Statistical Methods for Characterizing Occurrences and Impacts of Climate

by

Miyabi Ishihara

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Samuel Pimentel, Co-chair
Professor Solomon Hsiang, Co-chair
Associate Professor Avi Feller

Fall 2023

Statistical Methods for Characterizing Occurrences and Impacts of Climate

Abstract

Statistical Methods for Characterizing Occurrences and Impacts of Climate

by

Miyabi Ishihara

Doctor of Philosophy in Statistics

University of California, Berkeley

Assistant Professor Samuel Pimentel, Co-chair

Professor Solomon Hsiang, Co-chair

This thesis explores statistical methods for characterizing the occurrences and impacts of climate. Data on climate and the environment more broadly show unique characteristics, posing methodological challenges.

Part I presents an overview of an emerging, accessible earth observation data source, satellite images, and explores its potential applications in statistical and causal inference. We introduce an approach for incorporating image data, represented by image features, into a regression framework as a potential proxy for confounding variables. The study examines the impact of image features on regression estimates, focusing on the characterization of bias with and without the inclusion of images, as well as the conditions under which the inclusion of image data reduces or amplifies bias.

Part II introduces an empirical and methodological problem related to estimating the economic values of the environment. This question is pertinent for understanding how various environmental qualities, such as flood risk and air pollution, are currently reflected in residential property values and how society manages climate and environmental risks. However, a methodological challenge in estimating capitalization lies in the difficulty of accounting for confounders, which invites the application of the method studied in Part I. The results indicate that while some risk-related factors are associated with lower housing values, others, such as PM 2.5 and flood risk score, are associated with higher housing prices. This provides a lens through which to discuss how risks could be reflected in property values.

Part III focuses on statistical methods for characterizing occurrences of extreme climate events. These data are spatio-temporal in nature, and effectively visualizing and summarizing such data is challenging, though crucial for monitoring and identifying hazardous events. A primary challenge stems from the context-dependent nature of extreme events, where definitions vary over time and space. While many adopted approaches involve pre-defining criteria for anomalous events, typically by setting an exceedance threshold, determining appropriate values for the exceedance threshold, time window, and spatial boundary is a non-trivial task. The study applies functional principal component analysis to characterize spatio-temporal trends of extreme precipitation. The method shows potential as a flexible way to identify both the temporal window and geographic location of anomalous events.

To my family and friends.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Sol Hsiang, Avi Feller, and Sam Pimentel for their patience, kind advice, and support in navigating me through various aspects during my years in graduate school. They have shared perspectives and offered guidance on working interdisciplinary research questions as well as more general life matters. I would also like to thank Jon Proctor for engaging in discussions with me on both large and small questions of research.

Peng Ding served as my mentor during the early years of my graduate studies. His course on causal inference, along with conversations with him, Sol, Avi, and Sam, helped formulate my unwieldy interest into a more concrete one. Fernando Perez served on my exam committee and provided thoughtful comments.

I also appreciate the interdisciplinary communities that have enriched my journey, such as the Environment and Society: Data Science for the 21st Century (DS421), the Global Policy Lab, and the course Algorithms and Inequality. In the DS421 program, I met Sol and worked with PhD students in the Global Policy Lab on a satellite data project. This experience encouraged the exploration of satellite data's usage in inference, as discussed in Chapters 2, 3, and 4. Engaging in discussions with fellow members of the lab has also encouraged me to think across disciplines and learn from different frameworks. In the Algorithms and Inequality course, Rediet Abebe and fellow graduate students shared the value of bottom-up perspectives on algorithmic and statistical tools, and the importance of contextual nuance in engaging with difficult problems.

The work included in Chapters 2, 3, and 4 was conducted in collaboration with Jon, Sol, and Avi. The research included in Chapter 5 was a collaborative work with Chris Paciorek, Mark Risser, and Michelle Yu. I am grateful to all of them for their insights, but any errors are my own.

I owe many thanks to Daphna Harel and the applied statisticians I was fortunate to meet at NYU. Their humility, respect for complex problems, and encouragement helped me to continue my studies at Berkeley. I am grateful for the conversations with Sara Stoudt and Michelle Yu, who have also shared their experiences in working on interdisciplinary questions. The staff members of the Statistics Department and the Global Policy Lab have helped me in navigating Berkeley as an institution.

Lastly, I would like to express my gratitude to my family and friends for their support and everything they have shared with me.

Chapter 1

Introduction

This thesis explores the emerging accessible data of earth observation captured by satellite images and explores how it affects statistical and causal inference. The study discusses an approach for incorporating image features into a regression framework and applies it to study an empirical question: estimating the economic values of the environment, a key aspect necessary for discussing how different environmental qualities are currently considered and how risks are managed in society. The final part of the thesis focuses on extreme climate events and applies a statistical method for characterizing spatio-temporal extremes, an aspect necessary for monitoring and identifying anomalous events.

1.1 Regression using satellite imagery

Since the early 2000s, satellite images have become an accessible source of data, contributing to the monitoring of land-based attributes on a large scale. With over 1,000 earth observation satellites in orbit [85], satellite imagery offers a viable means of obtaining observations in locations where ground-based measurements face limitations, whether due to cost, risk, or time constraints.

Therefore, the usage of satellite image data, particularly in policy domains, involves aiding in the understanding and measurement of specific issues with detailed spatial and temporal resolution on a large scale. Additionally, image data have been used to assist in formulating intervention plans, including time-sensitive resource allocation and program design, while also providing information to evaluate the estimated impact of such interventions.

In using satellite image data for analysis, a methodological challenge lies in effectively transforming unstructured image data into a structured format. Many image

featurization techniques focus on estimating specific variables, such as tree cover. With the development of a generalizable featurization approach proposed by Rolf et al [73], it has become possible to use a single encoding of images to estimate various types of variables, resulting in more widely accessible structured image data for research.

While predictive performance has been studied for various variables, including considerations of error structure and the nature of bias, questions remain regarding how the inclusion of image data affects statistical and causal inference. Assessing the impact of satellite data on such analyses is necessary to understand how and under what circumstances the use of image data helps with inference.

Chapter 2 studies the impact of incorporating image features into a regression framework. Specifically, we investigate the potential role of satellite images as proxies for confounding variables, given their capacity to capture rich information observable from space. The chapter discusses the conditions under which the inclusion of images either reduces or amplifies bias compared to the omission of confounding variables.

We observe that incorporating image features reduces the bias of coefficient estimates for the variable of interest when their inclusion makes that variable or the outcome less confounded —meaning it is less associated with the confounder. Even if imagery captures a large portion of the variations in the confounder, the potential for bias amplification exists if there is a shared variation in the outcome, variable of interest, and confounder that is not captured by the imagery. While the conditions require the knowledge of unmeasured confounders and a diagnostic check remains a perpetual objective, our study using empirical data suggests that image inclusion tends to reduce bias when the number of observed covariates is limited.

1.2 Economic value of environment

Chapters 3 and 4 are driven by specific empirical and methodological questions, and we use the method introduced in Chapter 2 to study these questions. Globally, the frequency and cost of weather and climate disasters are increasing due to a combination of factors, including the heightened frequency of extreme climate events, increased exposure, and vulnerability. While the reasons for increased exposure are multifaceted and shared among multiple parties [42], one domain in which climate risks and broader environmental qualities are capitalized is the housing market. From one perspective, the housing finance system has the potential to incorporate the risks associated with specific regions and play a role in discouraging risky developments, such as the rapid construction of properties in high climate-risk areas.

Given the increasing importance of risk management and the interconnected nature of these risks, it is essential to gain a comprehensive understanding of the current state of incorporating environmental and climate conditions into property values on a national scale and across multiple variables simultaneously. Previous studies on capitalization have primarily concentrated on specific regions or environmental qualities, and results vary widely. The substantial variability in results may, in part, be attributed to the methodological challenge in regression, which involves the difficulty of effectively accounting for all confounding variables—a crucial aspect for obtaining accurate estimates of capitalization.

Our study adds to the existing thread of research by estimating the extent to which diverse environmental qualities are incorporated or overlooked on a national scale. The environmental qualities under study include flood risk, storms, precipitation, surface water, air pollution, wildfire risk, temperature, sunlight, dew point, elevation, and tree cover. In estimating the capitalization of each, we use regression analysis that controls for satellite data, as introduced in Chapter 2.

Chapter 3 situates the empirical problem, synthesizing the existing research in the valuation of the environment. The chapter also summarizes examples of how these research results were used outside the academic community, particularly in providing one perspective for designing or assessing environmental policy. The chapter further presents a catalog of the nature of the housing, environmental, and satellite data used to study the capitalization problem.

In Chapter 4, the regression using satellite images is implemented, and the implications of the estimates are discussed. While some risk-related factors, such as the fire potential index and wind speed as indicators of storms, show a general downward trend, indicating that greater risk is associated with lower pricing, we observe that certain risk-related factors, such as PM 2.5 and flood risk score, are associated with higher housing prices. While a more detailed study is necessary to understand how the trend varies at the state or county level, on a national scale, the current housing pricing raises questions about how risks should be reflected.

1.3 Characterizing spatio-temporal trends of extreme precipitation

Despite the contributions of satellite data and spatial smoothing techniques in estimating missing information and achieving spatially complete data, in situ measurements continue to be imperative. This significance becomes particularly evident when studying rare events that deviate from the norm, such as extreme climate

events. Firstly, assessing the accuracy and error characteristics of satellite data requires ground-truth measurements. The capacity of satellite data is limited to what can be observed from space. Additionally, interpolation techniques used to generate spatially complete data rely on input from ground-truth data, and they have a tendency to dampen extreme values, potentially leading to misleading results [27] [83] [71]. In situ measurements, collected directly from the specific locations of interest over time, remain an essential source for providing data at the native scale.

When studying spatio-temporal data of extremes, one of the initial challenges is in effectively visualizing and summarizing the information to help with the monitoring and identification anomalous events. A commonly adopted approach involves pre-defining the criteria for anomalous events, typically by setting an exceedance threshold. This involves calculating the frequency of events surpassing the threshold and aggregating the results within predefined spatial subdivisions. However, determining the appropriate values for the exceedance threshold, time window, and spatial boundary is a non-trivial task. The challenge arises from the context-dependent nature of extreme events, where definitions vary over time and space. What may be considered as extreme in one region may not be considered the same in another location.

Chapter 5 provides a concise overview of the exploratory analysis of visualizing and summarizing spatio-temporal extremes, specifically focusing on extreme precipitation in the contiguous United States. We use functional principal component analysis, a methodology in the literature of functional data analysis, commonly applied in health and biomedical contexts [84] for grouping multiple functional or longitudinal datasets without spatial information. This method is used to characterize extreme precipitation measurements collected over the last century, each obtained from multiple weather stations across the United States.

To accommodate the diverse definitions of extreme events across regions and time, the method is implemented to minimize the need for pre-specified criteria. Specifically, the analysis does not require predefined specifications for dividing the contiguous US or determining temporal windows. The resulting principal component (PC) function, coupled with maps of PC scores, offers a way to identify both the temporal window and geographic location of anomalous events.

Part I

Regression Using Satellite Data

Chapter 2

Use of Satellite Imagery in Regression

2.1 Introduction

Causal effects can be estimated using regression if the model is accurate and includes all confounding variables. Failure to account for confounding variables leads to biased estimated coefficients. In settings where there are spatially correlated unmeasured confounders, a common approach to mitigate bias is by incorporating spatial information. The underlying assumption is that units in close geographic proximity share similarities in terms of unmeasured confounding factors, and thus involves treating spatial information as a proxy for these confounders in the estimation process. Papadogeorgou has developed methods in this domain, including a propensity score matching procedure which takes into account units' spatial proximity, ensuring appropriate matching of units with similar covariates and spatial proximity [56]. In a regression framework, Druckenmiller and Hsiang proposed a spatial analogue of the first differencing approach commonly used in time series analysis. This method involves regressing the spatial first differences of the outcome variable on the treatment variable, thereby differencing out these covariates [20].

In many cases, we do not know whether unmeasured confounders are spatially correlated or not. Certain variables such as land use and socio-demographics can differ considerably in neighborhoods that are geographically close to each other. On the other hand, neighborhoods that are far away from each other may share similar confounding (Figure 2.1). We are interested in a proxy for unmeasured confounding that is defined for each location unit (and thus free from the assumption that nearby units are similar) and study its effect on coefficient estimates. For such a proxy,

we use satellite images, which capture a range of environmental and socio-economic characteristics that are visible from space.

Figure 2.1: **Example showing the relationship between physical features and geographic proximity of three neighborhoods captured by satellite images.** Image A shows a neighborhood with high tree coverage and a few houses, whereas images B and C show residential neighborhoods. Images A and B are 1.5 miles apart and images B and C are 3000 miles apart.

In this chapter, we provide an overview of earth observation data as represented by satellite images and its potential contribution in causal inference, along with machine learning methods for summarizing image data. The chapter presents a basic framework for incorporating image information within a regression context. We explore the effect of images on regression estimates, including the characterization of bias with and without the inclusion of images, as well as simulation studies that examine scenarios where controlling for images can help reduce or amplify bias. We comment on constraints associated with satellite images, underscoring the need of context-specific considerations and future prospects in this domain.

2.2 Earth observation for causal inference

Current usage of satellite imagery in policymaking

Satellite-based data has been used in policymaking across a range of functions. Satellite data plays a role by: a) aiding individuals to understand and measure specific problems at a detailed spatial and temporal resolution over a large scale; b) assisting in the formulation of intervention plans, including time sensitive resource allocation and program design; and c) providing information to evaluate the impact of interventions.

Satellite data enables individuals to understand and measure specific problems that pose challenges for traditional data collection methods. These difficulties can stem from the nature of the variable being measured or the location where data is required, which may involve risks or financial constraints. For example, since the early 2000s, satellite images have played a role in monitoring land-based attributes like forest cover [24] [7], which were previously assessed through ground surveys. By using satellite sensors and imaging systems that adhere to standardized techniques and calibration procedures, globally comprehensive measurements are obtained, ensuring comparability and consistency across different regions.

The fine spatial resolution of satellite imagery enables the acquisition of localized information at a large scale. For instance, agricultural data that was previously accessible only at larger scales such as state, district, or county levels can be obtained at the field scale by satellites [37]. This development allows for more targeted interventions to be implemented, as stakeholders gain a detailed understanding of specific areas and can tailor their strategies accordingly.

The temporal resolution and scale of satellite data facilitate the study of variable changes over time, spanning both small and large time scales. Satellites have been available since the 1970s, and depending on the sensor, they offer time resolutions as fine as daily measurements. This capability enables the comparison of ground conditions before and after specific events, such as evaluating the extent of damage caused by natural disasters or conflicts [88] [11], and assessing the impact of development programs [36] [68] and conservation policies [9]. The effects in these examples can manifest over various time spans, ranging from immediate impacts following the event to longer-term consequences.

The combination of granularity and scale in satellite data also serves in the estimation of missing data. Satellite data has been used to learn and fill in data gaps for regions or variables where ground measured information is lacking. For instance, machine learning methods have been used to generate poverty maps by learning the relationship between satellite data and ground-measured poverty levels, and estimating poverty levels in regions with incomplete data, including countries with low-resource or affected by conflicts [39] [32] [93] [79] [14]. Another application involves the delineation of crop field boundaries in regions where field boundary datasets were previously unavailable [90], which has been contributing to the creation comprehensive crop type mapping and yield mapping for smallholder farmers.

The examples provided highlight the capability of satellite images to swiftly estimate information in previously inaccessible areas, enabling real-time monitoring and facilitating prompt actions. They can bring long-standing problems and emerging problems more salient to public attention. In this context, satellite imagery serves as a tool for estimating missing values of a specific variable Y in particular locations

by leveraging data from areas where ground truth information is available. The estimated outcome variables $\hat{Y}$ have been used to inform the selection of appropriate interventions T and monitor the impact resulting from the implementation of the chosen intervention T.

Challenges of using satellite imagery in statistical analysis

While satellite imagery data presents many opportunities, the use of imagery data is subject to certain challenges, which can be categorized into multiple aspects. First, there are nuances inherent to imagery as a data product, including measurement errors and its unstructured nature. The unstructured nature necessitates methods to transform image data into a numerical summary, while the indirect measurement nature of the data requires treating the numerical summary as a proxy of the variable that is intended to capture. Second, there are relatively unexplored realms concerning how these qualities of image data impact the implications drawn from downstream statistical analysis and causal inference.

There is a limit regarding the extent to which satellite imagery can estimate certain variables. This limitation arises both from the machine learning method and the presence of measurement errors inherent in the satellite imagery itself. Therefore, although the inclusion of additional training data and imagery sources can improve predictive performance, there is a threshold to the extent of improvement.

Several studies have been conducted to explore the manifestation of these errors across various variables and geographic locations. For example, predictions of variables using imagery often demonstrate reduced variability compared to the true values, resulting in reduced predictive performance, particularly at the lower and upper extremes [73] [68] [63] [77]. The systematic tendency to over-predict low values and under-predict high values across all variables can in part be attributed to the choice of the objective function in the model. This selection favors predictions that gravitate towards the mean, leading to such patterns in the predictions.

Predictive accuracy can also vary depending on the geographic location. A study examining the accuracy of satellite-based poverty maps indicates the challenges associated with accurately predicting wealth levels within urban and rural areas [2]. For instance, satellite-based poverty maps may fail to identify urban populations living in poverty, which may be attributed to the limited capacity of satellite data in capturing more detailed wealth information beyond distinguishing whether an area is urban or not.

With regard to how the inclusion of satellite imagery affect implications drawn from statistical or causal analysis, studies are limited. Proctor et al focuses on the measurement error of image data and quantifies the extent to which variables

estimated via satellite imagery affect bias and uncertainty when used in regression analysis either as an independent or dependent variable [63]. Jerzak et al studies the potential of using imagery as proxy for confounding variables in observational causal inference [69] [40] [41].

In this study, we consider using satellite imagery as a proxy for confounding variable in a regression analysis, and quantify its effect on bias both from theoretical and simulation point of view.

2.3 Controlling for imagery in regression

We consider the problem of estimating causal effects using regressions in which we observe proxies for latent confounders.

Setting and notation

Let Y_i denote the continuously scaled outcome of the i^{th} of n location units, for example, the average housing price in locational region i. Let T_i be the treatment, which can be categorical or continuous, for example, the annual average air pollution in location i. Let $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ denote the vector of p observed covariates, for example, the average number of rooms or building age of houses in location i. Let $\mathbf{U}_i = (U_{i1}, ..., U_{ik})$ denote the vector of k latent confounding covariates that are related to the outcome and/or treatment variable. This may be urban form, which encompasses the configuration of buildings, transportation corridors, and other structural elements that shape the physical aspects of a city.

Consider the following "correct" but infeasible hedonic regression model:

$$Y_i = \beta_0 + T_i\beta_T + \mathbf{Z}_i\,\beta_Z + \mathbf{U}_i\,\beta_U + \epsilon_i \tag{2.1}$$

If the confounding variable is not observed, we can fit the model

$$Y_i = \beta_0^* + T_i\beta_T^* + \mathbf{Z}_i\,\beta_Z^* + \epsilon_i^*. \tag{2.2}$$

Omission of confounding variables $\mathbf{U}$ induces bias of coefficient estimates and may affect conclusions about the extent to which housing values reflect air pollution information. Unobserved variables are often difficult to define specifically or may not be available as ground measured data.

A common approach to alleviating the problem of unobserved heterogeneity has been to use spatial fixed effects [1].

This problem invites the use of remotely sensed data sources, in particular satellite images, to serve as proxy for unmeasured confounding variables. Satellite images capture various aspects of socioeconomic and environmental characteristics to the extent that they can be observed from space. Given their fine spatial scales, images can be used to detect small-scale heterogeneous variations.

Conceptually, we are interested in obtaining a regression estimate of interest while controlling for images – an infeasible task given that images are unstructured and cannot be regressed on. Hence we use a statistical summary of images, as represented by a vector of variables (hereafter, image features). Because the interest is in using images as proxy of unmeasured confounders, we rely on a highly descriptive set of features with the capacity to estimate ground conditions across diverse variables.

In this study, we use the image featurization technique called Multi-task Observation using Satellite Imagery and Kitchen Sinks (MOSAIKS), as proposed by Rolf et al [73]. This method involves a one-time unsupervised image featurization process using random convolutional features [65]. Image features measure the similarity between the image and smaller patches of imagery, which are selected randomly. hese features can be used for predicting ground conditions through linear regression, where the variable of interest is regressed against the features. The predictive performance of these features has been studied across various environmental and socioeconomic variables, both within and outside the U.S. [63] [73] [77]. The predictive performance of MOSAIK's image features is comparable with existing deep-learning methods, whilst offering faster computational speed [73]. Each satellite image in this study is represented using 4000 features. Further details regarding the featurization of images can be found in Section 3.4.

Convolutional random features, which we denote as $\mathbf{R}_i = (R_{i1}, ..., R_{if}) = r(I_i)$, is a vector of $f = 4000$ elements that is transformed from the daytime image I_i. Because $\mathbf{R}$ is descriptive, we expect that it captures variations of T, $\mathbf{Z}$, and $\mathbf{U}$ and thus can be represented as a function of these variables with some noise:

$$R_i = f(T_i, \mathbf{Z}_i, \mathbf{U}_i) + \varepsilon_i$$

When $\mathbf{R}$ is used as proxy for unmeasured confounders of $\mathbf{U}$, we can fit the model

$$Y = \beta_0' + T_i\beta_T' + Z_i\beta_z' + R_i\beta_R' + \varepsilon_i'. \tag{2.3}$$

2.4 Effects of controlling for imagery

Sources of Bias

One approach for understanding the effect of controlling for imagery is to study its effect on the bias in the coefficient estimate of the variable of interest, β'_T. To do this, it helps to consider sources of bias that is induced from controlling for image information. There are three sources: measurement error bias, omitted variable bias, and included variable bias.

Measurement error bias

Satellite images inherently involve measurement error as they serve as proxies for confounding variables rather than direct representations of those variables. The quality of images is influenced by the characteristics of the satellite sensor and its angles [47]. These factors influence the measurement of radiance, which is the amount of light reaching the satellite sensor. Atmospheric conditions, such as air pollution or rainfall, can also contribute to errors by introducing additional gases and aerosols into the atmosphere. Typically, including the images used for this analysis, image processing is applied to correct for variations caused by factors, such as atmospheric conditions, sensor characteristics, and sensor angles. Radiance data is also converted to surface reflectance data, a more consistent measure of the proportion of sunlight reflected by the Earth's surface.

The choice of how images are represented by features also plays a role in measurement errors. In this analysis, every image is characterized by 4000 random convolutional features, although alternative methods of feature creation or different number of features are possible. The accuracy of representing the confounder can vary depending on the method of image featurization. If an essential confounder is subject to measurement error, it can result in considerable bias in the estimate [81].

Omitted variable bias

Even with descriptive features extracted from high-quality images, omitted variable bias can persist. The limitations lie in the characteristics of satellites to capture specific types of confounding variables, where variables visibly discernible from space, such as trees, are more accurately represented. On the other hand, certain variables, such as air quality, are captured with a reduced accuracy by satellite imagery. The reduction of bias through the inclusion of images depends on the extent to which the relevant confounding variable is captured in the images.

Included variable bias

The inclusion of images can introduce new problems. One issue arises when images contain too much information, in particular if they can predict the outcome

Y or treatment variable T at a high level. The overlap assumption requires that the propensity score to remain away from 0 and 1, which ensures that the probability of treatment assignment is not entirely determined by covariates. Adding more variables for control can improve the accuracy of predicting treatment assignment, potentially leading to the violation of the overlap assumption. When the outcome variable is entirely dependent on information within the images, estimating the treatment effect becomes difficult. This problem can be assessed by checking how much image features predict Y and T.

Depending on the situation, controlling for variables can result in bias amplification phenomenon [16] [80]. One potential problem could manifest as M-bias, where certain covariates U_1 and U_2, although independent of one another, affect a common variable X and each impact T and Y. It has been studied that adjusting for such a covariate X may introduce bias [78] [57] [19].

Another potential issue is Z-bias, or instrumental variable bias. The situation arises when there is an instrumental variable X which affects the treatment T but not the outcome Y directly and there is an unmeasured confounding U. If X is not controlled for, then it can bias the coefficient on T. However, adjusting for X may result in a larger bias in the adjusted estimator [8] [92] [80] [52]. The stronger the instrumental variable, that is, if X is more predictive of T, the higher the relative bias becomes. Controlling for X can have an effect of removing essential variation in T.

Quantification of bias in omitted variable model

To understand the effect of image inclusion on the estimate of interest, we characterize the bias of the estimate under two scenarios: 1) when $\mathbf{U}$ is omitted and 2) when $\mathbf{R}$, a remotely sensed proxy of $\mathbf{U}$, is included. Quantifying bias allows us to gain insight into conditions under which controlling for image features is desired.

Bias when latent confounding covariate is omitted

Recall that if the confounding covariate $\mathbf{U}$ is not measured, then one can fit the model with treatment T and observed covariates $\mathbf{Z}$:

$$Y_i = \beta_0^* + T_i\beta_T^* + \mathbf{Z}_i\,\beta_Z^* + \epsilon_i^*. \tag{2.4}$$

To understand the bias of β_T^*, it helps to define a third regression on the omitted variable:

$$U_i = \gamma_0 + T_i\gamma_T + \mathbf{Z}_i\,\gamma_Z + \nu_i. \tag{2.5}$$

If we substitute this representation of U into the correct equation and rearrange terms, we get

$$Y_i = (\beta_0 + \beta_U\gamma_0) + (\beta_T + \beta_U\gamma_T)T_i + (\beta_Z + \beta_U\gamma_T)Z_i + (\epsilon_i + \beta_2\nu_i) \tag{2.6}$$

Equating the coefficients of T in (2.1) and (2.6) yields the omitted variable bias term

$$\beta_T^* - \beta_T = \beta_U\gamma_T, \tag{2.7}$$

which is large when there is a large association between the treatment and the confounder (i.e. large γ_T) and/or if there is a large association between the outcome and the confounder (i.e. large β_U).

Bias when image features are included as proxy

Next, we quantify the bias when image features $\mathbf{R}$ are controlled as proxy for unmeasured confounders $\mathbf{U}$. Because $\mathbf{R}$ is a function of T, $\mathbf{Z}$, and $\mathbf{U}$ with some noise, it helps to consider variables T, $\mathbf{Z}$, and $\mathbf{U}$ as comprised of two components: variation that is estimated by $\mathbf{R}$ and the remaining variation that is not estimated by $\mathbf{R}$. For example, $T = T_1 + T_0$, where T_1 denotes part of the treatment variation that is estimated by image features and T_0 denotes the the residual. We refer T_1 as visible variation and T_0 as invisible variation of T.

Recall the three models and our set up:

$$\begin{aligned}
Y_i &= \beta_0 + T_i\beta_T + Z_i\beta_Z + U_i\beta_U + \varepsilon_i && \text{Correct Model}\\
Y_i &= \beta_0^* + T_i\beta_T^* + Z_i\beta_Z^* + \varepsilon_i^* && \text{Omitted Variable Model}\\
Y_i &= \beta_0' + T_i\beta_T' + Z_i\beta_z' + R_i\beta_R' + \varepsilon_i' && \text{Proxied Model}\\
\\
U_i &= \gamma_0 + T_i\gamma_T + Z_i\gamma_Z + \nu_i && \text{Confounder Model}\\
T_i &= T_{1,i} + T_{0,i} && \text{Variable Decomposition}
\end{aligned}$$

We know that the true effect of T is β_T and the omitted variable bias is $\beta_U\gamma_T$. Now, we want to quantify the bias of $\beta_T^{'}$, which is the bias incurred by including proxy. By Frisch–Waugh–Lovell theorem [25] [46], we know that the estimate of $\beta_T^{'}$ is equivalent to those obtained by running the following regression:

$$\varepsilon^Y = \beta_T'\varepsilon^T + \varepsilon', \tag{2.8}$$

where

$$\begin{aligned} \varepsilon^Y &= Y - \hat{\beta}_Z^Y Z - \hat{\beta}_R^Y R \\ \varepsilon^T &= T - \hat{\beta}_Z^T Z - \hat{\beta}_R^T R \end{aligned}$$

ε^Y and ε^T denote the residual of Y and T, respectively, that are not predicted by observed covariates Z or proxy R. $\hat{\beta}_Z^Y$ and $\hat{\beta}_R^Y$ are the regression coefficient estimates of Z and R from running the regression:

$$Y = \beta_Z^Y Z + \beta_R^Y R + \epsilon^Y$$

Similarly, $\hat{\beta}_Z^T$ and $\hat{\beta}_R^T$ are the regression coefficient estimates of Z and R from running the regression:

$$T = \beta_Z^T Z + \beta_R^T R + \epsilon^T$$

Therefore, by design, $\varepsilon^Y = Y_0$ and $\varepsilon^T = T_0$. Eq. (2.8) can be rewritten as

$$Y_0 = \beta_T' T_0 + \varepsilon'$$

We can compute the bias of β_T' in the same way as we did with the omitted variable bias and obtain that it is $\beta_T' - \beta_T = \beta_{U_0}\gamma_{T_0}$, where β_{U_0} and γ_{T_0} are regression coefficients int he following relationships between invisible variations of Y, U, T, and Z:

$$\begin{aligned} Y_0 &= \beta_{00} + \beta_{T_0} T_0 + \beta_{Z_0} Z_0 + \beta_{U_0} U_0 + \eta_0 \\ U_0 &= \gamma_{00} + \gamma_{T_0} T_0 + \gamma_{Z_0} Z_0 + \nu_0 \end{aligned}$$

The bias of the proxy model is high when the invisible variation of the treatment T_0 is highly associated with the invisible variation of Y and/or U. That is, even if imagery captures a large extent of the variations of U, if there is a common variation in Y, T, and U that is not captured by imagery, then the bias can be high. On the other hand, if a large extent of the variation in confounding U is captured by imagery and the invisible variation U_0 to be close to noise, then its association with T_0 would be minimal and thus we can expect the bias to be small.

Condition in which image inclusion reduces bias

The bias of the treatment effect has been characterized when a confounding variable is omitted from regression and when it is proxied by imagery. By comparing the two biases, it can be observed that the bias after controlling image features is smaller than the omitted variable bias when the following condition is satisfied:

$$|\beta_{U_0}\gamma_{T_0}| < |\beta_U\gamma_T|, \tag{2.9}$$

where β_U is the association between U and Y, β_{U_0} is the association between invisible variations U_0 and Y_0, γ_T the association between T and U, γ_{T_0} the association between invisible variations T_0 and U_0.

Conceptually, image features reduce bias of the treatment effect estimate if including them in the regression makes the treatment variable and/or the outcome variable *less confounded*, that is, less associated with the confounder. In other words, even if imagery captures a large extent of the variations of U, if there is a common variation in Y, T, and U that is not captured by imagery, then including imagery can result in high β_{U_0} and/or high γ_{T_0}, potentially making the bias larger than when the imagery was not included.

To provide an alternative representation, a simulation is carried out to visualize the theoretical condition outlined in Equation 2.9. The simulation illustrates the influence of including image features $\mathbf{R}$ on bias in relations to varying values of γ_T, γ_{T_0}, β_U, and β_{U_0}.

Simulation example
Consider a simple setup with a single treatment T and a single confounder U.

1. Visible components T_1 and U_1 were generated as functions of image features R:

$$\begin{aligned} T_1 &= \alpha_t R \\ U_1 &= \alpha_u R, \end{aligned}$$

with

$$\begin{pmatrix} \alpha_t \\ \alpha_u \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_1 \\ \sigma_1 & 1 \end{pmatrix} \right)$$

2. Invisible components T_0 and U_0 were drawn from the bivariate normal distribution:

$$\begin{pmatrix} T_0 \\ U_0 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_0 \\ \sigma_0 & 1 \end{pmatrix} \right)$$

3. Variables T and U were defined as a linear sum of visible and invisible components:

$$\begin{aligned} T &= n_1 T_1 + n_0 T_0 \\ U &= m_1 U_1 + m_0 U_0, \end{aligned}$$

where n_1, n_0, m_1, and m_0 are constants.

The outcome variable Y was defined as a linear combination of T and U with some noise:

$$Y = \beta_T T + \beta_U U + \varepsilon$$

4. Parameters β_{U_0}, γ_T, and γ_{T_0} were estimated by running the following regressions:

$$\begin{aligned} U &= \gamma_0 + \gamma_T T + \nu \\ U_0 &= \gamma_{00} + \gamma_{T_0} T_0 + \nu_0 \\ Y_0 &= \beta_{00} + \beta_{T_0} T_0 + \beta_{U_0} U_0 + \eta_0 \end{aligned}$$

5. Treatment effect was estimated under the following two specifications:

$$\begin{aligned} Y &= \beta_0 * + \beta_T^* T + \epsilon^* \\ Y &= \beta_0{}' + \beta_T^{'} T + \beta_R^{'} R + \epsilon^{'} \end{aligned}$$

The procedure was carried out under all combinations of the following parameter values:

$$\begin{aligned} \beta_T = \beta_U &= 1 \\ \sigma_1, \sigma_0 &\in \{-0.9, 0.8, ..., 0.8, 0.9\} \\ m_1, m_0 &\in \{0, 0.1, ..., 1.9, 2\} \end{aligned}$$

σ_1 and σ_0 are varied as σ_1 modulates the strength of association between the visible components of T and U, while σ_0 governs the association between their invisible components. m_1 and m_0 are varied as they influence the amount of variation in U that is captured by image features.

Simulation Result

In general, including image control reduces bias in comparison to not controlling for images, when the relationships between the invisible variations of T and U or the invisible variations of Y and U are relatively weak.

The condition $\beta_{U_0} = \beta_U$ illustrates a situation in which the degree of association between the invisible variations U_0 and Y_0 is the same as the association between U and Y, which occurs when $m_1 = 1$. In this context, including images reduces bias when the association between the invisible variables U_0 and T_0 is weaker than the

association between U and T ($\gamma_{T_0} < \gamma_T$), which occurs in approximately half of the data points in the simulation (Figure 2.2). Conceptually, $\gamma_{T_0} < \gamma_T$ when including images explains shared variations in U and T.

The condition $\beta_{U_0} > \beta_U$ illustrates a situation in which the degree of association between the invisible variations U_0 and Y_0 is greater than the association between U and Y, which occurs when $m_1 > 1$. In this context, the association between the invisible variations must to be considerably weaker than the association between U and T, which occurs less than half of the data points int he simulation. The risk of amplifying bias is higher compared to the case when $\beta_{U_0} = \beta_U$, and the degree of amplification can be large.

The condition $\beta_{U_0} < \beta_U$ illustrates a situation in which the degree of association between the invisible variations U_0 and Y_0 is less than the association between U and Y, which occurs when $m_1 < 1$. In this context, there are considerably more cases of bias reduction due to image inclusion. Even when the inclusion amplifies bias, the extent of amplification is relatively minor.

The issue with the theoretical condition described in Equation 2.9 and illustrated in the simulation is it requires the knowledge of unobserved confounders U, and therefore, it is not practical to confirm whether the condition is satisfied using available data. A perpetual goal is the design of diagnostic checks to assess in advance whether incorporating image information contributes to the accuracy of the estimate.

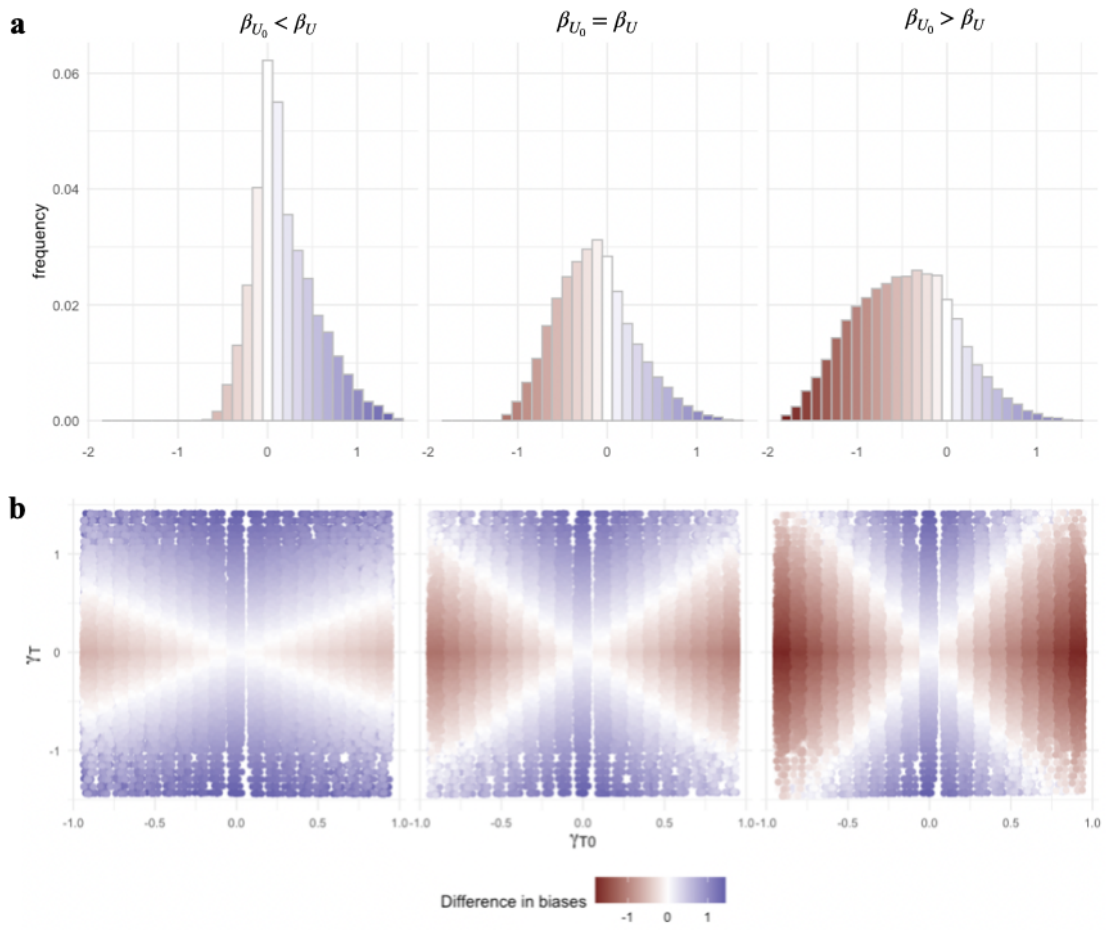

Figure 2.2: **Differences in biases computed under varying degrees of associations between γ_T, γ_{T_0}, β_U, and β_{U_0}.** (A) Distribution of the differences biases under three scenarios: 1) $\beta_{U_0} < \beta_U$, 2) $\beta_{U_0} = \beta_U$, and 3) $\beta_{U_0} > \beta_U$. The difference in biases is calculated as omitted variable bias − image proxied bias. (B) The relationship between γ_T and γ_{T_0} under the same three scenarios. The colors denote differences in biases; blue represents positive values when including images reduces bias, white indicates no change in bias with the inclusion of images, and red indicates the inclusion of images amplifies bias.

Simulated experiment on the effect of controlling imagery in relation to omitted variables

An additional simulated experiment is conducted to examine how the inclusion of images affects bias across different numbers of omitted variables. This involves a hybrid simulation that generates a true outcome model using empirical data containing covariates and image features. The effect of treatment variable is estimated under different model specifications, categorized into two groups: one without image features and the other including image features.

Simulation example: A Single True Outcome Model with Varied Model Specifications. One variation of the simulation involves defining an outcome model as a linear combination of the treatment and all p covariates. Subsequently, a comprehensive list of model specifications is generated by systematically witholding covariates in all possible combinations (e.g., temperature, temperature + precipitation, temperature + precipitation + tree cover, and so on). With p covariates in place, this process generates 2^p model specifications. The treatment effect is estimated within each of these model specifications, both with and without the inclusion of image features R. The model specifications are grouped based on the number omitted variables, or the number of covariates withheld in the analysis. Within each group of omitted variables, various model specifications are evaluated, resulting in distributions of two biases: omitted variable bias and the image-proxied bias. For each number of omitted variables, the average difference between the omitted variable bias and the image-proxied bias is computed. This comparative analysis is used to assess the effect of the inclusion of image features across varying number of omitted variables.

Figure 2.3 suggests that as more covariates are omitted from the model, the difference between omitted variable bias and image-proxied bias tends to increase. The difference is positive across all treatment variables, indicating that the inclusion of image features tends to reduce bias compared to the omission of variables. The magnitude of this difference varies depending on the specific treatment variable under consideration. For example, for certain treatment variables such as elevation, dew point, and building age, the reduction in bias incurred by image features is relatively large. For other variables such as house units, vacant house percentage, and flood risk score, the reduction in bias is comparatively low, even when most of the covariates are omitted from the analysis.

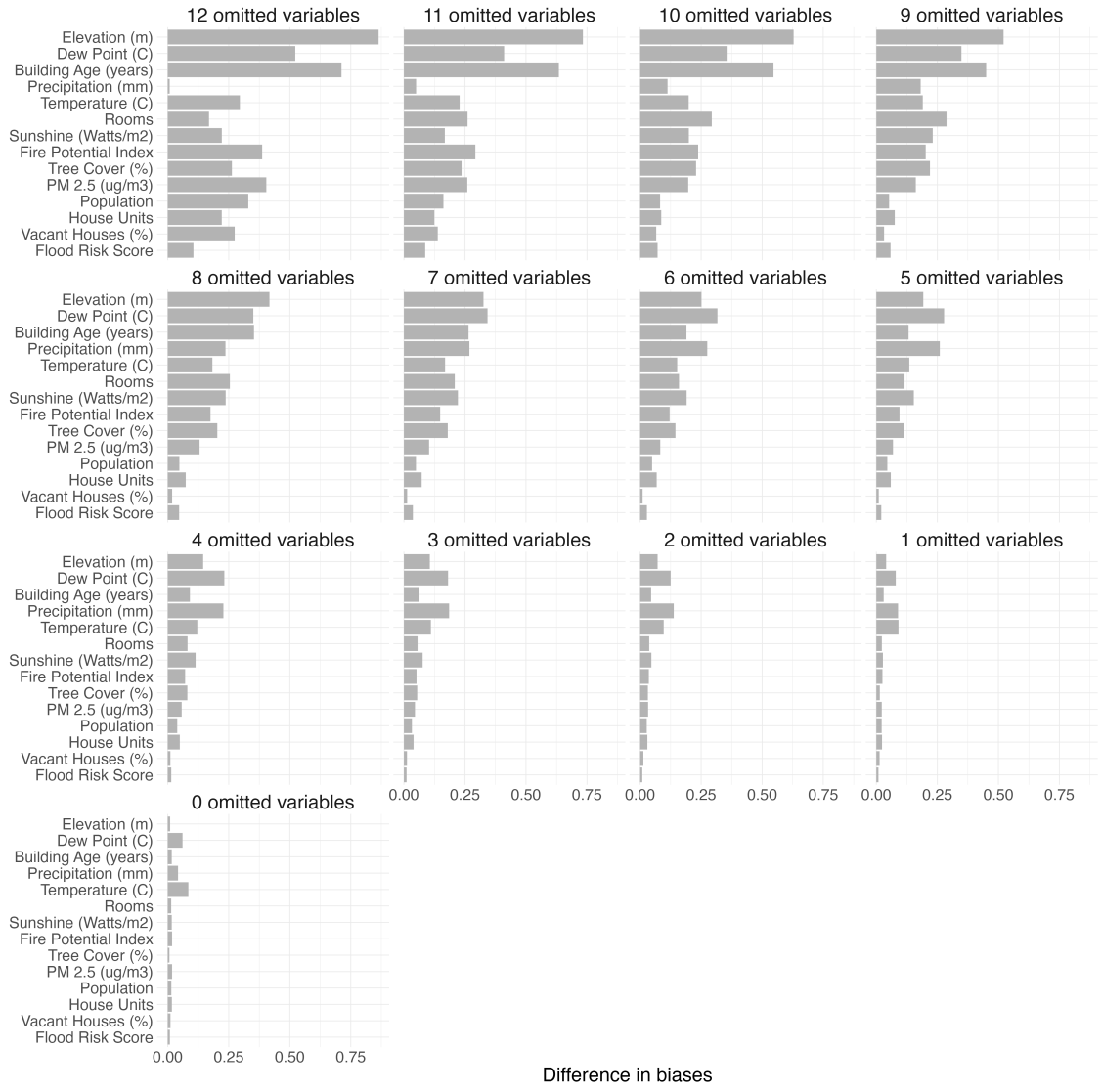

Figure 2.3: **Differences in biases computed for varying number of omitted variables and treatment variables.** The difference in biases is computed as omitted variable bias subtracted by image-proxied bias. The overall trend shows an increase in the difference between omitted variable bias and image-proxied bias as more covariates are omitted from the model.

2.5 Conclusion

In this chapter, we introduced the potential for incorporating information obtained from satellite imagery, specifically as a proxy for confounding variables, which previously may have been difficult to measure using other sources of measurements. One approach involves integrating image information by using image features as a control in a regression framework. The study examines the impact of incorporating image features, considering their potential to either reduce or amplify bias.

Theoretically, the inclusion of image features reduces bias of the variable of interest when incorporating these features in the regression makes the treatment variable and/or the outcome variable less confounded, meaning they are less associated with the confounder. However, if there is common variation in the outcome, treatment, or confounder that is not captured by imagery, including imagery can potentially amplify bias, even if it captures a significant portion of the variation in the confounding variables.

Since the theoretical condition requires the knowledge of the unmeasured confounding variable and cannot be verified using empirical data, one of the perpetual goals is to formulate a diagnostic study. The goal is to assess whether image features contribute to the accuracy of estimating the parameter of interest. One experiment which can be conducted using empirical data is to generate a true outcome model using empirical data and image features, then assessing whether the inclusion of images, on average, reduces bias. Such a simulated experiment is carried out using environmental and housing data. The result indicates that including imagery is likely to reduce bias, especially when there is a high number of omitted variables.

While the primary focus of this chapter was on examining the impact of including image features on bias of the estimate of interest, it is pertinent to study the tradeoff between bias and standard error. Additionally, though the imagery information was controlled within the context of a regression framework, an avenue for future work involves exploring matching techniques based on images. For example, matching land units based on their satellite images could be used as a way to pair units with similar physical land characteristics. This introduces considerations for measuring image similarity, identifying key factors contributing to such similarity, and exploring the relationship between image similarity and confounding variables.

Chapters 3 and 4 provide an empirical application of the proposed regression method using satellite imagery. The application focuses on estimating the economic value of environmental qualities.

Part II

Estimating the Economic Value of Environment

Chapter 3

History of Economic Valuation of the Environment

3.1 Introduction

The frequency and cost of weather and climate disasters are on the rise globally, including in the United States, driven by a combination of factors such as increased frequency of extreme climate events, heightened exposure, and vulnerability, with a significant role played by the concentration of people and properties in high-risk regions.

For instance, in 2018, approximately 42% of the US population resided in coastal shoreline counties, which cover around 10% of the continental US landmass and are prone to known risks of coastal storms and flooding. The population living in areas at risk of wildfires has grown, with estimates showing an increase from 10 million in 1990 to 20.8 million in 2010 [67]. The rise has been in part attributed to the emerging expansion of the wildland-urban interface, where human-made structures and flammable vegetation converge, thereby increasing the risk of wildfires [67] [64].

While the reasons for the increased exposure is multifaceted and responsibility is shared across multiple parties [42], one of the few domains where climate risks and broader environmental qualities are explicitly capitalized is housing markets. One perspective is that the housing finance system has the potential to incorporate the risks associated with specific regions and can play a role in discouraging risky behaviors, such as rapid development of properties in high climate-risk areas.

With the increasing significance of risk management, it is crucial to gain insights into the current status of how these environmental and climate conditions are integrated into property values at a national scale. Given the interconnected nature

of these risks, it is also important to study these variables simultaneously. Prior research on capitalization has primarily focused on specific regions or environmental quality, yielding mixed results. Additionally, a methodological challenge lies in adequately accounting for all confounding variables, an aspect for obtaining accurate estimates of capitalization.

Our study contributes to the existing body of research by estimating the current state of the extent to which various environmental qualities are incorporated or overlooked on a nationwide scale. The environmental qualities we study include: flood risk, storm, precipitation, surface water, air pollution, wildfire risk, temperature, sun light, dew point, elevation, and tree cover. In estimating the capitalization, we use regression analysis that incorporates satellite data, as introduced in Chapter 2.

The rest of the chapter is structured as follows: Section 3.2 provides a summary of the current understanding of how environmental and climate information is capitalized in residential properties. Section 3.3 examines a commonly used hedonic property model, its historical context, and its implications for policy. Section **??** formulates the capitalization problem. Lastly, Section 3.4 introduces the data we use to study the problem, which encompasses housing, neighborhood, environmental, and satellite imagery data. This section also details the data processing steps taken to prepare for the modeling process.

3.2 Valuation of environment

Among climate risks, flood risk has been the subject of considerable research, and a few nationwide studies have been conducted to evaluate the extent to which existing residential properties reflect flood risk. For instance, a study by Hino and Burke in 2021 [33] examined the impact of flood risk information provided by the federal government on property values. The findings indicated that houses located in flood zones are currently overvalued by a total of $43.8 billion (95% confidence interval: $32.6 to $55.6 billion). This overvaluation was calculated by comparing the empirical findings on the flood zone discount with the estimated efficient flood zone discounts, which represent the cost of full insurance. Furthermore, a study conducted by Gourevitch et al in 2023 [28] estimated the overvaluation to be between $121 to $237 billion, after taking into account future climate change impacts. In a meta-analysis conducted by Beltran et al [6], it was found that many studies focusing on localized effects within a single county or city suggest a price discount for properties located in floodplains. However, the results of these studies show significant variation, ranging from a discount of -75.5% to +61.0% price premium.

Other climate risks and broader environmental qualities have also been the focus of extensive studies. While not an exhaustive overview, we highlight some recent findings from studies conducted in these areas. Regarding wildfire risk, specific regional studies have indicated that properties located near hazardous topography tend to be associated with higher sales prices [13], while the provision of information shocks on wildfire can elevate risk perceptions and result in reduced housing prices [45], although the effect may be short-term [49].

3.3 Hedonic model

Origins and overview of hedonic model

The hedonic property model is a widely used approach in valuation studies for estimating the economic value of the environment. Initially, the hedonic pricing model was introduced in 1929 by Frederick V. Waugh, an agricultural economist [91]. Waugh conducted research on vegetable pricing, specifically focusing on asparagus. He used regression analysis to examine how the price of asparagus correlated with three qualities: color, size of stalks, and uniformity of spears. The primary motivation behind this study was to understand customers' preferences for these characteristics, which held interest for producers in the agricultural industry.

During the mid-1960s, the application of hedonic analysis expanded to include the economic valuation of the environment. Federal agencies funded research projects aimed at estimating people's willingness to pay for improved air quality, using data from residential property values. In 1967, Ridker and Henning published a paper that initiated discussions about using real estate prices as a means to estimate the economic value of air quality and, by extension, environmental quality [70].

Since the 1970s, there has been a surge in the number of publications focusing on hedonic pricing and environmental valuation. In this field of research, hedonic analysis commonly adopts a regression approach, assuming that the price of a house is a composite of various implicit prices associated with its individual characteristics. These characteristics may include: (a) Physical structure of the house, e.g. the number of rooms and the age of the building; (b) Surrounding neighborhoods, e.g. the socioeconomic status of the area and the availability of public services; (c) Environmental conditions, e.g. the air quality and the presence of green spaces. A regression analysis of housing prices in relation to these characteristics provides estimates of the marginal implicit prices attributed to each particular characteristic of interest.

Social implications of hedonic model

Estimates derived from hedonic models have had some practical applications beyond the research community. Case studies conducted by Palmquist and Smith [54] indicate that hedonic studies can serve as a basis for establishing and assessing improvements in various environmental standards. In litigation contexts, hedonic models have provided legal framework for assessing damage appraisals.

In policymaking, hedonic analysis has played a role in the examination of air pollution damages, providing motivation for improving air quality and setting national ambient standards. For instance, the 1970 report to the National Air Pollution Control Administration in the U.S. Department of Health, Education, and Welfare used estimates from four property value studies to derive a national annual cost estimate of sulfate air pollution. In addition, the 1974 report titled "Air Quality and Automobile Emission Control," published jointly by the National Academy of Science and the National Academy of Engineering, featured several research studies focused on the welfare effects (i.e., non-health consequences) of air pollution, which collectively provided the basis for the development of national ambient standards [54].

Hedonic analyses have been utilized in public litigation cases involving residential properties and the determination of liability for environmental impacts. In such litigation, the aim is often to seek compensation for matters of public interest. However, the records of these analyses are not always readily accessible to the public, and obtaining statistics on the frequency of hedonic analysis usage in such cases can be challenging. Nonetheless, an illustrative case study presented in [54] offers an example of its application in this context.

In summary, the case involved the pollution of a river by hazardous substances stemming from an abandoned silver mine in Colorado during the year 1985. The litigation concerned the cost of damages that the mine was liable. The defense computed the estimated loss incurred by homeowners due to the inability to use well water, using the price of bottled water as a reference. However, the lack of clean water is just one facet of the problems caused by the pollution. The plaintiff's expert conducted hedonic analysis to estimate the market value that was impacted by the pollution. This involved regressing the reported sales prices of 150 properties against a binary proximity variable, indicating whether each property was located within six miles of the mine site. Additionally, five covariates related to the structural housing attributes were included in the analysis, serving to control for potential confounding variables. The hedonic analysis produced an estimated loss that was two and a half times larger than the mitigation cost proposed by the defense. This discrepancy, where the mitigation cost was significantly lower than the estimated loss of market value, raised doubts about the plausibility of the defense's estimate. Consequently,

this prompted a thorough examination of other claims presented in their report.

Though not exhaustive, examining case studies that use hedonic models offers insights into the nuances of how estimates from these models can serve in discussions. In specific contexts, like public litigation addressing a particular waste disposal incident, hedonic analysis can directly impact actions by offering quantifiable and manageable assessments of damages. In more general scenarios, such as national policies involving diverse regions and populations with varying characteristics, the impact tends to be more indirect, serving as one aspect of the discussion.

Challenges of hedonic model

One of the challenges encountered with hedonic property models is the difficulty in effectively controlling for confounding variables [15] [55] [50] [1]. When these variables are not accounted for or exhibit spatial correlation, the model can be prone to omitted variable bias. Hedonic analyses are susceptible to confounders, which may encompass physical, environmental, and socioeconomic characteristics of the neighborhood. Some of these variables may be latent in nature, posing challenges in direct measurement. For instance, higher levels of air pollution may be linked to the urban form, which encompasses the configuration of buildings, transportation corridors, and other structural elements that shape the physical aspects of a city. Failing to adequately control for such variables can lead to either over- or underestimation of the implication of air pollution on housing prices.

3.4 The ACS, environmental, and satellite data

Data

This section describes datasets we use to estimate the extent to which housing values capitalize various environmental qualities. Factors were chosen to represent housing attributes and their surrounding environment, subject to the condition that high resolution and up-to-date data are available across the continental US. For environmental variables that are known to exhibit temporal variability, we use the normals data that reflect the average condition in the recent two decades. Below we describe the data sources.

Task	Units	Spatial resolution	Temporal period	Data source
Temperature	degrees Celsius	$\sim$ 4km $\times$ 4km	2000-2019	[62]
Precipitation	mm	$\sim$ 4km $\times$ 4km	2000-2019	[62]
Dew point	degrees Celsius	$\sim$ 4km $\times$ 4km	2000-2019	[62]
Sunshine	Watts / m^2	$\sim$ 111km $\times$ 111km	March 2000 - December 2019	[53]
Fire potential index	Index	1 acre	2001 - 2009	[86]
PM 2.5	$\mu g/m^3$	$\sim$ 1.11km $\times$ 1.11km	2000-2019	[5]
Tree cover	% tree cover	$\sim$ 30m $\times$ 30m	2010	[31]
Elevation	meters	$\sim$611.5m $\times$ 611.5m	2010	[4]
Surface Water occurrence	% water presence over time	$\sim$30m $\times$ 30m	March 1984 - October 2015	[22]
Wind speed	m / s	0.1° $\times$ 0.1°	1950 - 2008	[35]
Flood risk	score 1-10	census tract	2020	[3]

Table 3.1: **Data sources for environmental and climate variables.** The spatial resolutions mentioned for all raster data sets (forest cover, elevation) apply to grid cells located at the equator. Due to the Earth's curvature, the raster size in Euclidean distance will vary with latitude.

Temperature, Precipitation, Dew Point We use the normal data modeled by the PRISM Climate Group at the Oregon State University. The data reflect the mean temperature (degrees Celsius), mean precipitation (mm), and dew point (degrees Celsius) across the recent two decades for the period 2000 - 2019. The PRISM (Parameter-elevation Relationships on Independent Slopes Model) uses a regression-based spatial interpolation method that generates estimates at a resolution of 25 arcmin (approximately 4km) grid cells. The original data used in the estimation consists of nearly 10000 surface weather observation stations for temperature, 13000 for precipitation, and 4000 for dew point. Stations used in the model are weighted according to the physiographic similarity of the station to the grid cell. [18] [1]

We use dew point as opposed to relative humidity (RH), because dew point directly affects general human comfort levels outside [44]. Dew point is the temperature to which the air needs to be cooled to become saturated with water vapor, or RH of 100%. Higher dew point indicates higher amount of moisture in the air.

Sunshine To measure sunlight, we use shortwave flux down data at the Earth's surface, provided by The Clouds and the Earth's Radiant Energy System (CERES) Science Team. Shortwave flux down data is a satellite-based estimate of the flow of solar energy per unit area, measured in Watts/m^2. The estimate has a spatial resolution of 1 degree ($\sim$ 111km). We take the average of monthly data from March

[1] We accessed these data via the R function `get_prism_monthlys` from the `prism` package. Code and documentation can be found here: `https://cran.r-project.org/web/packages/prism/prism.pdf`

2000 to December 2019. [2]

Fire Potential Index We use wildland fire potential index (FPI) data maintained by the US Geological Survey (USGS). The index, ranging from 0 to 150, estimates two aspects of fire danger: the probability that a 1 acre ignition will result in a 100+ acre fire, and the probabilities of having at least 1, 2, 3, or 4 large fires [61] [3]. The estimates are available daily at 1km resolution. For this study, we use average flammability across the period 2001 - 2019. The estimates are produced based on a logistic regression model which relates historical fire occurrence data to vegetation and weather data. Vegetation inputs include the proportion of live to dead vegetation ad dead fuel moisture, which is the measure of amount of water in a fuel (vegetation) available to fire. Weather inputs include temperature, precipitation, and wind speed.

PM2.5 We use ground-level fine particulates matter (PM2.5) from Atmospheric Composition Analysis Group at Washington University in St. Louis. The data has a spatial resolution of $0.01° \times 0.01°$ ($\sim$ 1.11km $\times$ 1.11km) and reflects the average PM2.5 value over the period 2000-2019. The underlying data sources include a combination of satellite observations, chemical transport modeling, and ground-based measures. Satellite-derived aerosol optical depth (AOD) provides a measure of the amount of solar beam prevented from reaching the ground by aerosol particles. The chemical transport model relates AOD and ground monitor data, allowing for a spatially complete representation that is consistent with ground-based measurements [30] [87]. [4]

Tree Cover We use tree cover data from [31], which estimates the percentage of maximum tree canopy cover per 30m $\times$ 30m pixel in the year 2010. Trees in the data were defined as vegetation taller than 5m in height. These estimates are based on a regression model of growing season Landsat 7 ETM+ data as inputs. [5]

Elevation We use data on elevation provided by Mapzen, and accessed via the Amazon Web Services (AWS) Terrain Tile service and the Open Topography global datasets API. The data is available in raster format with resolution of different zoom

[2]The data is available at `https://ceres-tool.larc.nasa.gov/ord-tool/jsp/SYN1degEd41Selection.jsp`

[3]Fire Potential Index can be accessed at `https://www.usgs.gov/fire-danger-forecast/wildland-fire-potential-index-wfpi`

[4]The data is available at `https://sites.wustl.edu/acag/datasets/surface-pm2-5/`

[5]Data can be accessed from the University of Maryland, Department of Geographical Sciences and USGS at `https://glad.umd.edu/dataset/global-2010-tree-cover-30-m`

levels, ranging from 1 to 14. We use zoom level 8 (611.5 meters at the equator) to align the resolution to that of our satellite imagery. [6]

Mapzen Terrain Tiles are compilations of several major open data sets. The data that covers the continental US is based on the light detection and ranging (lidar) derived data, powered by 3DEP (3D Elevation Program) in the U.S. Geological Survey. Lidar data is collected using a laser scanner, typically mounted on an aircraft, which transmits pulses of light to the ground surface. The pulses are reflected back and their travel time is used to estimate the distance between the laser scanner and the ground.

Surface Water Occurrence We use water occurrence, produced under the Copernicus Programme [58]. The data measures the frequency (expressed in percentage) with which water was present on the surface from March 1984 to October 2015. Presence of water is estimated by classifying Landsat data as open water, as land or as a non-valid observation. Open water is defined as any stretch of water larger than 30m by 30m. [7]

Wind Speed We use surface-level exposure to tropical cyclone winds derived from the Limited Information Cyclone Reconstruction and Integration for Climate and Economics (LICRICE) [35]. The data is the average (across years) maximum wind speed for all 6,712 storms during 1950-2008. The observed cyclone data is provided by the International Best Track Archive for Climate Stewardship (IBTrACS) [8]. The unit of measurement is meters per second (m/s) and the spatial resolution is $0.1^\circ \times 0.1^\circ$ (11.1 km $\times$ 11.1km). The surface-level exposure provides an estimate of the "storm experience" of individuals on the ground.

Flood Risk We use flood risk indicator data offered by First Street Foundation, available through AWS Data Exchange. The flood risk score, which ranges from 1-10, encapsulates both the likelihood and the severity of flooding due to rainfall (pluvial), riverine flooding (fluvial), and coastal surge flooding. The risk scores are discretized based on the First Street Foundation Flood Model, which shows the distribution of expectation of flooding in the current year and in 30 years. The data are aggregated at different regional levels. For this study, we use the data that is aggregated at the census tract level.

[6] We accessed these data via the R function `get_elev_raster` from the `elevatr` package. Code and documentation can be found here: `https://cran.r-project.org/web/packages/elevatr/elevatr.pdf`

[7] Water data is available at `https://global-surface-water.appspot.com/download`

[8] These data are available through the National Climate Data Center at `https://www.ncei.noaa.gov/products/international-best-track-archive`

The model relies on the combination of multiple models and data sources; ... [9]

ACS variables Housing attributes and neighborhood characteristics are represented by 7 variables from the American Community Survey. [10] The survey is conducted every year across the US, covering a broad range of topics about social, economic, demographic, and housing characteristics of the US population. We study housing attributes and neighborhood characteristics that are considered to act as indicators of the quality of housing and the housing market.

We use the 5-year estimates, which represents the average characteristics over 5-year period of time and are available at the block group level. The data contains a few missing values, which are caused by geographic restrictions, unacceptable statistical reliability, or the Census Bureau's Disclosure Review Board requirements. Taking complete cases preserves about 98% of the data. We calculate grid cell level values by computing the weighted average value of the variable across the grid cell. The ACS variables we use are listed and described in Table 3.2.

[9] Flood risk data was accessed by following the instruction `https://firststreet.org/data-access/getting-started-with-first-street-data/how-to-get-the-aggregate-data-from-aws/`

[10] We accessed ACS daa via the R function `get_acs` from the package `tidycensus` package. Code and documentation can be found here: `https://cran.r-project.org/web/packages/tidycensus/tidycensus.pdf`

Name	**Code**	**Description**
Median house value	B25077	For owner-occupied housing units.
Number of housing units	B25001	"A housing unit may be a house, an apartment, a mobile home, a group of rooms or a single room that is occupied (or, if vacant, intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live separately from any other individuals in the building and which have direct access from outside the building or through a common hall. Both occupied and vacant housing units are included in the housing unit inventory. Boats, recreational vehicles (RVs), vans, tents, railroad cars, and the like are included only if they are occupied as someone's current place of residence."
Percent vacant	B25002	"A housing unit is vacant if no one is living in it at the time of the interview, unless its occupants are only temporarily absent. In addition, a vacant unit may be one which is entirely occupied by persons who have a usual residence elsewhere"
Number of rooms	B25017	"For each unit, rooms include living rooms, dining rooms, kitchens, bedrooms, finished recreation rooms, enclosed porches suitable for year-round use, and lodger's rooms. Excluded are strip or pullman kitchens, bathrooms, open porches, balconies, halls or foyers, half-rooms, utility rooms, unfinished attics or basements, or other unfinished space used for storage."
Building age	B25035	Data reported is the median year structure built. Building age is calculated as 2020 – median year structure built.
Income	B07011	"Median income in the past 12 months (in 2020 inflation-adjusted dollars)."
Population	B01003	Total population

Table 3.2: **Description of variables from the American Community Survey (ACS) used in the analysis).** Quoted descriptions of variables are from: `https://www.socialexplorer.com/data/ACS2020_5yr/metadata/?ds=ACS20_5yr` and `https://www.census.gov/housing/hvs/definitions.pdf`

Grid definition

The grids are designed with a consistent resolution of $0.01^\circ \times 0.01^\circ$ equal-angle, equivalent to approximately 1.11 km $\times$ 1.11 km at the equator. The grids are positioned to provide comprehensive coverage across the entire continental US. As a result, there are $8,307,981$ individual grids, each uniquely positioned without any spatial overlap between them. We use the entirety of these grids in our study instead of the sampled grids, as our empirical objective is to obtain an estimate of nationwide capitalization. Each image grid is associated with distinct features and their corresponding variable values.

Obtaining variable values for each image grid cell

To derive variable values for each grid cell, we perform a spatial overlay of our raw variable data and the grid cells. Due to the variations in format and spatial resolution of the variable data, it is necessary to calculate values based on each specific variable. Here, we will outline the approach used for each variable.

The raw tree cover, elevation, fire potential index, PM2.5, surface water occurrence, and wind speed data are initially provided as rasters with the same or higher spatial resolution compared to our grid cells. In these specific variables, we compute the mean value by averaging all the pixels in the variable data that are located within the grid cell boundaries. This process allows us to obtain the mean values across the image grid cell.

The temperature, precipitation, dew point, sunshine, and flood risk data are available at a spatial resolution smaller than our image grid cells. We extract the value that corresponds to each image grid cell, resulting in variables that represent the raw value present across the image grid cell.

The ACS variables are provided at the block-group level, which can vary in total area compared to our image grid cells across different regions in the US. In certain areas, block-groups might be larger in total area than our image grid cells, while in other regions, they may be smaller. To treat both cases consistently, we employ a weighted averaging approach for these variables. The weights used in the calculation are determined by the normalized area of intersection between the image grid cell and the polygons representing the variable data. This method ensures that the variable values are adjusted to account for the varying sizes. The resulting variables indicate the area-weighted average values across the grid cell.

Satellite imagery

We use daytime image data obtained from Planet's Surface Reflectance Basemaps product from 2019 [59]. This data offers a range of monitoring frequencies and spatial resolutions. For our study, we selected data from quarter 3 with a spatial resolution of 4.77m $\times$ 4.77m at the equator. This particular time point was chosen due to its minimal ice coverage in the northern hemisphere, making it suitable for our analysis which focuses on the continental US.

The image data is processed by Planet to achieve consistency in radiance values across satellite sensors and to minimize the influence of clouds, haze, atmospheric effects, and other sources of image variability [60]. These processing choices enhance images to effectively capture diverse physical characteristics of the land, including forestry, vegetation, and land cover.

Featurization of satellite imagery

To create statistical summarization of imagery, we use the image featurization technique called Multi-task Observation using Satellite Imagery and Kitchen Sinks (MOSAIKS), as proposed by Rolf et al [73]. This method involves a one-time unsupervised image featurization process using random convolutional features [65]. These features can then be used for predicting ground conditions through linear regression, where the variable of interest is regressed against the features. The features for the Planet images used in our analysis can be accessed via API at https://www.mosaiks.org/access [12].

In the MOSAIKS framework, each satellite image grid cell is associated with a set of features, which measure the similarity between the image and smaller patches of imagery. The process of creating these features involves several steps.

First, a large sample of N satellite images is gathered. Additionally, a random sample of K patches of imagery is selected. These patches are smaller portions of the satellite images.

Next, each of these K patches is convolved over each satellite image, sliding the patch over the image and calculating the similarity at each position. The result is a set of pixel matrices, one for each patch, for each satellite image. These pixel matrices are then passed through a nonlinear activation function to obtain K activation maps. The activation function enhances the features by introducing nonlinearity and flexibly capturing relationships between the patches and the image. Each activation map signifies regions of similarity between the corresponding patch and the image.

Finally, the activation maps are averaged over the pixels, resulting in a K-dimensional feature vector for each image. Each element of the feature vector repre-

sents the average similarity between the corresponding patch and the image.

Selection of patch sizes

In our study, we represent each satellite image using 4000 features, which were generated by Sherman et al [77]. These features are generated using patches of varying sizes. Specifically, 4000 patches are being used, half of which have a size of $4 \times 4 \times 3$ pixels, while the other half have a size of $6 \times 6 \times 3$ pixels. The values in the patch size respectively indicate the width, height, and the number of spectral bands, which correspond to the red, green, and blue channels of the image. Smaller patches capture local-level image structures, whereas larger patches capture broader larger-scale structures. These specific patch sizes were selected from a range of patch sizes available, which ranged from $3 \times 3 \times 3$ to $10 \times 10 \times 3$. The chosen combination of patch sizes showed the highest performance in predicting socioeconomic and environmental variables at a global level. These variables include night light intensity, road length, and forest cover.

The patches themselves are randomly sampled from the empirical distribution of patches derived from our training dataset of satellite images. By drawing patches from the empirical distribution, as opposed to generating them randomly or drawing them from a fixed distribution, we are able to sample from the distribution of sub-images encountered within the dataset. This approach contributes to the high descriptive capabilities of MOSAIKS features.

Evaluation of MOSAIKS features

Performance of MOSAIKS has been examined across a range of socioeconomic and environmental variables, including forest cover, elevation, population density, nighttime lights, income, road length, and house price in the continental US. The results indicate that the MOSAIKS demonstrates competitive predictive performance while also offering faster computational speed compared to existing deep-learning methods [73].

Nevertheless, there exists important limitations regarding the ability of images to capture ground conditions comprehensively. When image features were used to predict various ground conditions, the model consistently exhibited reduced performance in predicting extreme values [73] [77]. There was a tendency to overestimate lower values and underestimate higher values. This behavior may be attributed to both the implementation of ridge regression, which favors predictions converging towards the mean due to the l_2 penalty, and the limited availability of observations

containing extreme values. This limitation may also stem from the inherent constraints associated with the information that can be observed from space.

Additionally, satellite images used in the study are cross-sectional and do not account for temporal changes. Therefore, it is likely that the images are more suitable for capturing confounding variables that are known to have relatively low time variability, while they are less appropriate for variables that exhibit rapid temporal variability.

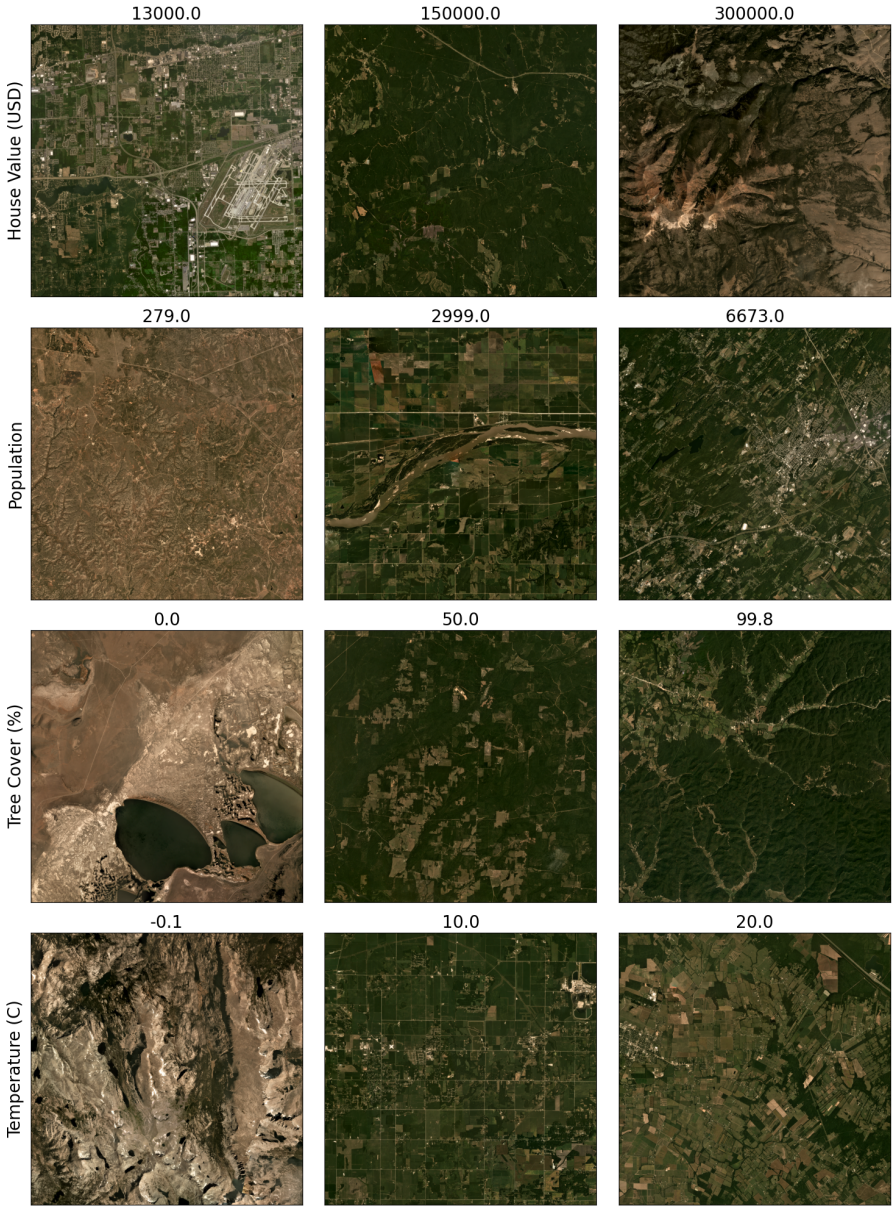

Figure 3.1: **Sample satellite images.** Each image is linked to distinct variable values. For instance, within the US sample, 12 images are chosen, with three images representing each of the four variables: house value, population, tree cover, and temperature. These images are ordered to display a spectrum of values, ranging from low, middle, to high for each variable.

Exploratory data analysis and data processing

While the majority of environmental and climate variables are spatially complete, approximately 2% of the ACS variables contain missing values. Due to the challenging nature of imputing these missing values, we have chosen to exclude these observations from the analysis. Additionally, to derive a general trend of capitalization, we have removed extreme values. This includes observations with buildings constructed before 1800. After applying these exclusions, the resulting dataset comprises x observations. The spatial distribution of each of the variables is shown in Figure 3.2.

To deal with strong skewness in certain variables (including house value, house units, vacancy, population, tree cover, and elevation, as shown in Figure 3.3), a log transformation was applied (Figure 3.4). Before the transformation, a value of 1 was added to each observation to handle cases where the original values were 0. The log transformation was selected to enable the interpretation of model coefficients as percent changes in housing value.

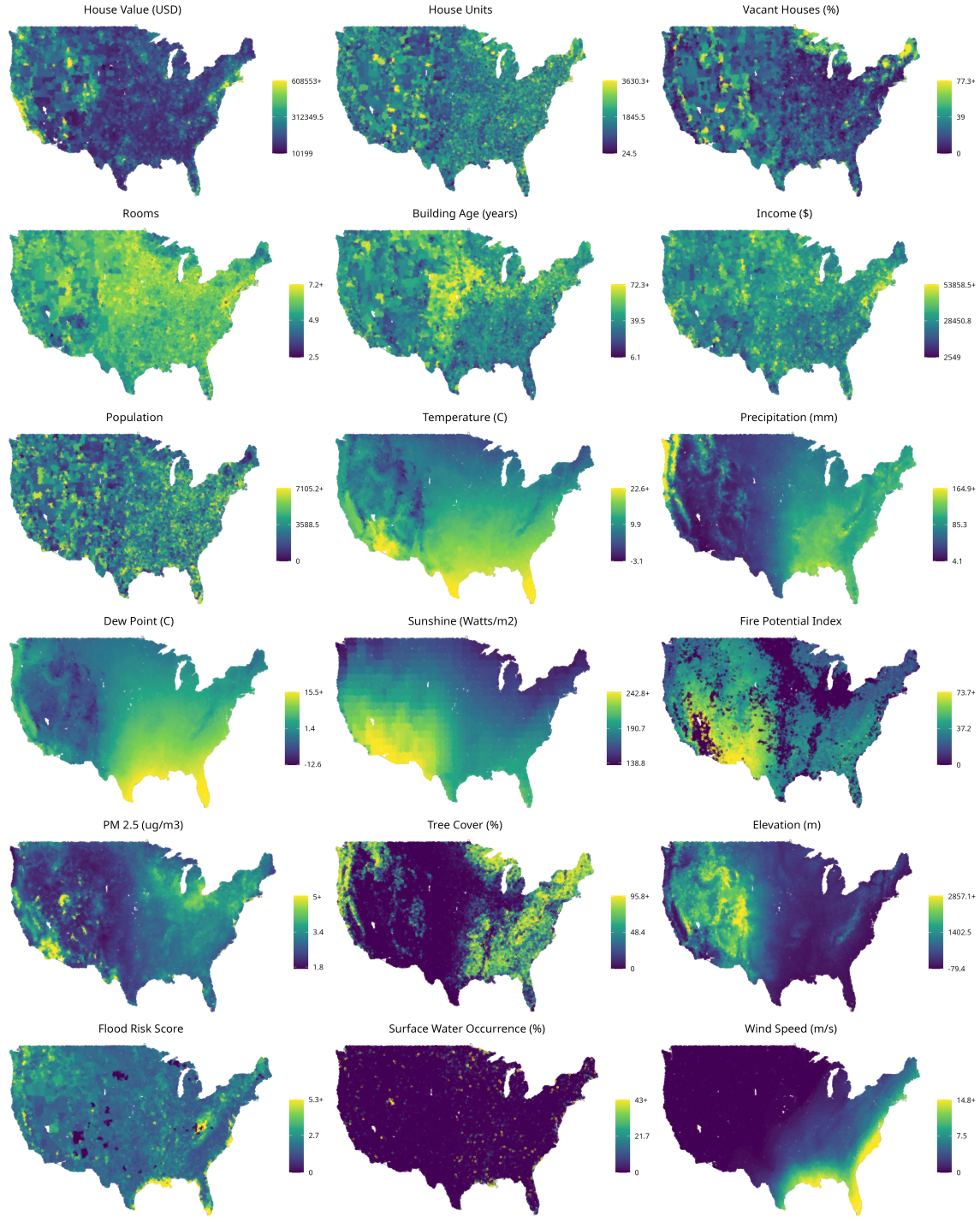

Figure 3.2: **0.01° × 0.01° resolution values of variables across the continental US.** There are total of 8,307,981 image grids that cover the continental US. The maps show values on randomly sampled one million grids for display.

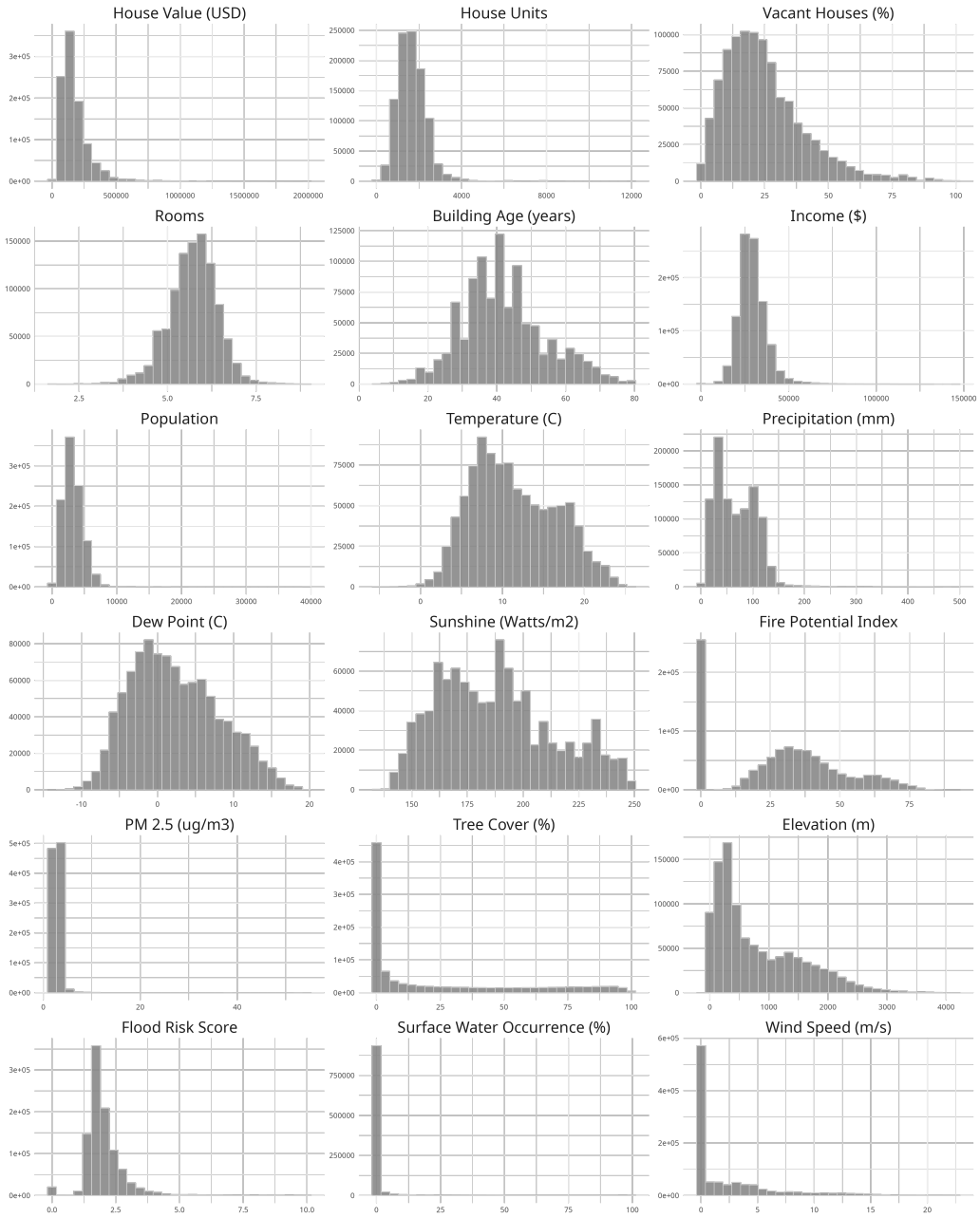

Figure 3.3: **Distribution of variables in raw values.** Histograms show the distribution of variables in their raw values across a sample of image grid cells. Among a total of 8,307,981 observations, we exclude missing and erroneous data, as well as the lower and upper 1% extremes, resulting in 7,614,441 observations. 1 million grid cells were randomly sampled for visualization purposes.

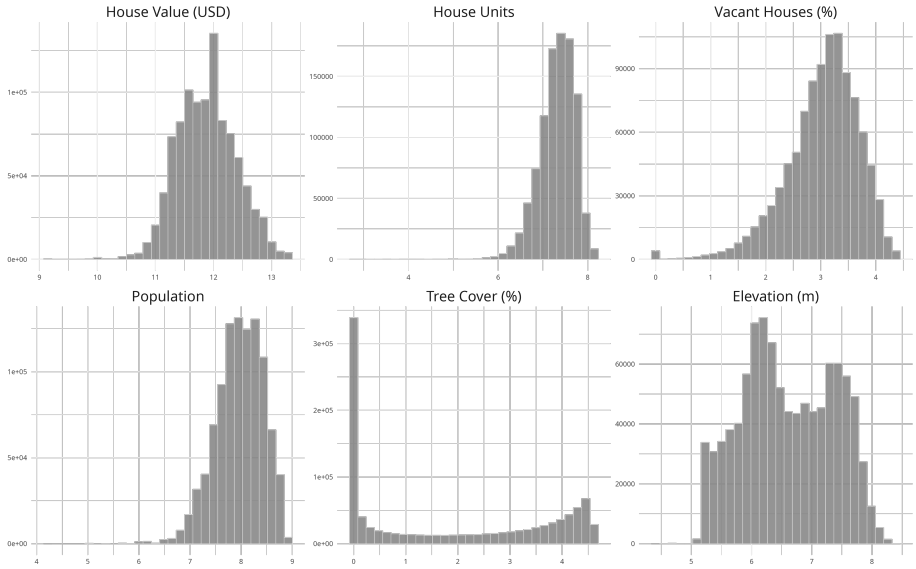

Figure 3.4: **Distribution of variables in transformed values.** Histograms show the distribution of log transformed variables. These variables are transformed due to the presence of skewness in the distribution of their raw values. For percent vacancy, population, tree cover, and elevation, log transformation was taken after adding 1, due to the presence of observations with a value of 0.

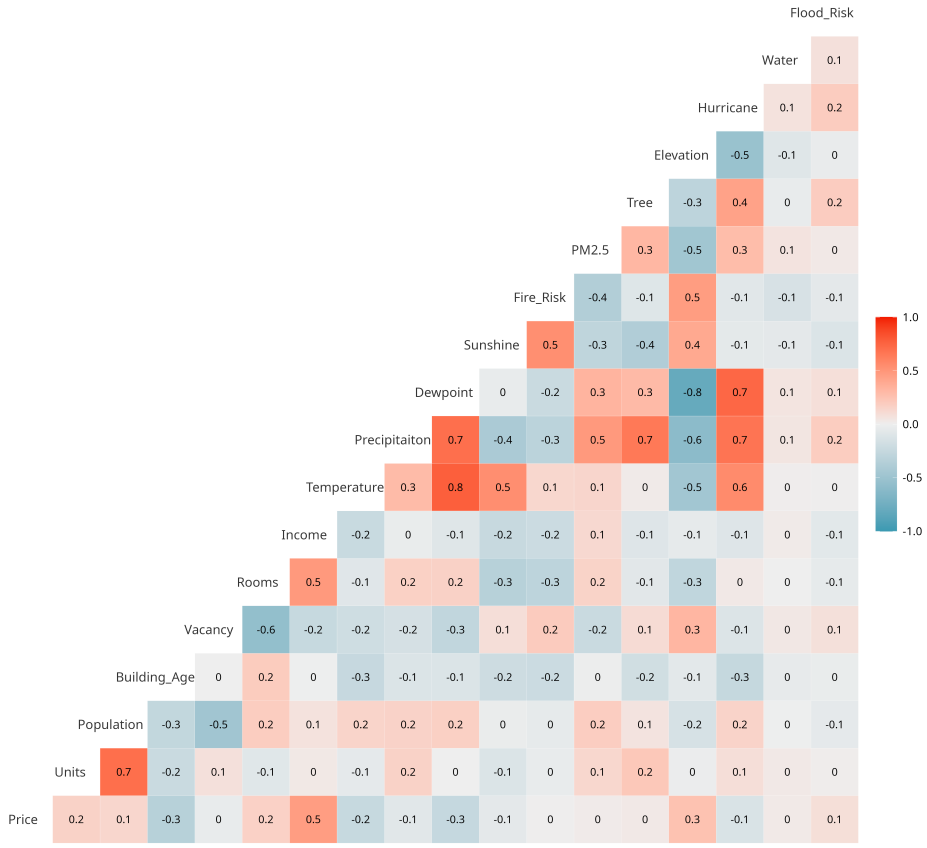

Figure 3.5: **Correlation of variables.** The matrix shows Pearson's correlation coefficient between all pairs of variables. The correlation coefficients are calculated using a random sample of 1 million grid cells.

Relationship between image features and variables

To examine the relationship between image features and the variables under study, we compute the proportion of variation in each variable that can be accounted for by image features alone. This is done by conducting ridge regression of a variable X against a vector of image features $\mathbf{R}$, visualizing the observed and predicted values of X, and computing R^2 as a measure of the explained variability (Figure 3.6).

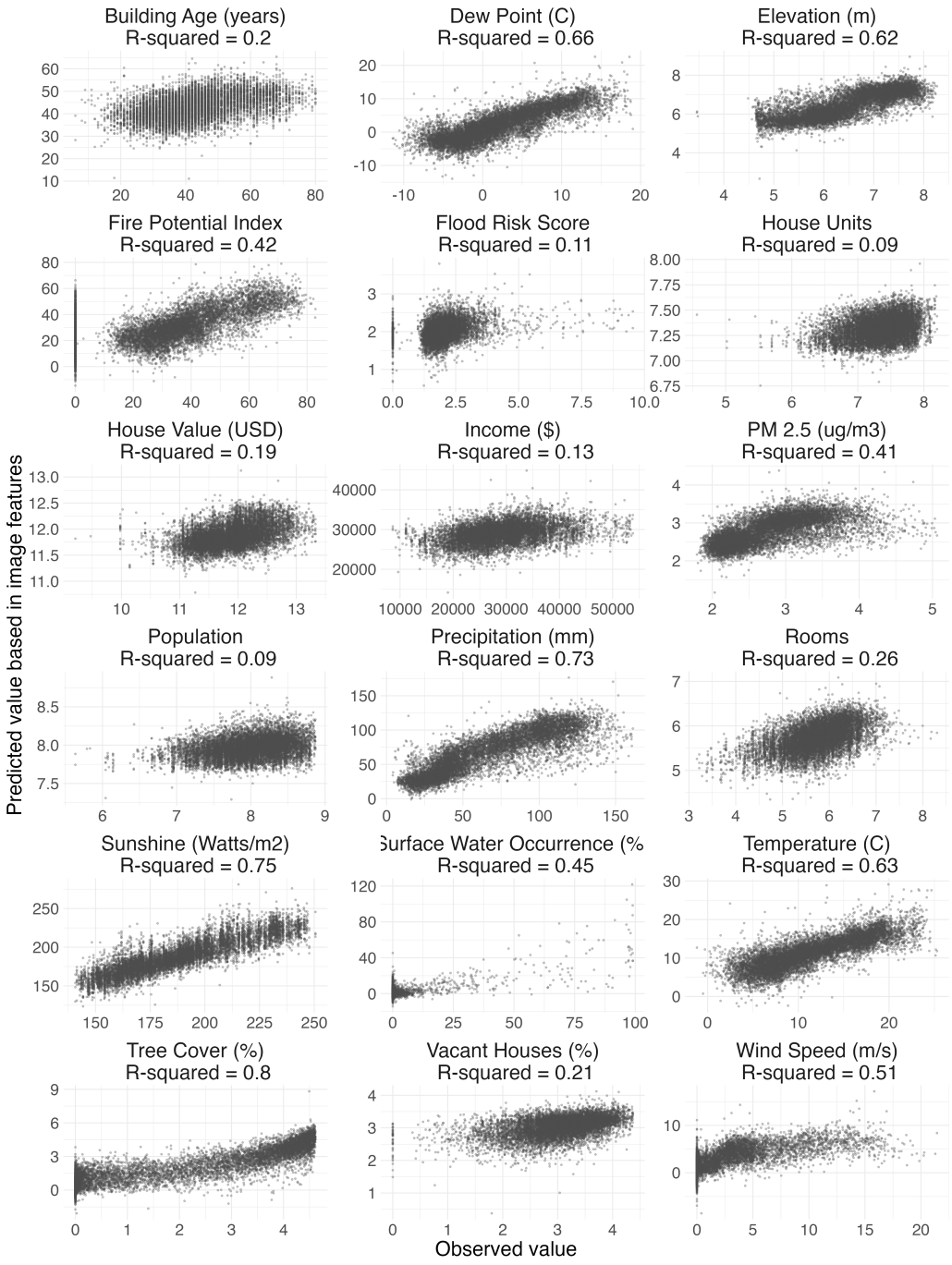

Figure 3.6: **Scatter plots of observed values (horizontal axis) and predicted values (vertical axis) for each of the 18 variables under study.** The predicted values represent the variable values estimated from image features alone. R^2 indicates the proportion of variation in the variable explained by image features. Randomly sampled 400000 points are shown for visualization.

3.5 Conclusion

While the reasons for the increased exposure to climate risks are multifaceted and responsibility is shared across multiple parties, one of the few domains where climate risks and broader environmental qualities are explicitly capitalized is housing markets. An aspect to consider is that the housing finance system has the potential to factor in the risks associated with specific regions and can contribute to discouraging risky behaviors, such as the rapid development of properties in areas with high climate risks. With the increasing significance of risk management, it is essential to gain insights into the current status of how these environmental and climate conditions are integrated into property values at a national scale.

Previous studies have used the hedonic property model as an approach in estimating the economic value of the environment. In the US, results obtained from hedonic analysis have provided a basis for establishing and evaluating progress in various environmental standards. The current body of research on capitalization has predominantly focused on specific regions or environmental quality, resulting in mixed outcomes. Additionally, a methodological challenge exists in accounting for confounding variables, an important aspect for obtaining accurate estimates of capitalization.

To study how different environmental qualities are capitalized in housing prices, housing variables, environmental variables, and satellite image features are aggregated for each 1km image grid across the continental US. The data cleaning, preprocessing steps and explanatory analyses are catalogued. These analyses outline the associations between individual housing attributes and environmental variables, as well as explore the extent to which satellite images explain variations in each of the variables.

Chapter 4

Estimating Economic Value of Environment

Chapter 2 introduced an approach to integrate satellite image data into a regression framework, while Chapter 3 discussed motivation and history of research in estimating the economic valuation of the environment. This included an endeavor to obtain nationwide estimates for various environmental qualities, which motivated the collection and aggregation of housing, environmental, and satellite image data for the continental US. The current chapter applies the method introduced in Chapter 2 to study the empirical question posed in Chapter 3. The study estimates the extent to which environmental qualities are capitalized in housing markets, using a hedonic model that includes controls for satellite images.

Section 4.1 outlines the model used to estimate capitalization, and Section 4.2 provides visualizations and interpretations of the estimated capitalization. Section 4.4 discusses the effect of adding image features on the results and provides guidance for future research.

4.1 Specification

Model Specification

We begin by introducing a simple model in which we assume that environmental effects are homogeneous across the US housing market and that variables affect housing value linearly. The model is restrictive, but it provides a baseline for more flexible models.

$$log(Y_i) = \beta_0 + T_i\beta_T + \mathbf{Z}_i\beta_Z + \varepsilon_i, \tag{4.1}$$

where Y_i is the median housing price in location i (1km by 1km region), T_i is the environmental variable of interest, $\mathbf{Z}_i = (Z_{i1}, ..., Z_{ip})$ is the vector of p observed covariates, which include the set of controls, such as housing attributes and environmental qualities, which are associated with either the outcome or the environmental variable of interest.

Equation (4.1) consists of several challenges, and the following outlines corresponding approaches to alleviate these problems.

- There is a need to select a functional form $f(\cdot)$ for how the variables enter into T_i and $\mathbf{Z}_i$. In particular, variables such as temperature and precipitation have been studied to have nonlinear associations with housing prices. To allow for flexible functions, natural splines are used to model $f(\cdot)$ per

$$f(x) = \sum_{j=1}^{m} \beta_j g_j(x), \tag{4.2}$$

  where g_j are the set of basis functions that span the space of kth order splines with knots at $s_1, ..., s_m$ and β_j are the associated spline coefficients. Natural splines are used as they impose the piecewise polynomial function to have a lower degrees to the leftmost and rightmost knots, reducing the issue of estimating functions with high variance at the boundaries of the domain. The current study uses 3 degrees of freedom.

- The equation (4.1) assumes homogeneity of the relationship between housing prices and covariates. However, a more realistic estimation approach involves accounting for spatial heterogeneity; for instance, the extent to which temperature is capitalized in housing prices may vary by region. One approach to account for such heterogeneity involves including state fixed effects.

- Unbiased estimation of β_T requires that ε_i is uncorrelated with T_i and Z_i, which is violated if unmeasured confounders exist. Despite the inclusion of covariates and spatial fixed effects, the presence of unmeasured confounders may still be likely. To account for variations observable from space, such as the physical characteristics of the land, satellite image features $\mathbf{R}$ are incorporated as a potential proxy for confounders. As the orthogonality assumption cannot be tested, sensitivity analysis is conducted to study the variability of estimated results across alternative specifications.

- The aforementioned modifications result in an increase in the feature set, including natural splines with degrees of freedom set to 3 for 17 variables, 47 spatial fixed effects, and 100 image features. Spatial fixed effects and image features serve as controls and the main interest is in interpreting coefficient estimates pertaining to variables under study. This invites the use deferentially-weighted ridge regression, which penalizes coefficients with different strength by putting weights to the loss function.

  When fitting ridge regression using `glmnet` in R, the default setting assigns equal weights of 1 to all coefficients. In this study, penalty factors of 0 are assigned to coefficients related to the 17 variables under study, ensuring that they are included in the model. On the other hand, penalty factors of 1 are assigned to coefficients associated with spatial fixed effects and image features, so that their coefficients are subject to shrinkage. These penalty factors are scaled to ensure their sum equals the number of regressors.

The specified model is

$$\begin{aligned} log(Y_i) &= f(T_i) + f(\mathbf{Z}_i) + S_i\beta_S + \mathbf{R}_i\beta_R + \varepsilon_i \\ &= \sum_{j=1}^{m} \beta_{t,j}\ g_j(t) + \sum_{k=1}^{p}\sum_{j=1}^{m} \beta_{z_k,j}\ g_j(z_k) + \mathbf{S}_i\,\beta_S + \mathbf{R}_i\beta_R + \varepsilon_i, \end{aligned} \tag{4.3}$$

where Y_i is the housing price, T_i is the environmental variable of interest, $\mathbf{Z}_i$ is the vector of covariates, $mathbf R_i$ is the vector of image features, and S_i is state fixed effects.

Capitalization

Coefficient estimates from the hedonic model specified in Equation (4.3) represent the marginal values of environmental qualities capitalized by market prices for homes. The capitalization of the environmental variable of interest can be computed as the product of the variable and its associated coefficient estimate:

$$\begin{aligned} log(\hat{Y}_i) &= \underbrace{T_i\hat{\beta}_T}_{\text{capitalization of } T} + \mathbf{Z}_i\,\hat{\beta}_Z + \mathbf{S}_i\,\hat{\beta}_S + \mathbf{R}_i\hat{\beta}_R \\ &= \hat{\beta}_0 + \underbrace{\sum_{j=1}^{m} \hat{\beta}_{t,j}\ g_j(t)}_{\text{capitalization of } T} + \sum_{k=1}^{p}\sum_{j=1}^{m} \hat{\beta}_{z_k,j}\ g_j(z_k) + \mathbf{S}_i\,\hat{\beta}_S + \mathbf{R}_i\hat{\beta}_R \end{aligned}$$

Since the outcome is log transformed, $T\hat{\beta}_T$ is converted so that it can be interpreted as capitalization in its original unit, USD. A natural way to do this is by exponentiating the value:

$$\hat{Y}_i = \underbrace{e^{T_i\hat{\beta}_T}}_{\text{capitalization of } T} e^{\mathbf{Z}_i\,\hat{\beta}_Z}\ldots$$

However, it is more interpretable to view housing values as a sum of implicit prices for each of its characteristics rather than as a product. To achieve this, instead of exponentiating, capitalization is multiplied by the median housing price. This approach allows the interpretation of capitalization in terms of USD while maintaining the model additive:

$$\tilde{Y}log(\hat{Y}_i) = \underbrace{\tilde{Y}T_i\hat{\beta}_T}_{\text{capitalization of } T} + \tilde{Y}\,\mathbf{Z}_i\,\hat{\beta}_Z + \cdots$$

For enhanced interpretability, relative capitalization is calculated, representing the change in capitalization of T relative to its reference T_r. The reference T_r is set as the median value of the variable T.

$$\text{Relative capitalization of } T \text{ in location } i = \tilde{Y}T_i\hat{\beta}_T - \tilde{Y}T_r\hat{\beta}_T$$

The relative capitalization reflects a thought experiment comparing two regions that are identical in all aspects, including observed covariates and satellite images, except for one characteristic T. For example, one region may have national median temperature of 11° and another with 21°. Relative capitalization of temperature indicates the economic value of the added 10°.

4.2 Estimated capitalization

Response functions

The response function in Figure 4.1 shows the percentage change in housing price relative to the reference, which is defined as the housing price of a location experiencing the median value of variables (Table 4.1). For example, in comparison to the housing price in a region with a national median temperature of 11°, there is a trend of increased housing prices, holding all other variables constant. On the other hand, houses in regions with lower temperatures generally tend to be priced lower. In the case of precipitation, regions facing deficits are estimated to have lower prices

compared to regions with an average of 60 mm of annual precipitation, typically observed in dry areas.

Other variables that indicate increased capitalization with higher values include dew point, PM 2.5, elevation, flood risk score, and surface water. The overall trend, where houses situated in areas with higher flood risk scores are estimated to be priced higher, while houses with minimal flood risk are priced lower, aligns with nationwide studies reported by [33] and [28]. These studies report the overvaluation of property values located in flood zones. Houses in areas with increased PM 2.5 levels, indicative of higher air pollution, are valued more highly.

Variables showing increased capitalization with lower values include sunshine and wind speed. Houses located in areas with high wind speed, often associated with storms, are priced much lower compared to areas with lower wind speed. As for fire potential index, greater potential for fire corresponds to lower-priced houses, but lower fire potential does not correspond to higher pricing. For tree cover, there seems to be an optimal amount of around 20% coverage, and houses located in areas with no or less trees are priced lower.

Environmental variable	Median value
Temperature	$10.6°C$
Precipitation	61.2 mm
Dew point	$1.66°C$
Sunshine	185 Watts/m^2
Fire potential index	31.2
PM 2.5	2.73 $\mu g/m^3$
Tree cover	3.15 %
Elevation	511 m
Flood risk score	1.86
Surface water	0 %
Wind speed	0.08 m/s

Table 4.1: **Median value of each environmental variable.** The median value is used as a reference for computing the relative value of capitalization.

Capitalization maps

While response functions show which variable values contribute to higher housing prices, capitalization maps in Figure 4.2 show the locations that experience relatively higher prices. For example, houses in warmer regions in the south and west coast tend to be priced higher relative to regions with a median temperature value of around $11°C$. In general, increased tree cover appears to be positively reflected in housing prices on the right side of the US, as well as along the west coast.

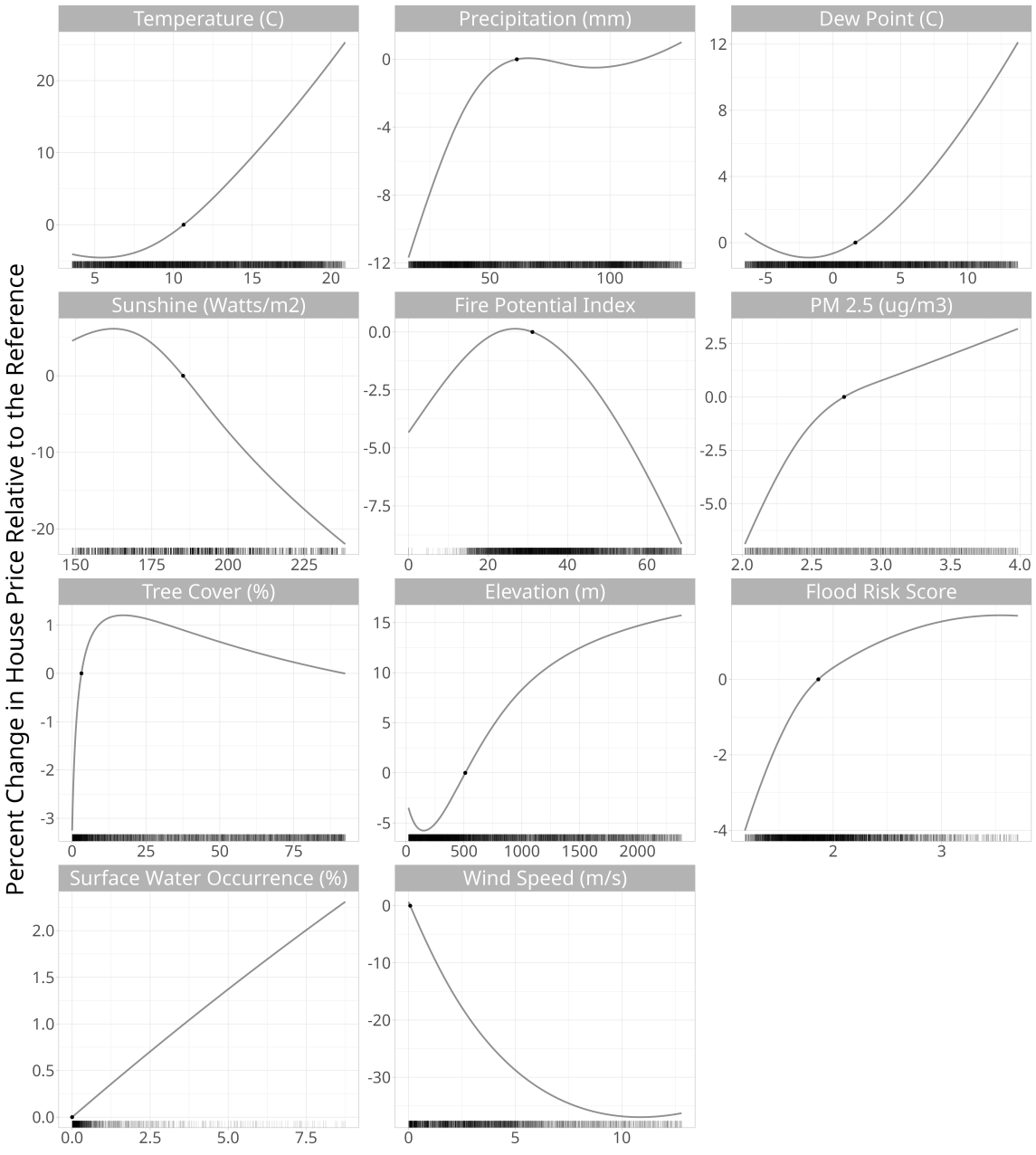

Figure 4.1: **Response functions.** The percentage change in housing price relative to its reference is plotted for each variable. The reference is defined as the housing price of a location experiencing the median value of a specific variable. The rug plot at the bottom of each figure shows the distribution of variable values, with 10,000 data points selected for visualization. As the data used in the study includes all 1km × 1km grids covering the continental US (i.e., it represents the population and not a sample), standard errors of the estimates are not plotted.

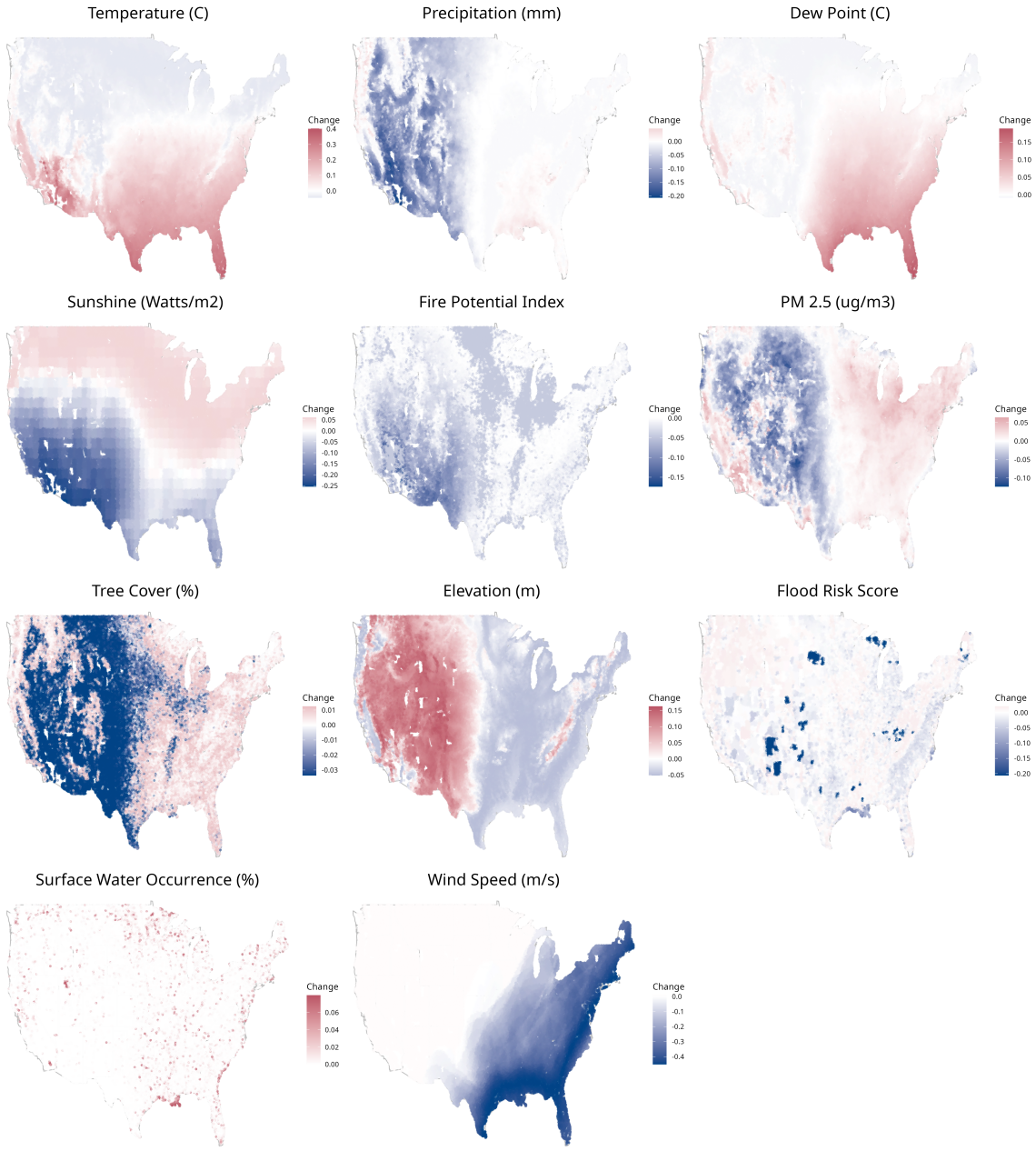

Figure 4.2: **The distribution of value capitalized by 11 environmental qualities in the United States.** The maps show the percent change in the relative capitalization value for each environmental quality. Areas with reference values are displayed in white, those with positive changes in red, and those with negative changes in blue. The maps show randomly sampled 100,000 units for display.

4.3 Explained variation in housing prices

Explained variation in housing prices

To quantify the extent to which environmental qualities as a collective factor accounts for housing prices, the study also examines the proportion of variation in housing prices that is explained by these qualities. This involves estimating housing prices using coefficients obtained from a regression model that incorporates all variables and image features. The 17 variables are categorized into three groups, which includes housing attributes, environmental qualities, and neighborhood characteristics (Table 4.2), and the estimation of housing price is obtained from each group.

Group	Variable
Housing attributes	House units
	Vacant houses
	Rooms
	Building age
Neighborhood	Income
	Population
Environmental	Temperature
	Precipitation
	Dew point
	Sunshine
	Fire potential index
	PM 2.5
	Tree cover
	Elevation
	Flood risk score
	Surface water
	Wind speed

Table 4.2: **Grouping of housing, neighborhood, and environmental variables.**

To maintain additivity within the model, the housing price is kept in its transformed state:

$$log(\hat{Y}_i) = \underbrace{T_i\hat{\beta}_T + X_1\hat{\beta}_{X_1} + \cdots + X_{10}\hat{\beta}_{X_{10}}}_{\mathbf{X}_{\text{environment}}\hat{\beta}_{\text{environment}}} + \underbrace{X_{11}\hat{\beta}_{X_{11}} + \cdots + X_{15}\hat{\beta}_{X_{15}}}_{\mathbf{X}_{\text{house}}\hat{\beta}_{\text{house}}} + \underbrace{X_{16}\hat{\beta}_{X_{16}} + X_{17}\hat{\beta}_{X_{17}}}_{\mathbf{X}_{\text{neighborhood}}\hat{\beta}_{\text{neighborhood}}} + \mathbf{S}_i\,\hat{\beta}_S + \mathbf{R}_i\hat{\beta}_R$$

The model, incorporating variables related to housing attributes, neighborhood characteristics, environmental qualities, state fixed effects, and satellite images, explains 63% of the variation in housing prices. Housing attributes account for 14% of the variation, neighborhood characteristics contribute to 18%, and environmental conditions contribute to 12% (Figure 4.3). The computation of explained variation in this manner offers a view into the relative contributions of each of the three groups of variables that are distinctly associated with housing prices.

4.4 Effects of controlling for images

In examining the impact of controlling for image information in this regression study, two aspects are considered. The first aspect involves comparing the regression results obtained with and without the inclusion of image information, while the second aspect explores how this effect varies depending on the amount of image information incorporated. For discussions of how image information reduces or amplifies bias from a theoretical standpoint, please refer to Chapter 2. The current section empirically studies the extent to which image information explains variation in housing prices.

Images and explained variation

To examine the extent to which including image information explains additional variation, R-squared values are compared between nested models: one without image features and one with image features R (Table 4.3). When environmental variables are included in the regression, the addition of image features does not greatly explain additional variation in housing prices. This outcome is expected due to the high correlation between some of the remotely measured environmental variables, such as tree cover and elevation, and image features (Figure 3.6). On the other hand, when only housing attributes are included in the model, the inclusion of image features makes a moderate contribution by explaining an additional 15% of the variation. A similar observation can be made for the model that solely incorporates neighborhood variables, as well as the model that incorporates both housing and neighborhood variables.

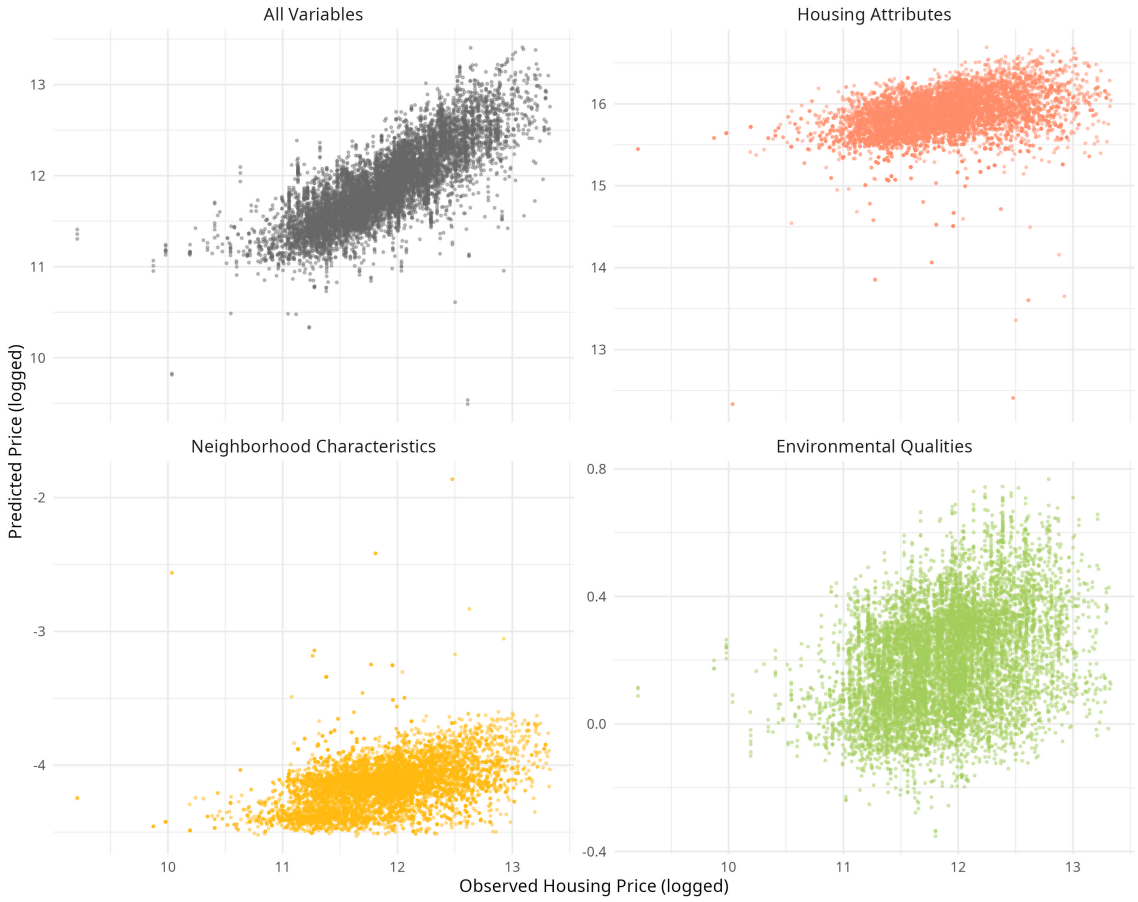

Figure 4.3: **Predicted housing price vs observed housing price.** Each figure shows a scatter plot of predicted housing prices based on different sets of variables.

Response functions in relation to the number of image features

In this analysis, we examine the sensitivity of the response functions to the number of features used (Figure 4.4). As the features in MOSAIKS are generated randomly, there is no theoretical basis for selecting a specific number of features. We fit a model with the specification outlined as Model 4.3, but vary the number of features among the values $\{0, 10, 100, 1000, 4000\}$. For each set of features and task, we perform

Model	R^2	Change in R^2
$Y \sim R$	0.16	
$Y \sim X_{\text{environment}}$	0.27	
$Y \sim X_{\text{environment}} + R$	0.3	+0.03
$Y \sim X_{\text{house}}$	0.3	
$Y \sim X_{\text{house}} + R$	0.45	+0.15
$Y \sim X_{\text{neighborhood}}$	0.23	
$Y \sim X_{\text{neighborhood}} + R$	0.4	+0.17
$Y \sim X_{\text{house}} + X_{\text{neighborhood}}$	0.4	
$Y \sim X_{\text{house}} + X_{\text{neighborhood}} + R$	0.52	+0.12

Table 4.3: **Proportion of variance explained by different sets of variables.** $X_{\text{environment}}$ denotes a set of eleven environmental variables, X_{house} includes five housing attributes, $X_{\text{neighborhood}}$ includes two neighborhood socioeconomic variables, R indicates a set of 100 image features. Change in R^2 is calculated as the difference between the R^2 of a smaller model which excludes R and a larger model that includes R.

3-fold cross-validation to determine the optimal hyperparameter λ.

There is generally a difference between the results obtained when image information is entirely excluded compared to when it is included to some extent. The greatest variation across feature sizes tends to occur in situations where observations are limited, such as at the lower and upper extremes.

In most of the variables, the response functions for features of sizes 10, 100, 1000, and 4000 show significant overlap, except for tree cover and surface water. This implies that the amount of image information does not appear to have a discernible effect on the interpretation of these variables, particularly in data domains where observations are abundant. This finding aligns with the sensitivity analysis conducted in Rolf et al [73], which compares the predictive performance of seven tasks using R^2 values across different numbers of features ranging from 100 to 8192. The analysis indicates that with features of size 100 captures over 80% portion of the variation explained by the full 8192 features for all seven variables under study.

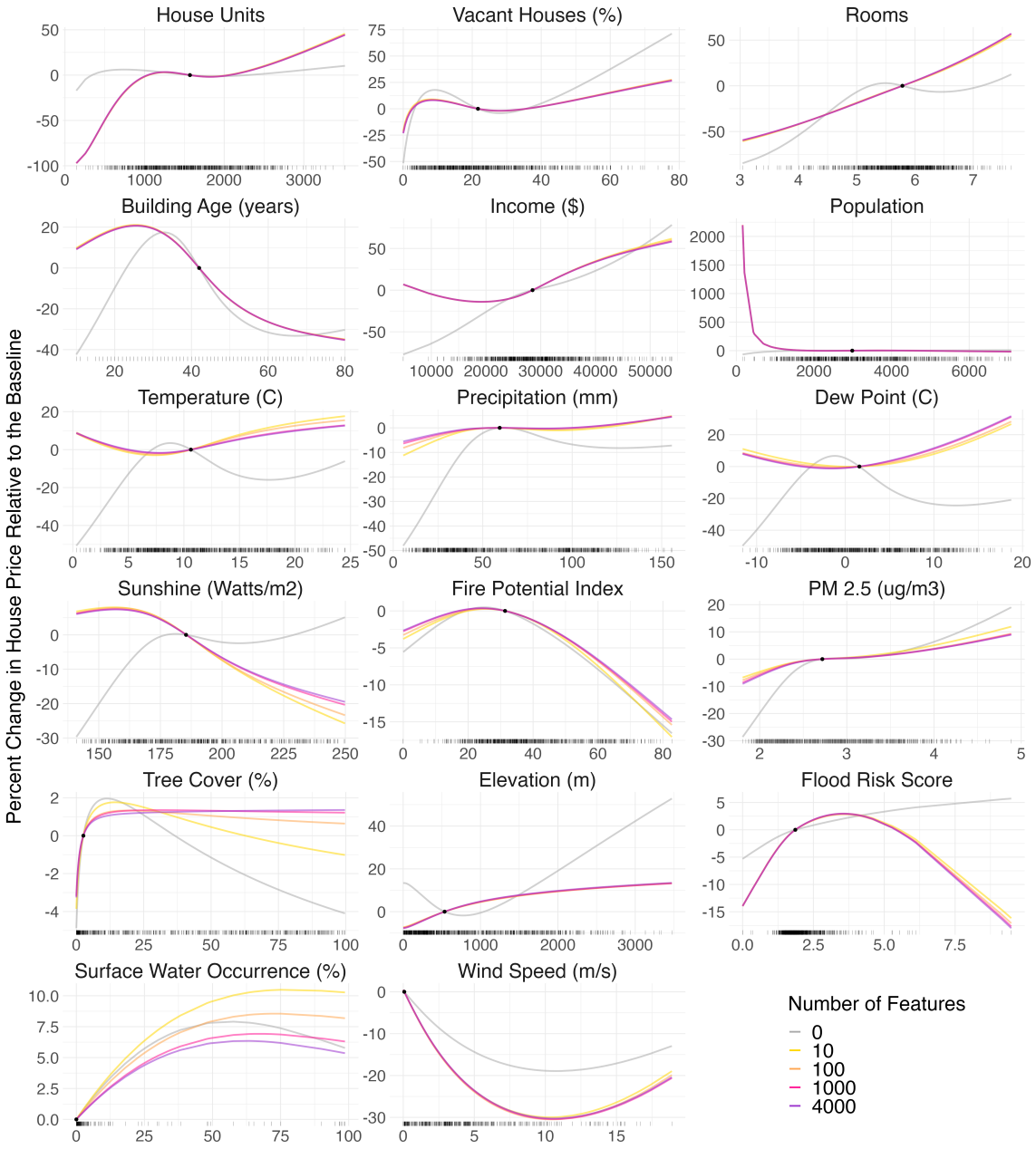

Figure 4.4: **Response functions across varying number of image features.** For each variable, five response functions are overlaid, each corresponding to a different number of features: 0, 10, 100, 1000, and 4000, while keeping the number of observations fixed at 400,000.

Limitations and paths forward

One of the limitations of our study is the presence of measurement error, specifically regarding the use of imagery. Since imagery is time-specific, incorporating imagery from different time points may result in the capture of different information. Additional research is needed to investigate the variability of imagery across various time points and assess the sensitivity of our findings to the use of imagery from different temporal contexts.

Secondly, the capitalization of risk is one of several approaches to communicate the risks involved to home buyers. The discussion surrounding risk capitalization should be contextualized within the intricate dynamics of housing finance systems, which can encourage development in unsuitable areas. In many regions across the country, low-income individuals live in areas where housing is relatively affordable. Merely reducing prices in risk-prone areas can inadvertently shift the burden of risk onto low-income individuals. What is needed is a coordinated effort to achieve a long-term reduction in risk that does not transfer the risk from one group to another.

4.5 Conclusion

The current chapter provides estimates of the extent to which various environmental attributes contribute to the valuation of housing prices in the continental US. Nationwide estimates offer a comprehensive perspective on the current state of how climate and environmental qualities are reflected in residential property values. Housing prices are studied, as they represent one of the few contexts where environmental quality is capitalized and also play a role in shaping the incentive structure concerning the scale and geographical placement of residential development.

The estimation uses a hedonic model, where housing prices are regressed on environmental qualities, additional covariates, and satellite image features. The inclusion of satellite image information in the analysis is considered a potential method for controlling latent confounding variables. One of the finding suggests that certain risk factor variables, including flood risk score and PM 2.5, are associated with higher housing prices. Contrary to the expectation that risks would lead to property devaluation, these results suggest an overvaluation of housing prices.

Part III

Characterizing Spatio-Temporal Trends of Extremes

Chapter 5

Characterizing spatio-temporal trends of extreme precipitation

5.1 Introduction

Characterizing variability and changes in precipitation, including extreme precipitation, is important for understanding and monitoring natural hazards. Numerous studies report increased variability in extreme precipitation events in the US [29] [89] [21] [82].

A widely used approach for characterizing variability and changes in extreme events involves pre-specifying an exceedance threshold and defining an anomalous event by calculating the frequency of threshold exceedance and aggregating across space [43] [38] [21] [82]. As an example, the Bukovsky region is one subregionalization that divides the US into 29 groupings with similar temperature and precipitation characteristics [10]. However, determining the appropriate threshold value, time window, and spatial boundary for defining anomalies is not a trivial task. One challenge arises from the fact that defining what is considered extreme and catastrophic depends on the specific area under consideration and the timeframe involved. What may be deemed extreme in one region may not hold the same characterization in another location.

In the study of extreme precipitation events, it is also common to use gridded data products, which uses smoothing techniques to produce spatially complete data. However, the process of smoothing tends to reduce the magnitude of extreme values. There is a thread of research reporting issues with gridded data products in accurately characterizing climatological extremes [26] [23] [72]. In contrast, in situ measurements from weather stations offer important native-scale data for extremes.

Therefore, using in situ measurements and using a method that allows for the characterization of precipitation with minimal need for prior specification of anomaly criteria, such as regional boundaries or fixed temporal windows, would be beneficial. In this study, functional principal component analysis (FPCA) is used to characterize both seasonal mean and extreme precipitation using data from the Global Historical Climatology Network Daily over the contiguous United States. FPCA provides a flexible method for identifying modes of temporal variability and spatial patterns of precipitation variability across various scales. Additionally, the method characterizes nonlinear trends in the distribution of precipitation and help detect anomalous spatio-temporal events.

5.2 Precipitation data

We use in-situ measurements of daily precipitation (mm) from 1880 to 2017, collected by 21,232 stations across continental United States. Data is collected by Global Historical Climatology Network (GHCN) [51]. For each year, seasonal mean and seasonal extreme precipitation are computed. Seasons are defined as Spring (March, April, May), Summer (June, July, August), Fall (September, October, November), and Winter (December, January, February). Seasonal mean is defined as the total precipitation in a given season, which is computed as the product of seasonal mean and the number of days in the season. Seasonal extreme is defined as the maximum precipitation in a given season.

5.3 Methods

Frequency decomposition

Spline smoothing is a smoothing technique that estimates the underlying function of a data by performing a regularized regression over the natural spline basis, placing knots at all points $x_1, ..., x_n$, where n is the number of knots which we pre-specify. As a process, it 1) breaks up the domain by n knots, 2) places a special type of piecewise polynomial function called *natural spline*, and 3) estimates a coefficient of each polynomial function such that it minimizes the sum of the squared differences and a penalty.

A kth order *spline* is a piecewise polynomial function of degree k that is continuous and has continuous derivatives of orders $1, ..., k-1$, at its knot points. Natural splines are piecewise polynomial functions that have a lower degree on the leftmost

and rightmost intervals. This is to remedy the problem that is often confronted in regression smoothing, where the boundaries of the domain get estimates with high variance (which gets worse as the order k gets larger).

For a cubic smoothing splines where $k = 3$, the coefficients β are chosen to minimize the following objective function

$$||y - G\beta||_2^2 + \lambda\beta^T\Omega\beta,$$

where $G \in \mathbb{R}^{n\times n}$ is the basis matrix defined as

$$G_{ij} = g_j(x_i), \quad i, j = 1, ..., n,$$

and $\Omega \in \mathbb{R}^{n\times n}$ is the penalty matrix defined as

$$\Omega_{ij} = \int g_i''(t)g_j''(t)dt, \quad i, j = 1, ..., n.$$

The *smoothing spline* estimate at x is defined as

$$\hat{r}(x) = \sum_{j=1}^{n} \hat{\beta}_j g_j(x).$$

Smoothing splines have a regularization term that determines how much to penalize the second differential of the basis function g_j and imparts more shrinkage on the coefficients $\hat{\beta}_j$ that correspond to wigglier functions g_j. The parameter $\lambda \geq 0$ is the smoothing parameter. If $\lambda \to 0$, there is no shrinkage or smoothing. If $\lambda \to \infty$, there is more shrinkage and the estimate converges to a linear least squares estimate.

Functional principal component analysis

A goal of functional principal component analysis (FPCA) is to find the dominant modes of variation in the data (e.g. sinusoidal nature). We also want to know how many of these modes of variations are required to adequately approximate the original data. PCA for functional data works in a similar way as PCA for multivariate data. [66]

1. Find *principal component weight* function $\xi_1(s)$ for which the *principal components scores*

   $$f_{i1} = \int \xi_1(s)x_i(s)ds$$

   maximize $\sum_i f_{i1}^2$ subject to

   $$\int \xi_1^2(s)ds = ||\xi_1||^2 = 1.$$

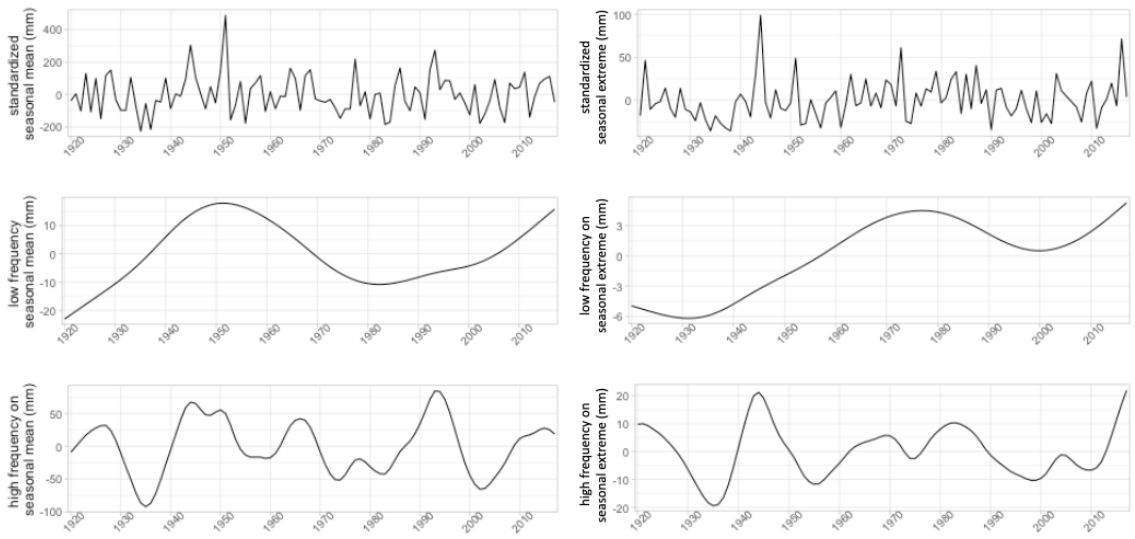

Figure 5.1: **Time series of selected station in Kansas for seasonal mean (left) and seasonal extreme (right) for summer.** The upper plot displays standardized seasonal precipitation measurements (mm). In the middle plot, a low pass filter is applied through spline smoothing to the standardized measurements, using a smoothing parameter of 0.9 to capture a long-term trend. Subsequently, a second smoothing spline is applied with a smoothing parameter of 0.5 to the residuals to obtain the medium pass filter. The residuals represent the difference between the standardized measurement and the low pass filter. The bottom plot shows the high pass filter, obtained by applying spline smoothing to the residuals, where the residual is the difference between the medium pass filter and the low pass filter. A smoothing parameter of 0.5 is used to capture shorter-term trends.

2. Compute the weight function $\xi_2(s)$ and principal component scores f_{i2} that maximize $\sum_i f_{im}^2$ subject to the normalizing constraint $||\xi_2||^2 = 1$ and the orthogonal constraint:
$$\int \xi_2(s)\xi_1(s)ds = 0.$$

3. and so on.

In essence, we wish to find principal components, which are linear combination of functions that maximally explain variation in data. Each component should be

orthogonal to each other, meaning that they should explain variation that is not explained by other components.

To obtain simple and interpretable components, we rotate components. Let B be a $K \times n$ matrix consisting of the first K principal component functions $\xi_1, ..., \xi_K$. At row m, B has the values $\xi_m(t_1), ..., \xi_m(t_n)$. We can define a set of orthonormal components A as

$$A = TB,$$

where T is an orthonormal matrix of order K (i.e. $T'T = I$). T is chosen such that it maximizes the variation in the values a_{mj}^2, which occurs when each component have a few high loadings and the rest being zero or close to zero. Note that, because T is a rotation matrix, it redistributes the variance explained each of the components, and thus the overall variance explained by the K components remain the same.

The decisions made in this study include the following:

- Division of time into four seasons
- Decomposition of functional data into three different frequencies
- Determination of the number of principal components

These choices result in a total of 240 principal component (PC) outcomes: 4 seasons $\times$ 10 principal components $\times$ 2 decompositions $\times$ 3 frequencies. While some pre-specifications were made regarding the temporal window and the number of frequencies, no spatial boundary was specified.

5.4 Case study of Dust Bowl

A case study is conducted as an initial procedure to examine the capability of FPCA in detecting anomalous events. Specifically, the case study focuses on the Dust Bowl, a historical case of severe drought in the central US during the 1930s [48]. The period was marked by a decade of rainfall deficits and increased temperatures affecting nearly two-thirds of the country, particularly the central and southern Great Plains [74] [76]. The Dust Bowl is characterized by its atypical spatial pattern [34] [17], prolonged duration, and intensity, with rainfall deficit exceeding 0.1 mm/day for much of the central US, with peak deficits over 0.3 mm/day in Kansas [75]. Accompanying the precipitation deficits were major wind erosion and dust storms [48].

The FPCA results are shown for high-frequency time series of seasonal mean and extreme values in both summer and winter (Figure 5.2). Among the 10 principal components (PC), the outcomes for specific components capturing anomalies in the 1930s are selected. The PC function plot represents the overall mean time series and two curves derived by adding and subtracting a multiple of each PC curve. Representing components as perturbations of the mean makes it possible to identify the time windows in which a component explains high variability. The perturbations of the mean is constructed by the mean function plus/minus a constant factor times each eigenfunction separately:

$$\hat{\mu} \pm 0.2C\hat{\gamma}_j,$$

where $\hat{\mu}$ is the overall mean precipitation, C is a constant, and $\hat{\gamma}_j$ is the time series of PC's coefficients. The constant C is defined as the root-mean-square difference between $\hat{\mu}$ and its overall time average $\bar{\mu}$,

$$C^2 = \frac{1}{T} \|\hat{\mu} - \bar{\mu}\|^2,$$

where

$$\bar{\mu} = \frac{1}{T} \int \hat{\mu}(t)\, dt.$$

For example, considering the seasonal mean during summer, the first PC function explains about 10% of the variability in the data, with a relatively large variability in the 1930s, as indicated by the high amplitude in the perturbed functions.

The map shows the spatial distribution of individual weather stations along with their respective PC scores. The PC score indicates the strength of association between the time series observed at a particular station and the component under consideration. Weather stations in the central Great Plains exhibit large negative PC scores, which aligns with the region that experienced a severe precipitation deficit during the Dust Bowl. The combined use of PC functions and maps provides information about the temporal window and spatial distribution which experienced a anomalous case of precipitation deficit.

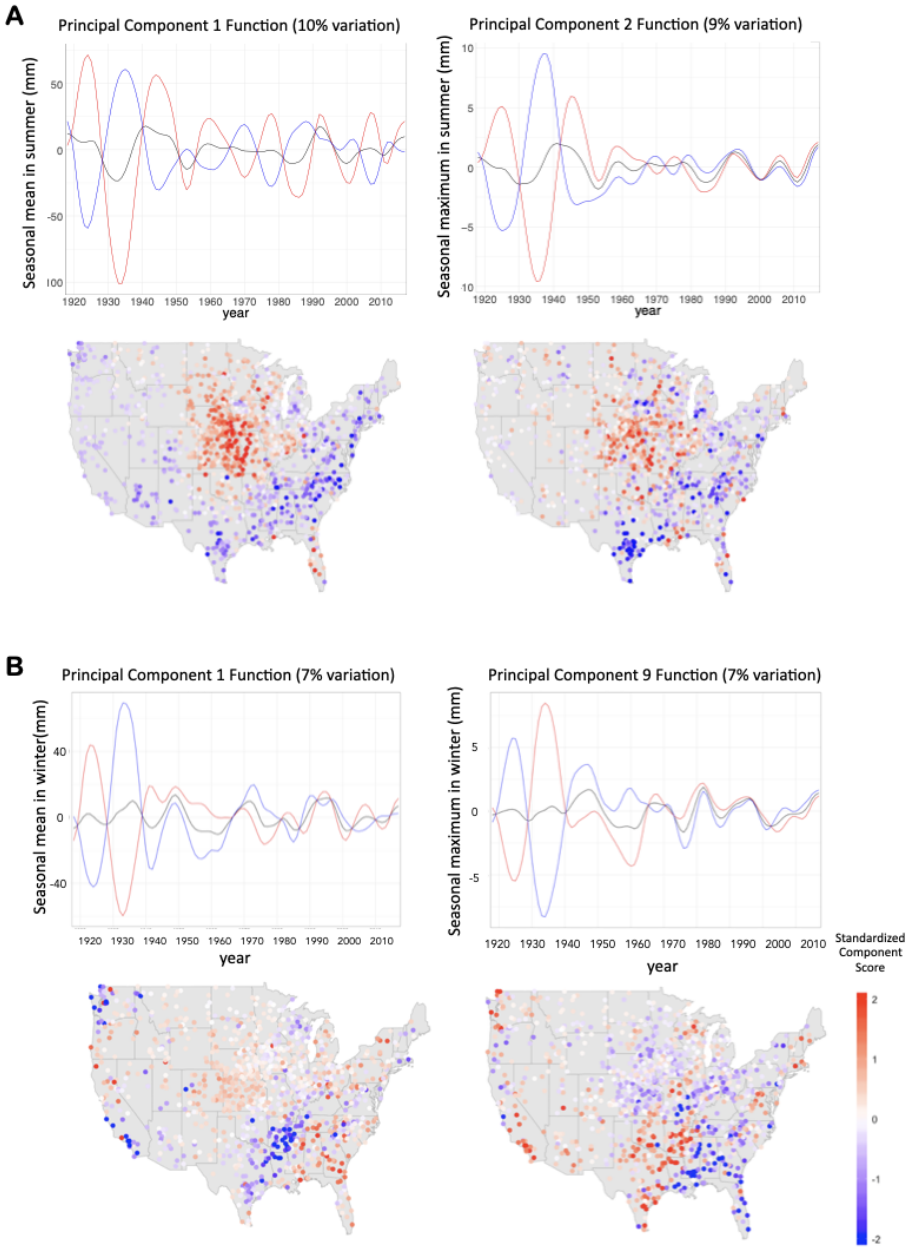

Figure 5.2: **PC functions and spatial distributions of PC scores for seasonal mean and extreme precipitation.** Time series plots: seasonal mean or extreme precipitation (mm) vs. year. For each scenario, one principal component result is selected for display. The plot features the mean precipitation (in black) and perturbations of the mean, obtained by adding and subtracting a suitable multiple of the eigenfunction (in red and blue). Maps: principal component scores of each weather station. (A) displays the PC functions and maps for the summer, while (B) displays the same for winter.

5.5 Conclusion

The study used functional principal component analysis to characterize extreme precipitation in the US over the last century. To account for the fact that definitions of extreme events differ across regions and time, the method involves a relatively small number of decisions. This includes determining how to divide seasons and the general frequency of precipitation measurements: short-term, medium-term, and long-term trends. Importantly, the analysis does not require specifications for dividing the contiguous US or determining the temporal window. The PC function, along with maps of PC scores, can be used to identify the temporal window and the geographic location of anomalous events.

While the study conducts a case study as an initial step to assess the effectiveness of the FPCA method in detecting extremes, a more comprehensive validation analysis is required. Specifically, summarizing instances where FPCA results align with observed anomalous events in raw data, and cases where FPCA incorrectly indicates anomalies not present in raw data, would be beneficial. Additionally, there is a need to evaluate the variability in results across different parameter specifications.

Effectively summarizing spatio-temporal events is challenging, particularly when the area and time under consideration are extensive. Extreme events introduce an additional layer of complexity as the definitions of extremes are context-dependent, varying with time and space, and requiring in situ measurements. The current study introduces a potential approach for conducting an exploratory analysis of spatio-temporal data.

Appendix A

Appendix

Term	Estimate
House Units 1	8.33
House Units 2	26.29
House Units 3	6.35
Population 1	-1.96
Population 2	-5.56
Population 3	-1.46
Building Age 1	-0.33
Building Age 2	0.48
Building Age 3	-0.46
Vacant Houses 1	-0.082
Vacant Houses 2	0.40
Vacant Houses 3	0.13
Rooms 1	0.85
Rooms 2	1.37
Rooms 3	0.13
Income 1	0.27
Income 2	-0.51
Income 3	0.46
Temperature 1	0.12
Temperature 2	0.11
Temperature 3	0.38
Precipitation 1	0.09
Precipitation 2	0.38
Precipitation 3	0.11
Dew Point 1	0.0009
Dew Point 2	-0.04
Dew Point 3	0.16
Sunshine 1	-0.16

Sunshine 2	-0.16
Sunshine 3	-0.35
Fire Potential Index 1	0.01
Fire Potential Index 2	-0.03
Fire Potential Index 3	-0.15
PM 2.5 1	0.09
PM 2.5 2	0.25
PM 2.5 3	0.12
Tree Cover 1	0.04
Tree Cover 2	0.06
Tree Cover 3	0.01
Elevation 1	0.09
Elevation 2	-0.18
Elevation 3	0.19
Wind Speed 1	-0.59
Wind Speed 2	-0.04
Wind Speed 3	-0.10
Surface Water Occurrence 1	0.07
Surface Water Occurrence 2	-0.01
Surface Water Occurrence 3	0.05
Flood Risk Score 1	0.16
Flood Risk Score 2	0.30
Flood Risk Score 3	-0.02

Table A.1: Spline coefficient estimates were obtained for housing and environmental variables. Natural splines with degrees of freedom set to 3 were used to model each variable.

Bibliography

[1] Joshua K Abbott and H Allen Klaiber. "An embarrassment of riches: confronting omitted variable bias and multiscale capitalization in hedonic price models". In: *The Review of Economics and Statistics* 93.4 (Nov. 2011), pp. 1331–1342. URL: http://direct.mit.edu/rest/article-pdf/93/4/1331/1615154/rest_a_00134.pdf.

[2] Emily Aiken, Esther Rolf, and Joshua Blumenstock. *Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy*. Tech. rep.

[3] Amazon Web Services. *First Street Foundation*. 2022.

[4] Amazon Web Services. *Terrain Tiles*. 2018.

[5] Atmospheric Composition Analysis Group. *Satellite-derived PM2.5*. 2021.

[6] Allan Beltrán, David Maddison, and Robert J.R. Elliott. "Is Flood Risk Capitalised Into Property Values?" In: *Ecological Economics* 146 (Apr. 2018), pp. 668–685. ISSN: 09218009. DOI: 10.1016/j.ecolecon.2017.12.015.

[7] Ariel BenYishay et al. "Indigenous land rights and deforestation: Evidence from the Brazilian Amazon". In: *Journal of Environmental Economics and Management* 86 (Nov. 2017), pp. 29–47. ISSN: 10960449. DOI: 10.1016/j.jeem.2017.07.008.

[8] Jay Bhattacharya, William B Vogt, and Senior Economist. *Do instrumental variables belong in propensity scores?* Tech. rep. National Bureau of Economic Research, 2007. URL: http://www.nber.org/papers/t0343.

[9] Allen Blackman. *Evaluating forest conservation policies in developing countries using remote sensing data: An introduction and practical guide*. Sept. 2013. DOI: 10.1016/j.forpol.2013.04.006.

[10] M.S. Bukovsky. *Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective*. Tech. rep. Boulder, CO: Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, 2011.

[11] Marshall Burke et al. *Using satellite imagery to understand and promote sustainable development*. Mar. 2021. DOI: 10.1126/science.abe8628.

[12] Tamma Carleton et al. "Multi-Task Observation Using Satellite Imagery and Kitchen Sinks (MOSAIKS) API". In: *https://mosaiks.org* (2022).

[13] Patricia Ann Champ, Geoffrey H. Donovan, and Christopher M. Barth. "Homebuyers and wildfire risk: A colorado springs case study". In: *Society and Natural Resources* 23.1 (Jan. 2010), pp. 58–70. ISSN: 08941920. DOI: 10.1080/08941920802179766.

[14] Guanghua Chi et al. "Microestimates of wealth for all low-and middle-income countries". In: (2023). DOI: 10.1073/pnas.2113658119/-/DCSupplemental.

[15] T L Chin and K W Chau. "A critical review of literature on the hedonic price model". In: *International Journal for Housing and Its Applications* 27.2 (2003), pp. 145–165.

[16] Carlos Cinelli, Andrew Forney, and Judea Pearl. *A Crash Course in Good and Bad Controls*. Tech. rep. 2020. URL: https://ucla.in/2ZcRpRq,.

[17] Benjamin I. Cook, Richard Seager, and Jason E. Smerdon. "The worst North American drought year of the last millennium: 1934". In: *Geophysical Research Letters* 41.20 (Oct. 2014), pp. 7298–7305. ISSN: 19448007. DOI: 10.1002/2014GL061661.

[18] Christopher Daly et al. "Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States". In: *International Journal of Climatology* 28.15 (2008), pp. 2031–2064. ISSN: 10970088. DOI: 10.1002/joc.1688.

[19] Peng Ding and Luke W. Miratrix. "To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias". In: *Journal of Causal Inference* 3.1 (Sept. 2014), pp. 41–57. ISSN: 2193-3677. DOI: 10.1515/jci-2013-0021.

[20] Hannah Druckenmiller and Solomon Hsiang. "Accounting for unobservable heterogeneity in cross section using spatial first differences". 2018. URL: http://www.nber.org/papers/w25177.

[21] D.R. Easterling et al. "Precipitation change in the United States". In: *Climate Science Special Report: Fourth National Climate Assessment* I (2017), pp. 207–230. DOI: 10.7930/J0H993CC. URL: https://science2017.globalchange.gov/chapter/7/.

[22] European Commission. *Global Surface Water - Data Access*. 2022.

[23] Muhammad Abrar Faiz et al. "How accurate are the performances of gridded precipitation data products over Northeast China?" In: *Atmospheric Research* 211 (Oct. 2018), pp. 12–20. ISSN: 01698095. DOI: 10.1016/j.atmosres.2018.05.006.

[24] Andrew D Foster and Mark R Rosenzweig. "Economic growth and the rise of forests". In: *The Quarterly Journal of Economics* 118.2 (May 2003), pp. 601–637. URL: https://academic.oup.com/qje/article/118/2/601/1899592.

[25] Ragnar Frisch and F.V. Waugh. "Partial time regression as compared with individual trends". In: *Econometrica* 1 (Oct. 1933), pp. 387–401.

[26] David Gampe and Ralf Ludwig. "Evaluation of Gridded Precipitation Data Products for Hydrological Applications in Complex Topography". In: *Hydrology* 4.53 (2017).

[27] Melissa Gervais et al. "Representing extremes in a daily gridded precipitation analysis over the United States: Impacts of station density, resolution, and gridding methods". In: *Journal of Climate* 27.14 (2014), pp. 5201–5218. ISSN: 08948755. DOI: 10.1175/JCLI-D-13-00319.1.

[28] Jesse D. Gourevitch et al. "Unpriced climate risk and the potential consequences of overvaluation in US housing markets". In: *Nature Climate Change* (Feb. 2023). ISSN: 1758-678X. DOI: 10.1038/s41558-023-01594-8. URL: https://www.nature.com/articles/s41558-023-01594-8.

[29] Pavel Ya Groisman et al. "Trends in Intense Precipitation in the Climate Record". In: *Journal of Climate* 18 (May 2005), pp. 1326–1350.

[30] Melanie S. Hammer et al. "Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998-2018)". In: *Environmental Science and Technology* 54.13 (July 2020), pp. 7879–7890. ISSN: 15205851. DOI: 10.1021/acs.est.0c01764.

[31] M C Hansen et al. "High-resolution global maps of 21st-century forest cover change." In: *Science (New York, N.Y.)* 342.6160 (Nov. 2013), pp. 850–3. ISSN: 1095-9203. DOI: 10.1126/science.1244693. URL: http://www.ncbi.nlm.nih.gov/pubmed/24233722.

[32] Andrew Head et al. "Can Human Development be Measured with Satellite Imagery?" In: Association for Computing Machinery (ACM), Nov. 2017, pp. 1–11. DOI: 10.1145/3136560.3136576.

[33] Miyuki Hino and Marshall Burke. "The effect of information about climate risk on property values". In: (). DOI: 10.1073/pnas.2003374118/-/DCSupplemental.y. URL: https://www.pnas.org/lookup/suppl/.

[34] Martin Hoerling, Xiao Wei Quan, and Jon Eischeidi. "Distinct causes for two principal U.S. droughts of the 20th century". In: *Geophysical Research Letters* 36.19 (Oct. 2009). ISSN: 00948276. DOI: 10.1029/2009GL039860.

[35] Solomon M. Hsiang. "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.35 (Aug. 2010), pp. 15367–15372. ISSN: 00278424. DOI: 10.1073/pnas.1009510107.

[36] Luna Yue Huang, Hsiang M. Hsiang, and Marco Gonzalez-Navarro. "Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-Poverty Programs". Cambridge, July 2021. URL: http://www.nber.org/papers/w29105.

[37] Meha Jain. "The benefits and pitfalls of using satellite data for causal inference". In: *Review of Environmental Economics and Policy* 14.1 (Jan. 2020), pp. 157–169. ISSN: 17506824. DOI: 10.1093/reep/rez023.

[38] Emily Janssen et al. "Observational- and model-based trends and projections of extreme precipitation over the contiguous United States". In: *Earth's Future* 2.2 (Feb. 2014), pp. 99–113. ISSN: 2328-4277. DOI: 10.1002/2013ef000185.

[39] Neal Jean et al. "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301 (2016), pp. 790–794. URL: https://www.science.org.

[40] Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. "Estimating Causal Effects Under Image Confounding Bias with an Application to Poverty in Africa". In: (June 2022). URL: http://arxiv.org/abs/2206.06410.

[41] Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. "Image-based Treatment Effect Heterogeneity". In: (June 2022). URL: http://arxiv.org/abs/2206.06417.

[42] Carolyn Kousky et al. "Flood Risk and the U.S. Housing Market". In: *Journal of Housing Research* 29.sup1 (Dec. 2020), S3–S24. ISSN: 1052-7001. DOI: 10.1080/10527001.2020.1836915.

[43] Kenneth E. Kunkel et al. "Monitoring and understanding trends in extreme storms: State of knowledge". In: *Bulletin of the American Meteorological Society* 94.4 (Apr. 2013), pp. 499–514. ISSN: 00030007. DOI: 10.1175/BAMS-D-11-00262.1.

[44] Mark G. Lawrence. "The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications". In: *Bulletin of the American Meteorological Society* 86.2 (Feb. 2005), pp. 225–233. ISSN: 00030007. DOI: 10.1175/BAMS-86-2-225.

[45] John Loomis. "Do nearby forest fires cause a reduction in residential property values?" In: *Journal of Forest Economics* 10.3 (Nov. 2004), pp. 149–157. ISSN: 11046899. DOI: 10.1016/j.jfe.2004.08.001.

[46] M.C. Lovell. "Seasonal adjustment of economic time series and multiple regression analysis". In: *Journal of the American Statistical Association* 58 (Dec. 1963), pp. 993–1010.

[47] Javed Mallick. "Land Characterization Analysis of Surface Temperature of Semi-Arid Mountainous City Abha, Saudi Arabia Using Remote Sensing and GIS". In: *Journal of Geographic Information System* 06.06 (2014), pp. 664–676. ISSN: 2151-1950. DOI: 10.4236/jgis.2014.66055.

[48] W.A. Mattice. "Dust storms, November 1933 to May 1934". In: *Monthly Weather Rev* 63 (1935), pp. 53–55.

[49] Shawn J. McCoy and Randall P. Walsh. "Wildfire risk, salience & housing demand". In: *Journal of Environmental Economics and Management* 91 (Sept. 2018), pp. 203–228. ISSN: 10960449. DOI: 10.1016/j.jeem.2018.07.005.

[50] Robert Mendelsohn and Sheila Olmstead. "The economic valuation of environmental amenities and disamenities: Methods and applications". In: *Annual Review of Environment and Resources* 34 (2009), pp. 325–347. ISSN: 15435938. DOI: 10.1146/annurev-environ-011509-135201.

[51] Matthew J. Menne et al. "An overview of the global historical climatology network-daily database". In: *Journal of Atmospheric and Oceanic Technology* 29.7 (July 2012), pp. 897–910. ISSN: 07390572. DOI: 10.1175/JTECH-D-11-00103.1.

[52] Joel A. Middleton et al. "Bias amplification and bias unmasking". In: *Political Analysis* 24.3 (2016), pp. 307–323. ISSN: 14764989. DOI: 10.1093/pan/mpw015.

[53] NASA. *CERES_SYN1deg_Ed4.1 Subsetting and Browsing*. 2021.

[54] Raymond B Palmquist and V Kerry Smith. "The Use of Hedonic Property Value Techniques for Policy and Litigation". In: *The International Yearbook of Environmental and Resource Economics*. Ed. by Tom Tietenberg and Henk Folmer. Edward Elgar Publishing, 2002. Chap. 3, pp. 115–164.

[55] Raymond B. Palmquist. *Chapter 16 Property Value Models*. 2005. DOI: 10.1016/S1574-0099(05)02016-4.

[56] Georgia Papadogeorgou, Christine Choirat, and Corwin M. Zigler. "Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching". In: *Biostatistics* 20.2 (Apr. 2019), pp. 256–272. ISSN: 14684357. DOI: 10.1093/biostatistics/kxx074.

[57] Judea Pearl. "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates". In: *Proceedings of UAI* (2010), pp. 417–424.

[58] Jean François Pekel et al. "High-resolution mapping of global surface water and its long-term changes". In: *Nature* 540.7633 (Dec. 2016), pp. 418–422. ISSN: 14764687. DOI: 10.1038/nature20584.

[59] Planet Team. *Planet Application Program Interface: In Space for Life on Earth*. 2017.

[60] Planet Team. *Planet imagery product specifications*. Tech. rep. Planet, June 2022. URL: https://assets.planet.com/products/basemap/planet-basemaps-product-specifications.pdf.

[61] Haiganoush K Preisler et al. "Forecasting Distribution of Numbers of Large Fires". In: *Proceedings of the large wildland fires conference* (2015), pp. 181–187. URL: http://www.wfas.net/index.php/.

[62] PRISM Climate Group. *PRISM Climate Data*. 2022.

[63] Jonathan Proctor, Tamma Carleton, and Sandy Sum. "Parameter recovery using remotely sensed variables". 2023. URL: http://www.nber.org/papers/w30861.

[64] Volker C. Radeloff et al. "Rapid growth of the US wildland-urban interface raises wildfire risk". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.13 (Mar. 2018), pp. 3314–3319. ISSN: 10916490. DOI: 10.1073/pnas.1718850115.

[65] Ali Rahimi and Benjamin Recht. "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems 21*. 2008.

[66] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. New York, NY, USA: Springer Series in Statistics. Springer-Verlag New York Inc., 2005.

[67] Krishna Rao et al. "Plant-water sensitivity regulates wildfire vulnerability". In: *Nature Ecology and Evolution* 6.3 (Mar. 2022), pp. 332–339. ISSN: 2397334X. DOI: 10.1038/s41559-021-01654-2.

[68] Nathan Ratledge et al. "Using machine learning to assess the livelihood impact of electricity access". In: *Nature* 611.7936 (Nov. 2022), pp. 491–495. ISSN: 14764687. DOI: 10.1038/s41586-022-05322-8.

[69] Abbavaram Gowtham Reddy and Benin Godfrey. *CANDLE: An Image Dataset for Causal Analysis in Disentangled Representations*. Tech. rep.

[70] Ronald G Ridker and John A Henning. "The Determinants of Residential Property Values with Special Reference to Air Pollution". In: *The Review of Economics and Statistics* 49.2 (May 1967), pp. 246–257. URL: https://about.jstor.org/terms.

[71] Mark D. Risser et al. "A probabilistic gridded product for daily precipitation extremes over the United States". In: *Climate Dynamics* 53.5-6 (Sept. 2019), pp. 2517–2538. ISSN: 14320894. DOI: 10.1007/s00382-019-04636-0.

[72] Mark D. Risser et al. "Detected changes in precipitation extremes at their native scales derived from in situ measurements". In: *Journal of Climate* 32.23 (Dec. 2019), pp. 8087–8109. ISSN: 08948755. DOI: 10.1175/JCLI-D-19-0077.1.

[73] Esther Rolf et al. "A generalizable and accessible approach to machine learning with global satellite imagery". In: *Nature Communications* 12.1 (Dec. 2021). ISSN: 20411723. DOI: 10.1038/s41467-021-24638-z.

[74] Cynthia Rosenzweig, Daniel Hillel, and D Lel. "Plant and Environment Interactions The Dust Bowl of the 1930s: Analog of Greenhouse Effect in the Great Plains?" In: *Journal of Environmental Quality* 22 (Jan. 1993), pp. 9–22.

[75] Siegfried D Schubert et al. "On the Cause of the 1930s Dust Bowl". In: *Science* 303.5665 (Mar. 2004), pp. 1855–1859. URL: http://science.sciencemag.org/.

[76] Richard Seager et al. "Modeling of Tropical Forcing of Persistent Droughts and Pluvials over Western North America: 1856-2000*". In: *Journal of Climate4065* 18 (Oct. 2005), pp. 4065–4088. URL: http://www.usbr.gov/uc/water/crsp/cs_gcd..

[77] Luke Sherman et al. "Global High-Resolution Estimates of the United Nations Human Development Index Using Satellite Imagery and Machine-learning". 2023. URL: http://www.nber.org/papers/w31044.

[78] Ian Shrier. "Letter to the editor: Propensity cores". In: *Statistics in Medicine* 28.8 (Apr. 2009), pp. 1315–1318. ISSN: 02776715. DOI: 10.1002/sim.3473.

[79] Isabella S Smythe and Joshua E Blumenstock. "Geographic microtargeting of social assistance with high-resolution poverty maps". In: *Proceeding of the National Academy of Sciences* 119.32 (Aug. 2022). DOI: 10.1073/pnas. URL: https://doi.org/10.1073/pnas.2120025119.

[80] Peter M. Steiner and Yongnam Kim. "The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases". In: *Journal of Causal Inference* 4.2 (Sept. 2016). ISSN: 2193-3677. DOI: 10.1515/jci-2016-0009.

[81] Stefan H. Steiner et al. "Planning and analysis of measurement reliability studies". In: *The Canadian Journal of Statistics* 39.2 (Sept. 2011), 344–n/a355. DOI: 10.1002/cjs.

[82] Daniel L. Swain et al. "Increasing precipitation volatility in twenty-first-century California". In: *Nature Climate Change* 8.5 (May 2018), pp. 427–433. ISSN: 17586798. DOI: 10.1038/s41558-018-0140-y.

[83] Ben Timmermans et al. "An evaluation of the consistency of extremes in gridded precipitation data sets". In: *Climate Dynamics* 52.11 (June 2019), pp. 6651–6670. ISSN: 14320894. DOI: 10.1007/s00382-018-4537-0.

[84] Shahid Ullah and Caroline F Finch. *Applications of functional data analysis: A systematic review*. Tech. rep. 2013. URL: http://www.psych.mcgill.ca/misc/fda/.

[85] Union of Concerned Scientists. *UCS Satellite Database*. 2023.

[86] USGS. *Wildland Fire Potential Index*. 2021.

[87] Aaron Van Donkelaar et al. "Regional Estimates of Chemical Composition of Fine Particulate Matter Using a Combined Geoscience-Statistical Method with Information from Satellites, Models, and Monitors". In: *Environmental Science and Technology* 53.5 (Mar. 2019), pp. 2595–2611. ISSN: 15205851. DOI: 10.1021/acs.est.8b06392.

[88] Stefan Voigt et al. *10 National Disaster Reduction Center of China Beijing, China. 11 Geneva International Centre for Humanitarian Demining*. Tech. rep., p. 2023. URL: https://www.science.org.

[89] J. Walsh et al. *Ch. 2: Our Changing Climate. Climate Change Impacts in the United States: The Third National Climate Assessment*. Tech. rep. Washington, DC: U.S. Global Change Research Program, 2014, pp. 19–67. DOI: 10.7930/J0KW5CXT. URL: https://nca2014.globalchange.gov/downloads.

[90] Sherrie Wang, François Waldner, and David B. Lobell. "Unlocking Large-Scale Crop Field Delineation in Smallholder Farming Systems with Transfer Learning and Weak Supervision". In: *Remote Sensing* 14.22 (Nov. 2022). ISSN: 20724292. DOI: 10.3390/rs14225738.

[91] Frederick V. Waugh. "Quality as a Determinant of Vegetable Prices; a Statistical Study of Quality Factors Influencing Vegetable Prices in the Boston Wholesale Market". In: *American Journal of Agricultural Economics* 11.4 (1929), pp. 525–702. ISSN: 0002-9092. DOI: 10.1093/ajae/11.4.679.

[92] Jeffrey M. Wooldridge. "Should instrumental variables be used as matching variables?" In: *Research in Economics* 70.2 (June 2016), pp. 232–237. ISSN: 10909443. DOI: 10.1016/j.rie.2016.01.001.

[93] Christopher Yeh et al. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". In: *Nature Communications* 11.1 (Dec. 2020). ISSN: 20411723. DOI: 10.1038/s41467-020-16185-w.