + \varepsilon_{i,u,r} \tag{3.2}$$

McKenzie (2012) discusses the benefits of a design that uses several post-treatment surveys to obtain more precise treatment effect estimates. Variables central to the analysis, such as profits and revenues, are likely to exhibit little auto-correlation. In this setting, statistical power in ANCOVA specifications is increased by pooling post-treatment observations. Section 3.3 describes that we conduct one midline follow up in addition to two endline surveys, resulting in three ($r \in \{1, 2, 3\}$) post-treatment surveys. Pooling those rounds, we estimate

$$y_{i,u,r} = \delta_r + \beta_1\texttt{treat}_{i,u} + \beta_2 y_{i,u,r=0} + \alpha_u + \varepsilon_{i,u,r} \tag{3.3}$$

where δ_r is a survey round fixed effect, and $r = 0$ indexes the baseline.

Effect heterogeneity

We are interested in analyzing heterogeneity in the ITT-effects along four independent, preregistered dimensions. First, we explore whether effects differ by an individual's field of study. Students in a business-related degree may have a higher ex ante likelihood of starting (successful) businesses due to higher entrepreneurial intentions or a different skill set (e.g., Solesvik, 2013; Bae et al., 2014). Second, we test whether effects differ by an individual's year in their degree. Students closer to graduation are more likely to move into (self-)employment in the near future. Thus, we test whether effects differ between students in their final (third year) and the remaining students. Third, we assess whether effects are different for students who report having sufficient financial means at baseline. Capital constraints have frequently been cited as the major obstacle to business growth in developing countries, and individuals who already possess the required funds may stand to benefit in a more immediate way (McKenzie and Woodruff, 2014). Fourth, we analyze differential effects by gender (Shinnar et al., 2014). Additional exploratory heterogeneity analyses (e.g., along self-reported motives and randomly assigned marketing themes, economic preferences or personality traits) will be clearly indicated as such.

Inference

Inference about the estimates in Equations (3.1) and (3.2) will be based on conventional heteroskedastic-robust Eicker-Huber-White standard errors. In case of Equation (3.3) standard errors will be clustered at the individual level since we use up to three observations per individual. Randomization of admission offers occurs at the individual level, and thus these standard errors are appropriate.

Selection into entrepreneurship

We describe and analyze selection at two steps before being (randomly) offered admission to the training program. In the *selection experiment*, random assignment to marketing messages during information sessions provides us with orthogonal variation which we exploit to study selection into applying for the training program along two salient motivations. Specifically, we use the following specification to analyze the differential effect of exposure to a specific marketing messages on a student's propensity to apply (Hypothesis 1 of Hypothesis Family 2.1):

$$\texttt{applied}_{i,u} = \beta_0 + \beta_1 \texttt{treat\_profit}_{i,u} + \alpha_u + W'_{i,u}\delta + \varepsilon_{i,u}. \tag{3.4}$$

Indices are defined as above; `applied` is an indicator equal to one if an individual submits an application for the training program, and zero otherwise; `treat_profit` is an indicator equal to one if an individual participated in an information session randomly emphasizing financial independence, and equal to zero if theme was creative freedom. The vector $W_{i,u}$ is included to increase the precision of estimates and it contains an individual's gender as well as indicators for years in the current degree (defined as above).

Hypotheses 2 through 4 of Family 2.1.1 capture the idea that selection patterns may differ relative to the underlying motivation for entrepreneurship. Denote a dimension of hypothesized heterogeneity in selection (cognitive ability, over-confidence, entrepreneurial self-assessment, see Hypotheses 2 through 4 of Family 2.1.1 in Table 3.1) with Z_i; we then estimate the following specification to test for different selection patterns:

$$\begin{aligned}\texttt{applied}_{i,t} = \beta_0 + \beta_1 \texttt{treat\_profit}_{i,u} + \beta_2 Z_{i,t} + \gamma Z_{i,t} * \texttt{treat\_profit}_{i,u} \\ + \alpha_u + W_{i,u}\delta + \varepsilon_{i,t}.\end{aligned} \tag{3.5}$$

Conclusions about differential selection will be based on assessing whether the estimated coefficients of our heterogeneity analyses are statistically significantly different from zero ($\mathsf{H}_0 : \gamma = 0$).

Effect heterogeneity

We do not anticipate to have sufficient power to study whether effects are heterogenous by individuals' field of study. However, we do intend to conduct exploratory analyses to assess whether the marketing messages induce differences in the composition of business and non-business students. In this case, we will follow Casey et al. (2012) and label the regressions as unregistered and exploratory.

Inference

The *selection experiment* is a clustered design in which all students participating in a given information session are exposed either to the financial independence or the creative freedom marketing message. Thus, standard errors should be clustered at the session level (Abadie et al., 2020); the level at which treatment varies. However, due to administrative issues, for some individuals we are unable to observe the exact session an individual attended and cannot cluster at this appropriate level. We attempt to overcome this by conservatively clustering at the training cohort level which is the next highest level. In the worst-case scenario of two waves, there are only 18 training cohorts and standard cluster-robust inference may over-reject. We thus pursue the wild bootstrap adjustment proposed by Cameron et al., 2008 to calculate standard errors and conduct inference.

Non-experimentally describing selection

Finally, we document selection into entrepreneurship by comparing those who were informed about the training program but did not attend an information session (non-interested subpopulation), to those who applied to the training program using baseline data. In addition, we document trends in how the non-interested subpopulation evolves over time relative to the subpopulation that expressed interest in the training. We do so by comparing them to those who applied but were not admitted—the control group—using Endline I ($r = 2$) data. Both comparisons are based on estimating the following specification.

$$\mathtt{y}_{i,u,r} = \beta_0 + \beta_1 \mathtt{applied}_{i,u} + \alpha_u + \varepsilon_{i,u,r}. \tag{3.6}$$

Indices are defined as above and we examine them at the baseline ($r = 0$), and again at the Endline I ($r = 2$). $\mathtt{applied}_{i,u}$ is an indicator equal to one if an individual applied to the training, and zero otherwise. There is no experimental variation at this stage and therefore $\widehat{\beta_1}$ does not measure a causal effect, but is merely informative of a correlation. We calculate heteroskedasticity-robust Eicker-Huber-White standard errors. For completeness, we also show results for individual index components.

Data processing

First, to establish that our results, especially those involving monetary outcomes, are not driven by extreme observations, we will report results with and without winsorizing outcomes at the 99^{th} percentile. Should a variable lack a natural lower bound (i.e., revenues are bound at zero, while profits are unbounded), we also winsorize at 1^{st} percentile.

Second, distributions of variables such as revenue and profits are likely be skewed to the right. We apply the inverse hyperbolic sine transformation to this data which is defined as $f(x) = log(x+\sqrt{x^2+1})$ (Burbidge et al., 1988). Note that this transformation is also defined for $x = 0$ and retains the interpretation of the classic linear-logarithmic regression model for all values of x — except for very small values.

Third, in order to limit noise caused by variables with minimal variation, questions for which 95 percent of observations have the same value within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests. In the event that omission decisions result in the exclusion of all constituent variables for an indicator, the indicator will be not be calculated. We explicitly exclude variables in Hypothesis 2 of Family 1.2 for "financial professionalization": Indicators, such as equity investment or business registrations are likely to be rare events and are insightful despite having little variation.

Fourth, whenever a survey's skip logic was triggered by a "yes" or "no" answer, we code the subsequent questions in the logical fashion.[24] Note that we account for the fact that people answer "don't know" or "don't want to answer"; We only impute the logical value if an explicit "yes" or "no" answer triggered the skip logic.

Section 3.3 describes how we construct indices to reduce the number of hypotheses tests. Note that the index value is missing if there is one or more missing values in the component variables (e.g., if a person answers "don't know" to one of the questions). We address this problem by providing two estimates in addition to the estimate based on the actually observed number of non-missing cases. First, we impute missing values using the mean value for the entire population, and then generate the index. For robustness, we also provide benchmarks for imputing minimum and maximum values for the entire population. Second, we implement an Inverse Probability Weighting (IPW) estimator in which each non-missing index value is weighted by the inverse probability of having data observed (Seaman and White, 2013). We model the incidence of observing an index value using a logit model with complete baseline characteristics (sex, employment status,

[24] For instance, if somebody does not know any entrepreneurs, then the number of friends and family members who are entrepreneurs is zero — although the skip logic would have result in this being a missing value.

self-employment; see Table 3.3), and use the predicted probability.

Fifth, in order to compare monetary values across time, we adjust values using Consumer Price Index data published by the Uganda Bureau of Statistics.

Multiple hypotheses testing

We construct several indices within each family of outcomes as detailed in Section 3.3 and Appendix C1. We employ two approaches to control the FWER, that is, controlling the probability of a false positive within each family. First, we implement the approach used by Aker et al., 2016 who use a traditional Bonferroni-type adjustment but account for correlations across variables used to test hypotheses.[25] Their method nests the classic Bonferroni adjustment when outcomes are uncorrelated. Second, we also employ the method outlined by Barsbai et al. (2020) who develop a regression-adjusted version of List et al. (2019). This is a bootstrap-based stepwise procedure designed to control the FWER in settings with multiple hypotheses.

Thus, for each hypothesis across our five families we obtain two p-values which control the FWER, on top of standard p-values. The p-values that correct for multiple hypothesis testing are of interest for researchers with no priors on the specific hypotheses we test. Our preferred procedure is the one by Barsbai et al., 2020 and our main conclusions will be based on being able to reject null hypotheses using those p-values. We report p-values using the procedure by Aker et al., 2016 for comprehensiveness.

Test for reporting errors being treatment independent

In business training interventions whose overall effectiveness is — among others — judged through financial outcomes and adherence to "good" management practices, reporting errors may not be independent of treatment assignment. Individuals who have gone through the training program may be better at accurately judging profits and sales. Alternatively, they may intentionally overstate profits (to suggest the training was helpful) or positively report on business practices because they are more likely to know what the "correct" answer is (McKenzie and Woodruff, 2016).

To address this concern, we construct a measure of sales minus profits which should equal costs and thus be weakly larger than zero. Should it be lower than zero, it likely signals a reporting error as costs cannot be negative. We then test whether treatment assignment predicts the incidence and magnitude of observed reporting errors. In a second step, we calculate implied revenue per customer, and compare the implied prices

[25] In our cases, we employ the correlation between index measures within each family.

of the goods and services across treatment and control and cross check with market prices.[26]

Conditional on detecting statistically significant treatment differences in reporting errors, we will conduct detailed in-person audits with a randomly selected subset of 100 treatment and 100 control group subjects. The audits will take place shortly after the endline data collection in the spirit of McKenzie (2017). We focus on business experience and business performance. This allows us to establish bounds of reporting errors for each of the variables studied (difference between endline self-reports and the audit data, separately by treatment and control groups). We will present the bounded results as robustness checks.

[26]We are aware of the possibility that new businesses may create goods and services of higher quality which command above-market prices. Nonetheless, implied prices should be largely comparable to market prices, assuming they are free of reporting errors.

Appendix C1 Construction of outcome indices

Table 3.1 provides an overview of the hypotheses. In the following, we detail which variables are used to construct those indices. Note that we spell out winsorization and transformation in Section 3.4, the creation of indexes based on z-scores in Section 3.3.

1. Entrepreneurship training study

1.1. Economic outcomes (four hypotheses)

1 *Business creation*

- Business exists (yes/no)
- Average hours contributed by the hour per week

2 *Business success*

- Monthly profits
- Monthly sales

3 *Capital and labor input*

- Value of physical assets
- Value of inventory
- Capital investment over past 3 months
- Number of full-time employees
- Number of part-time employees
- Number of partners in business

4 *Economic self-sufficiency*

- Earnings from self-employment (monthly profits)
- Earnings from wage employment
- Earnings from other sources

1.2 Business and personal input (eight hypotheses)

1 *Business practices*

- Share of business practices employed

2 *Financial professionalization*

- Taken out a loan (yes/no)
- Size of loan
- Business registration
- Local trade licenses
- Knowledge about funding initiatives
- Actual funding from initiatives

- Received equity investment
- Banking account
- Emergency borrowing
- Business banking account
- Hours of consulting services

3 *Marketing*

- Number of marketing channels used

4 *Innovation*

- Introduction of a new product (yes/no)
- Number of new products
- Main new product is a new product line (yes/no)
- Product improvement (yes/no)
- Product new to neighborhood (yes/no)
- Origin of idea (own idea vs. inspired vs. purchased/others idea)
- Process improvement (yes/no)
- Introduced a new method for pricing (yes/no)
- Website with functioning URL (yes/no)

5 *Networks*

- Number of contacts in friends and family
- Number of contacts in ”other”
- Scope of potential advice
- Scope of advice used
- Number of business partners

6 *Entrepreneurial mindset*

- Personal initiative
- Aspirations
- Entrepreneurial self-efficacy (general and task-specific separately)
- Entrepreneurial future

7 *Owner’s non-cognitive traits*

- Big-5
- Grit

8 *Preferences*

- Risk preferences
- Subjective risk preferences
- Loss aversion
- Time preferences

- Subjective time preferences

2. Selection study

2.1 Selection into entrepreneurship among those with interest (four hypotheses)

1 *Submitted application*

2 *Cognitive ability*

- Number of correctly solved Raven's matrices

3 *Over-confidence*

- Over-estimation
- Over-placement

4 *Entrepreneurial self-assessment*

- Believes about becoming a successful entrepreneur,
- Subjective rank of entrepreneurial ability,

2.2.1 Economic outcomes (non-experimental) [*identical to 1.1*]

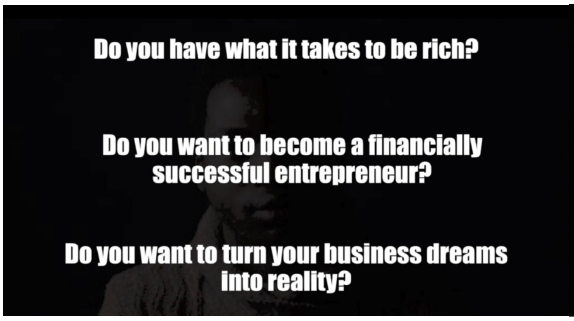
2.2.2 Business and personal input (non-experimental) [*identical to 1.2*]

Appendix C2 Marketing themes

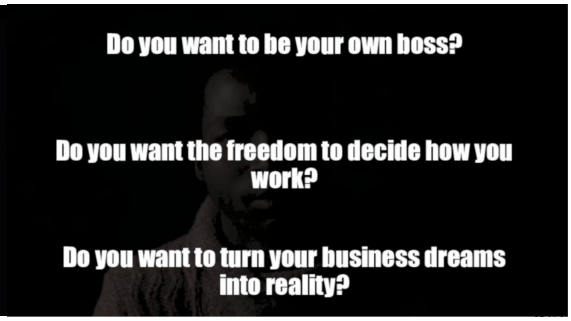
Section 3.2 describes how our design allows us to study selection into entrepreneurship. In order to apply to the entrepreneurship training program, students ought to attend information sessions where application forms can be obtained. We randomly vary the content of those information sessions by emphasizing either that entrepreneurship offers the possibility of achieving *financial independence*, or that entrepreneurship offers the *freedom to be creative*. Information sessions take approximately 15-20 minutes and the content is presented by a member of our partner organization. In each session, a presenter went through 12 presentation slides and two videos.

The videos constituted the main source of variation in the presentation. This guaranteed that students across sessions are exposed to the identical content. The first video differed in both visual and audio content. It was was 3 minutes 57 seconds in the profit condition, and 3 minutes 38 seconds in the creative freedom condition. The difference stems from the voice over being longer in the former. The second video only differed in audio content, and took 1 minute 53 seconds in both treatment conditions. Videos were embedded in the presentations to reduce technological complexity. The first video was presented on slide seven, while the second video was presented on the last slide. In between, slide nine presented different content.

In Figure C.1 we show examples of different content across the two treatments. In panels a and b, we show a still frame of the first video's first slide. Two of three statements differ, and the voice over emphasized the differences between the two treatments. Note that *not* entire presentation was kept in this black and white layout. In panels c and d we show the first frame of the second video. Again, the voice over emphasized the differences. Finally, we show the slide in which the presentations further differed in panels e and f.



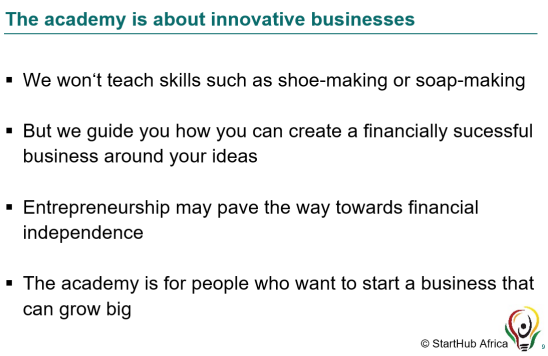
(a) First video, profit



(b) First video, creative freedom

(c) Second video, profit

(d) Second video, creative freedom



(e) Slide nine, profit

(f) Slide nine, creative freedom

Figure C.1: Example for treatment variation in information sessions

Bibliography

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). "Sampling-based versus design-based uncertainty in regression analysis". *Econometrica*, 88 (1), pp. 265–296.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). "The economics of density: Evidence from the Berlin Wall". *Econometrica*, 83 (6), pp. 2127–2189.

Aker, J., Boumnijel, R., McClelland, A., and Tierney, N. (2016). "Payment mechanisms and anti-poverty programs: Evidence from a mobile money cash transfer experiment in Niger". *Economic Development and Cultural Change*, 65 (1).

Aladangady, A. (2017). "Housing Wealth and Consumption: Evidence from Geographically-Linked Microdata". *American Economic Review*, 107 (11), pp. 3415–46.

Alan, S., Boneva, T., and Ertac, S. (2019). "Ever failed, try again, succeed better: Results from a randomized educational intervention on grit". *Quarterly Journal of Economics*, 134 (3), pp. 1121–1162.

Alaref, J., Brodmann, S., and Premand, P. (2020). "The medium-term impact of entrepreneurship education on labor market outcomes: Experimental evidence from university graduates in Tunisia". *Labour Economics*, 62, p. 101787.

Albouy, D. (2016). "What are Cities Worth? Land Rents, Local Productivity, and the Total Value of Amenities". *Review of Economics and Statistics*, 98 (3), pp. 477–487.

Andersen, S., Di Girolamo, A., Harrison, G. W., and Lau, M. I. (2014). "Risk and time preferences of entrepreneurs: evidence from a Danish field experiment". *Theory and Decision*, 77 (3), pp. 341–357.

Angrist, J. D. and Pischke, J.-S. (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics". *Journal of economic perspectives*, 24 (2), pp. 3–30.

Ashraf, N., Bandiera, O., Davenport, E., and Lee, S. S. (2020). "Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services". *American Economic Review*, 110 (5), pp. 1355–94.

Åstebro, T., Herz, H., Nanda, R., and Weber, R. A. (2014). "Seeking the roots of entrepreneurship: Insights from behavioral economics". *Journal of Economic Perspectives*, 28 (3), pp. 49–70.

Åstebro, T., Jeffrey, S. A., and Adomdza, G. K. (2007). "Inventor perseverance after being told to quit: The role of cognitive biases". *Journal of Behavioral Decision Making*, 20 (3), pp. 253–272.

Athey, S. (2018). "The Impact of Machine Learning On Economics". *The economics of artificial intelligence: An agenda*. University of Chicago Press, pp. 507–547.

Bae, T. J., Qian, S., Miao, C., and Fiet, J. O. (2014). "The relationship between entrepreneurship education and entrepreneurial intentions: A meta–analytic review". *Entrepreneurship Theory and Practice*, 38 (2), pp. 217–254.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). "CEO Behavior and Firm Performance". *Journal of Political Economy*, 128 (4), pp. 1325–1369.

Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015). "The miracle of microfinance? Evidence from a randomized evaluation". *American Economic Journal: Applied Economics*, 7 (1), pp. 22–53.

Barsbai, T., Licuanan, V., Steinmayr, A., Tiongson, E., and Yang, D. (2020). "Information and the formation of social networks". *NBER Working Paper #27346*.

Beaman, L., Magruder, J., and Robinson, J. (2014). "Minding small change among small firms in Kenya". *Journal of Development Economics*, 108, pp. 69–86.

Behaghel, L., Crépon, B., Gurgand, M., and Le Barbanchon, T. (2015). "Please call again: Correcting nonresponse bias in treatment effect models". *Review of Economics and Statistics*, 97 (5), pp. 1070–1080.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls". *Review of Economic Studies*, 81 (2), pp. 608–650.

Bernard, T. and Taffesse, A. S. (2014). "Aspirations: An approach to measurement with validation using Ethiopian data". *Journal of African Economies*, 23 (2), pp. 189–224.

Bjorvatn, K. and Tungodden, B. (2010). "Teaching business in Tanzania: Evaluating participation and performance". *Journal of the European Economic Association*, 8 (2-3), pp. 561–570.

Blattman, C., Fiala, N., and Martinez, S. (2014). "Generating skilled self-employment in developing countries: Experimental evidence from Uganda". *Quarterly Journal of Economics*, 129 (2), pp. 697–752.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3 (Jan), pp. 993–1022.

Bouchouicha, R. and Vieider, F. M. (2019). "Growth, entrepreneurship, and risk-tolerance: A risk-income paradox". *Journal of Economic Growth*, 24 (3), pp. 257–282.

Brooks, L. and Lutz, B. (2019). "Vestiges of Transit: Urban Persistence at a Microscale". *Review of Economics and Statistics*, 101 (3), pp. 385–399.

Brueckner, J. K., Fu, S., Gu, Y., and Zhang, J. (2017). "Measuring the Stringency of Land Use Regulation: The Case of China's Building Height Limits". *Review of Economics and Statistics*, 99 (4), pp. 663–677.

Bruhn, M., Karlan, D., and Schoar, A. (2018). "The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in Mexico". *Journal of Political Economy*, 126 (2), pp. 635–687.

Bruhn, M. and Zia, B. (2013). "Stimulating managerial capital in emerging markets: The impact of business training for young entrepreneurs". *Journal of Development Effectiveness*, 5 (2), pp. 232–266.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). "Alternative transformations to handle extreme values of the dependent variable". *Journal of the American Statistical Association*, 83 (401), pp. 123–127.

Burmeister, K. and Schade, C. (2007). "Are entrepreneurs' decisions more biased? An experimental investigation of the susceptibility to status quo bias". *Journal of Business Venturing*, 22 (3), pp. 340–362.

Calabrese, S., Epple, D., and Romano, R. (2007). "On the Political Economy of Zoning". *Journal of Public Economics*, 91 (1-2), pp. 25–49.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). "Bootstrap-based improvements for inference with clustered errors". *Review of Economics and Statistics*, 90 (3), pp. 414–427.

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H. C., McKenzie, D., and Mensmann, M. (2017). "Teaching personal initiative beats traditional training in boosting small business in West Africa". *Science*, 357 (6357), pp. 1287–1290.

Carlson, N. and Rink, A. (2019). "Starting fresh? Evidence from a field experiment with young entrepreneurs in Zimbabwe". *Working Paper*.

Casey, K., Glennerster, R., and Miguel, E. (2012). "Reshaping institutions: Evidence on aid impacts using a preanalysis plan". *Quarterly Journal of Economics*, 127 (4), pp. 1755–1812.

De Mel, S., McKenzie, D., and Woodruff, C. (2012). "Enterprise recovery following natural disasters". *Economic Journal*, 122 (559), pp. 64–91.

Dell, M. (2010). "The Persistent Effects of Peru's Mining Mita". *Econometrica*, 78 (6), pp. 1863–1903.

Dell, M., Lane, N., and Querubin, P. (2018). "The Historical State, Local Collective Action, And Economic Development In Vietnam". *Econometrica*, 86 (6), pp. 2083–2121.

Dettling, L. J. and Kearney, M. S. (2014). "House Prices and Birth Rates: The Impact of the Real Estate Market on the Decision To Have a Baby". *Journal of Public Economics*, 110, pp. 82–100.

Diamond, R. (2016). "The determinants and welfare implications of US workers' diverging location choices by skill: 1980-2000". *American Economic Review*, 106 (3), pp. 479–524.

Duckworth, A. L. and Quinn, P. D. (2009). "Development and validation of the Short Grit Scale (GRIT–S)". *Journal of Personality Assessment*, 91 (2), pp. 166–174.

Duflo, E., Banerjee, A., Finkelstein, A., Katz, L. F., Olken, B. A., and Sautmann, A. (2020). "In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics". *NBER Working Paper #26993*.

Evenson, B. and Wheaton, W. (2003). "Local Variation in Land Use Regulations". *Brookings-Wharton Papers on Urban Affairs*, 4, pp. 221–260.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). "Global evidence on economic preferences". *Quarterly Journal of Economics*, 133 (4), pp. 1645–1692.

Fehr, E. and Goette, L. (2007). "Do workers work more if wages are high? Evidence from a randomized field experiment". *American Economic Review*, 97 (1), pp. 298–317.

Fischel, W. A. (2001). *The homevoter hypothesis: How home values influence local government taxation, school finance, and land-use policies*. Harvard University Press Cambridge, MA.

Frese, M., Krauss, S. I., Keith, N., Escher, S., Grabarkiewicz, R., Luneng, S. T., Heers, C., Unger, J., and Friedrich, C. (2007). "Business owners' action planning and its relationship to business success in three African countries." *Journal of Applied Psychology*, 92 (6), p. 1481.

Frieden, B. J. (1979). "The New Regulation Comes To Suburbia". *The Public Interest*, 55, p. 15.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The Elements of Statistical Learning*. Vol. 1. 10. Springer Series in Statistics New York.

Furman, J. (2015). "Barriers to Shared Growth: The Case of Land Use Regulation and Economic Rents". Remarks. Urban Institute.

Ganguli, I., Huysentruyt, M., and Le Coq, C. (2018). "How do nascent social entrepreneurs respond to rewards? A field experiment on motivations in a grant competition". *SITE Working Paper 46*.

Ganong, P. and Shoag, D. (2017). "Why has regional income convergence in the US declined?" *Journal of Urban Economics*, 102, pp. 76–90.

Gelman, A. and Imbens, G. (2018). "Why High-order Polynomials Should Not Be Used In Regression Discontinuity Designs". *Journal of Business & Economic Statistics*, pp. 1–10.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). "Text as Data". *Journal of Economic Literature*, 57 (3), pp. 535–74.

Geshkov, M. V. and DeSalvo, J. S. (2012). "The Effect of Land-Use Controls on the Spatial Size of US Urbanized Areas". *Journal of Regional Science*, 52 (4), pp. 648–675.

Glaeser, E. L. and Gyourko, J. (2003). "The Impact of Building Restrictions on Housing Affordability". *Economic Policy Review*, 9 (2).

Glaeser, E. L., Gyourko, J., and Saks, R. (2005). "Why is Manhattan so Expensive? Regulation and the Rise in Housing Prices". *The Journal of Law and Economics*, 48 (2), pp. 331–369.

Glaeser, E. L. and Kahn, M. E. (2010). "The greenness of cities: carbon dioxide emissions and urban development". *Journal of Urban Economics*, 67 (3), pp. 404–418.

Glaeser, E. L., Gyourko, J., and Saiz, A. (2008). "Housing supply and housing bubbles". *Journal of Urban Economics*, 64 (2), pp. 198–217.

Glaeser, E. L. and Ward, B. A. (2009). "The Causes and Consequences of Land Use Regulation: Evidence from Greater Boston". *Journal of Urban Economics*, 65 (3), pp. 265–278.

Griffiths, T. L. and Steyvers, M. (2004). "Finding Scientific Topics". *Proceedings of the National Academy of Sciences*, 101 (suppl 1), pp. 5228–5235.

Guzman, J., Oh, J. J., and Sen, A. (2020). "What motivates innovative entrepreneurs? Evidence from a global field experiment". *Management Science*.

Gyourko, J., Mayer, C., and Sinai, T. (2013). "Superstar Cities". *American Economic Journal: Economic Policy*, 5 (4), pp. 167–99.

Gyourko, J. and Molloy, R. (2015). "Regulation and Housing Supply". *Handbook of regional and urban economics*. Vol. 5. Elsevier, pp. 1289–1337.

Gyourko, J., Saiz, A., and Summers, A. (2008). "A New Measure of the Local Regulatory Environment for Housing Markets: The Wharton Residential Land Use Regulatory Index". *Urban Studies*, 45 (3), pp. 693–729.

Hamilton, B. H. (2000). "Does entrepreneurship pay? An empirical analysis of the returns to self-employment". *Journal of Political Economy*, 108 (3), pp. 604–631.

Heblich, S., Redding, S. J., and Sturm, D. M. (2020). "The making of the modern metropolis: evidence from London". *The Quarterly Journal of Economics*, 135 (4), pp. 2059–2133.

Helsley, R. W. and Strange, W. C. (1995). "Strategic Growth Controls". *Regional Science and Urban Economics*, 25 (4), pp. 435–460.

Herz, H., Schunk, D., and Zehnder, C. (2014). "How do judgmental overconfidence and overoptimism shape innovative activity?" *Games and Economic Behavior*, 83, pp. 1–23.

Hilber, C. A. and Robert-Nicoud, F. (2013). "On the Origins of Land Use Regulations: Theory And Evidence From Us Metro Areas". *Journal of Urban Economics*, 75, pp. 29–43.

Hilber, C. A. and Turner, T. M. (2014). "The Mortgage Interest Deduction and its Impact on Homeownership Decisions". *Review of Economics and Statistics*, 96 (4), pp. 618–637.

Hilber, C. A. and Vermeulen, W. (2016). "The impact of supply constraints on house prices in England". *The Economic Journal*, 126 (591), pp. 358–405.

Hsieh, C.-T. and Moretti, E. (2019). "Housing constraints and spatial misallocation". *American Economic Journal: Macroeconomics*, 11 (2), pp. 1–39.

Hsieh, C.-T. and Olken, B. A. (2014). "The missing "missing middle"". *Journal of Economic Perspectives*, 28 (3), pp. 89–108.

Hurst, E. and Pugsley, B. W. (2011). "What do small businesses do?" *NBER Working Paper #17041*.

Hurst, E. G. and Pugsley, B. W. (2015). "Wealth, tastes, and entrepreneurial choice". *NBER Working Paper #21644*.

Ihlanfeldt, K. R. (2007). "The effect of land use regulation on housing and land prices". *Journal of Urban Economics*, 61 (3), pp. 420–435.

Irwin, E. G. and Bockstael, N. E. (2004). "Land Use Externalities, Open Space Preservation, and Urban Sprawl". *Regional Science and Urban Economics*, 34 (6), pp. 705–725.

Jackson, K. (2016). "Do land use regulations stifle residential development? Evidence from California cities". *Journal of Urban Economics*, 91, pp. 45–56.

Jackson, K. (2018). "Regulation, land constraints, and California's boom and bust". *Regional Science and Urban Economics*, 68, pp. 130–147.

Kahn, M. E., Vaughn, R., and Zasloff, J. (2010). "The Housing Market Effects of Discrete Land Use Regulations: Evidence From the California Coastal Boundary Zone". *Journal of Housing Economics*, 19 (4), pp. 269–279.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). "Experimental analysis of neighborhood effects". *Econometrica*, 75 (1), pp. 83–119.

Klinger, B. and Schündeln, M. (2011). "Can entrepreneurial activity be taught? Quasi-experimental evidence from Central America". *World Development*, 39 (9), pp. 1592–1610.

Koudstaal, M., Sloof, R., and Van Praag, M. (2016). "Risk, uncertainty, and entrepreneurship: Evidence from a lab-in-the-field experiment". *Management Science*, 62 (10), pp. 2897–2915.

LaLonde, R. J. (1986). "Evaluating the Econometric Evaluations of Training Programs With Experimental Data". *The American economic review*, pp. 604–620.

Leamer, E. E. (1983). "Let's Take the Con Out of Econometrics". *The American Economic Review*, 73 (1), pp. 31–43.

Lee, D. S. (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". *Review of Economic Studies*, 76 (3), pp. 1071–1102.

Levine, N. (1999). "The Effects of Local Growth Controls on Regional Housing Production and Population Redistribution in California". *Urban Studies*, 36 (12), pp. 2047–2068.

Levine, R. and Rubinstein, Y. (2017). "Smart and illicit: Who becomes an entrepreneur and do they earn more?" *Quarterly Journal of Economics*, 132 (2), pp. 963–1018.

Levine, R. and Rubinstein, Y. (2018). "Selection into entrepreneurship and self-employment". *NBER Working Paper #5350*.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). "Multiple hypothesis testing in experimental economics". *Experimental Economics*, 22 (4), pp. 773–793.

Mayer, C. J. and Somerville, C. T. (2000). "Land Use Regulation and New Construction". *Regional Science and Urban Economics*, 30 (6), pp. 639–662.

McKenzie, D. (2012). "Beyond baseline and follow-up: The case for more T in experiments". *Journal of Development Economics*, 99 (2), pp. 210–221.

McKenzie, D. (2017). "Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition". *American Economic Review*, 107 (8), pp. 2278–2307.

McKenzie, D. and Woodruff, C. (2014). "What are we learning from business training and entrepreneurship evaluations around the developing world?" *World Bank Research Observer*, 29 (1), pp. 48–82.

McKenzie, D. and Woodruff, C. (2016). "Business practices in small firms in developing countries". *Management Science*, 63 (9), pp. 2773–3145.

McMillen, D. P. and McDonald, J. F. (1991). "Urban Land Value Functions with Endogenous Zoning". *Journal of Urban Economics*, 29 (1), pp. 14–27.

Moore, D. A. and Healy, P. J. (2008). "The trouble with overconfidence." *Psychological Review*, 115 (2), p. 502.

Mullainathan, S. and Spiess, J. (2017). "Machine Learning: An Applied Econometric Approach". *Journal of Economic Perspectives*, 31 (2), pp. 87–106.

Ortalo-Magné, F. and Prat, A. (2014). "On the Political Economy of Urban Growth: Homeownership Versus Affordability". *American Economic Journal: Microeconomics*, 6 (1), pp. 154–81.

Pioneer Institute for Public Policy Research and Rappaport Institute for Greater Boston (2005). *Massachusetts Housing Regulation Database*. Prepared by Amy Dain and Jenny Schuetz. URL: http://www.masshousingregulations.com.

Porta, R. L. and Shleifer, A. (2008). "The unofficial economy and economic development". *NBER Working Paper #14520*.

Premand, P., Brodmann, S., Almeida, R., Grun, R., and Barouni, M. (2016). "Entrepreneurship education and entry into self-employment among university graduates". *World Development*, 77, pp. 311–327.

Quigley, J. M. and Raphael, S. (2005). "Regulation and the High Cost of Housing in California". *American Economic Review*, 95 (2), pp. 323–328.

Rammstedt, B. and John, O. P. (2007). "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German". *Journal of Research in Personality*, 41 (1), pp. 203–212.

Rigol, N., Hussam, R., and Roth, B. (2018). "Targeting high ability entrepreneurs using community information: Mechanism design in the field". *Working Paper*.

Saiz, A. (2007). "Immigration and Housing Rents in American Cities". *Journal of Urban Economics*, 61 (2), pp. 345–371.

Saiz, A. (2010). "The Geographic Determinants of Housing Supply". *The Quarterly Journal of Economics*, 125 (3), pp. 1253–1296.

Saks, R. E. (2008). "Job creation and housing construction: Constraints on metropolitan area employment growth". *Journal of Urban Economics*, 64 (1), pp. 178–195.

Schildberg-Hörisch, H. (2018). "Are risk preferences stable?" *Journal of Economic Perspectives*, 32 (2), pp. 135–54.

Schumpeter, J. (1911). "Theorie der wirtschaftlichen Entwicklung". *Joseph Alois Schumpeter*. Leipzig: Duncker & Humblot.

Seaman, S. R. and White, I. R. (2013). "Review of inverse probability weighting for dealing with missing data". *Statistical Methods in Medical Research*, 22 (3), pp. 278–295.

Severen, C. and Plantinga, A. J. (2018). "Land-use Regulations, Property Values, and Rents: Decomposing the Effects of the California Coastal Act". *Journal of Urban Economics*, 107, pp. 65–78.

Shane, S. (2009). "Why encouraging more people to become entrepreneurs is bad public policy". *Small Business Economics*, 33 (2), pp. 141–149.

Shertzer, A., Twinam, T., and Walsh, R. P. (2018). "Zoning and the economic geography of cities". *Journal of Urban Economics*, 105, pp. 20–39.

Shinnar, R. S., Hsu, D. K., and Powell, B. C. (2014). "Self-efficacy, entrepreneurial intentions, and gender: Assessing the impact of entrepreneurship education longitudinally". *International Journal of Management Education*, 12 (3), pp. 561–570.

Sims, K. R. and Schuetz, J. (2009). "Local Regulation and Land-use Change: The Effects of Wetlands Bylaws in Massachusetts". *Regional Science and Urban Economics*, 39 (4), pp. 409–421.

Solesvik, M. Z. (2013). "Entrepreneurial motivations and intentions: Investigating the role of education major". *Education+ Training*, 55 (3), pp. 253–271.

Streicher, M., Rosendahl Huber, L., Moberg, K., Jørgensen, C., and Redford, D. (2019). "Filling in the blanks? The impact of entrepreneurship education on European high school students". *Academy of Management Proceedings*.

Stroebel, J. and Vavra, J. (2019). "House Prices, Local Demand, And Retail Prices". *Journal of Political Economy*, 127 (3), pp. 1391–1436.

Tsivanidis, N. (2018). "The aggregate and distributional effects of urban transit infrastructure: Evidence from bogotá's transmilenio". Working Paper.

Turner, M. A., Haughwout, A., and Van Der Klaauw, W. (2014). "Land use regulation and welfare". *Econometrica*, 82 (4), pp. 1341–1403.

Ubfal, D., Arraiz, I., Beuermann, D. W., Frese, M., Maffioli, A., and Verch, D. (2019). "The impact of soft-skills training for entrepreneurs in Jamaica". *IZA Discussion Paper #12325*.

Wallace, N. E. (1988). "The Market Effects of Zoning Undeveloped Land: Does Zoning Follow the Market?" *Journal of Urban Economics*, 23 (3), pp. 307–326.

Zhou, J., McMillen, D. P., and McDonald, J. F. (2008). "Land values and the 1957 comprehensive amendment to the Chicago zoning ordinance". *Urban Studies*, 45 (8), pp. 1647–1661.