ESSAYS IN APPLIED MACHINE LEARNING AND CAUSAL INFERENCE.

by

CONNOR LENNON

A DISSERTATION

Presented to the Department of Economics
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2022

DISSERTATION APPROVAL PAGE

Student: Connor Lennon

Title: Essays in Applied Machine Learning and Causal Inference.

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Economics by:

Glen Waddell	Chair
Edward Rubin	Chair
Grant McDermott	Core Member
Humphrey Shi	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded September 2022

© 2022 Connor Lennon
This work, including text and images of this document but not including supplemental files (for example, not including software code and data), is licensed under a Creative Commons
Attribution 4.0 International License.

DISSERTATION ABSTRACT

Connor Lennon

Doctor of Philosophy

Department of Economics

September 2022

Title: Essays in Applied Machine Learning and Causal Inference.

This dissertation represents a study of how machine learning can be incorporated into existing econometric causal techniques, with explorations both in the costs and benefits of making that choice. The first chapter explores a simulated instrumental variables setting to evaluate the ease of incorporating unmodified machine learning techniques into the ”first stage“ problem. The first stage of two-stage least squares (2SLS) is a prediction problem—suggesting gains from utilizing ML in 2SLS's first stage. However, little guidance exists on when ML helps 2SLS—or when it hurts. We investigate the implications of inserting ML into 2SLS, decomposing the bias into three informative components. Mechanically, ML-in-2SLS procedures face issues common to prediction *and* causal-inference settings—and their interaction. Through simulation, we show linear ML methods (e.g.post-Lasso) work “well,” while nonlinear methods (e.g.random forests, neural nets) generate substantial bias in second-stage estimates—some *exceeding* the bias of endogenous OLS. This work was performed in conjunction with professors Edward Rubin and Glen Waddell. The chapter author wrote simulation code, excepting the substantial portions used for table creation and to iterate over differing methods, to evaluate and run the methods tested in this chapter, and we

designed the DGP function based on those found in Belloni, Chen, Chernozhukov, and Hansen (2012).

The second chapter is an applied use of Machine Learning to evaluate an existing causal estimate of property value on suppression costs in the Wildfire Economics space. Models in use currently rely on excluding class A-D wildfires that burn fewer than 300 acres, use property values as an input and feature differential estimates for per-acre suppression costs in the Eastern and Western United States. However, restricting suppression cost estimates to large fires ignores wildfires that have high per-acre costs due to aggressive initial-attack strategies, and fires occurring in well-managed forests with fewer suppression requirements, which may lead SCI-derived estimates of cost to be biased and potentially be overly responsive to changes in local wealth. Using double/debiased vision transformers, SCI parameters overestimate the impact of property value as a contributor to suppression costs.

This dissertation includes unpublished and co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Connor Lennon

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
University of Puget Sound, Tacoma, WA, USA

DEGREES AWARDED:

Doctor of Philosophy, Economics, 2022, University of Oregon
Master of Science, Economics, 2018, University of Oregon
Bachelor of Science, Economics, 2010, University of Puget Sound

AREAS OF SPECIAL INTEREST:

Econometrics
Machine Learning
Environmental Economics

PROFESSIONAL EXPERIENCE:

Instructor, Principles of Microeconomics, 2020-2022
Instructor, Intro to Econometrics, 2020-2021
Doctoral Economics Intern, Resources for the Future, 2020
Research Assistant, 2018
Graduate Employee, Grader and Lab Instructor, 2017-2022

GRANTS, AWARDS AND HONORS:

Economics EG Daniel Scholar, 2018

PUBLICATIONS:

ACKNOWLEDGEMENTS

I thank my advisors, Glen Waddell and Edward Rubin for constant insight and for helping me become a better econometrician, and who always gave me the right questions to guide my ideas in a positive direction. Without their input and influence this document would surely be condiderably less valuable.

I thank Glen in particular for showing me how to take my ideas, in every state and find the bright spots within it, which over repeated iterations can be turned into something of merit. His course in econometrics gave me the foundations on which I have built my understanding.

I thank Ed for his help learning how to take complex ideas and breaking them down into their component parts to explain them at a variety of levels, and offered me endless support of my teaching. His willingness to listen and always trying to hear what I had to say built my confidence to speak when there was something important to voice.

I also thank my committee, who have been understanding and very helpful in allowing me to grow into a better writer, presenter and researcher over my time at the university. I also must thank Grant McDermott and Edward Rubin for encouraging my early exploration of machine learning and Glen Waddell to encourage me to consider its use for causal applications.

To my co-chairs and advisors, Drs Edward Rubin and Glen Waddell, who helped me get my footing during covid, through transformative periods of my life and were advocates on my behalf for the better, and without whom I would most certainly have struggled substantially more. Lastly, to my wife Mandy for supporting me in undertaking this quixotic journey into academia, and my family who somehow still put up with me.

TABLE OF CONTENTS

Chapter Page

I. MACHINE LEARNING INSTRUMENTAL VARIABLES 1

1.1. Prelude 1

1.2. Ch.1 Introduction 1

1.3. Models 1

1.3.1. The problem 1

1.3.2. Estimators 1

1.4. Data-generating processes 1

1.4.1. A *low-complexity* case 1

1.4.2. A *high-complexity* case 1

1.5. Results 1

1.5.1. Which hammer? 1

1.5.2. Decomposing the bias 1

1.5.2.1. The $\beta_1 \text{Cov}(\hat{x}, e)$ component 1

1.5.2.2. The $\text{Cov}(\hat{x}, u)$ component 1

1.5.2.3. The $\frac{1}{\text{Var}(\hat{x})}$ component 1

1.6. Potential solutions 1

II. REVISITING THE STRATIFIED COST INDEX 2

2.1. Ch.2 Introduction 2

2.2. Model of Cost 2

2.3. Econometric Model 2

2.3.1. Wildfire spread 2

2.3.2. "Damages" from Wildfire 2

Chapter Page

- 2.3.3. Property Valuation and Environmental Amenities 2
- 2.4. Unbiasedness through Twin Vision Transformers 2
  - 2.4.1. Transformer 2
  - 2.4.2. Combining Minimum Travel Time, NVC and Transformer Decoders 2
  - 2.4.3. The adjustment model 2
  - 2.4.4. 'Causal' Model 2
- 2.5. Data 2
- 2.6. Results 2
  - 2.6.1. Difference in estimated ATE of Property Value on Suppression Costs 2
  - 2.6.2. Model Comparisons 2
    - 2.6.2.1. A note on implied efficiency 2
  - 2.6.3. Evaluating the Conditional ATEs 2
- 2.7. Conclusion 2

APPENDICES

- A. CH 1 FIGURE APPENDIX 3
- B. CH 1 TABLE APPENDIX 8
- C. CH 2 FIGURE APPENDIX 11
- D. CH 2 TABLE APPENDIX 28
- E. APPENDIX CHAPTER 1 34
  - E.1. Appendix: Math 35
    - E.1.1. Wedge a covariance and correlation 35
    - E.1.2. Appendix: MLP/Neural cross-validation procedure 35
    - E.1.3. Low-bias methods 37

Chapter Page

- E.2. Appendix: Neural approaches to measuring causal response 37
- E.3. Appendix: The complications of machine learning and the monotonicity assumption 39

F. APPENDIX CHAPTER 2 41

- F.1. Model Training and Hyperparameter Selection 41
  - F.1.1. Hyper Parameter Selection 41
  - F.1.2. On instability 41
- F.2. Integrated Gradient estimation of Cost-Important Factors in Selected Fires 42
- F.3. Out of Distribution Performance 42
- F.4. The Non-linear model Hypothesis (testing other regression models) 43

REFERENCES CITED 45

LIST OF FIGURES

Figure Page

A1. **The "high-complexity" DGP:** 100 instruments of varying strength . . 4

A2. Main results—$\hat{\beta}$ distributions across competing two-stage methods . . 4

A3. **Exclusion-restriction violations** via higher-order interactions among instruments ('low-complexity case' of 7 strong instruments) 4

A4. **Predictions *vs.* estimation:** Comparing cross-validated prediction performance with bias in random-forest-based 2SLS 4

A5. **Distributions of estimates** with "exclusion-restriction violations" from *k*-term interactions ('low-complexity case) 5

A6. Explaining unrestricted/narrow neural networks' bimodal distributions of $\hat{\beta}$ 6

A7. **Distributions of estimates** under heterogenous treatment effects . . . 7

C1. Example Spatial Input for Transformer: CZU Complex Fire 30×30 *km* 12

C2. Example Spatial Input for Transformer, with Cut Lines 13

C3. Example Single Patch Inputs 3×3 *km* 14

C4. **ViT Fit** Fire Model 15

C5. **Regional Estimates of** θ : Conditional Estimates of Elasticity with Respect to Property Value . . . 16

C6. SCM: Property Value's Effect on Fire Suppression 17

C7. SCM: Property Value's Effect on Fire Suppression 18

C8. Train/Validation Loss for Twin Transformers 19

C9. Effects of Property Value on Suppression Costs, Conditioned on Monitoring 20

C10. Conditional Effects of Property Value on Fires of Differing Sizes 21

Figure Page

C11. τ **Correlation A:** Non-Parametric Correlations Between Predictor Variables and Economic Variables of Interest 22

C12. τ **Correlation B:** Non-Parametric Correlations Between Predictor Variables 23

C13. **Regional Statistics:** Coordination-Center Level Figures 24

C14. **Integrated Gradients:** Holiday Farm Fire 25

C15. **Integrated Gradients:** Silverado Wildfire 26

C16. Compact Convolutional Multimodal Transformer 27

LIST OF TABLES

Table Page

A1. **Simulation results**: Mean and standard deviation for methods and DGPs . . . 9

A2. **Simulation results**: Decomposing bias components (means from simulation) . . . 10

E1. Data and Descriptive Statistics . . . 29

E2. Causal Equations . . . 30

E3. Results: All Fire Sizes . . . 31

E4. Results: Fires over 100 Acres . . . 32

E5. Results: Nonlinear Ignition-Point Models . . . 33

CHAPTER I

MACHINE LEARNING INSTRUMENTAL VARIABLES

1.1 Prelude

This work was completed in collaboration with Dr.s Edward Rubin and Glen Waddell.

1.2 Ch.1 Introduction

Today machine learning is everywhere—from exciting applications in image processing, linguistics, and forecasting, to obligatory sections in job-market papers, to the increasingly common seminar question: *Have you tried machine learning?* With this recent popularity, machine learning (ML) methods are appearing in an increasingly wide range of empirical econometric applications. Despite this excitement and frequent recommendations, the literature has little to say regarding the appropriateness of using two-stage least squares (2SLS) with ML methods (e.g.what are the benefits and costs of ML-augmented 2SLS with regards to unconfoundedness, exogeneity, and strength of instruments).[1]

In this paper, we focus on a potentially promising application of ML methods: curating and generating the first-stage predictions of two-stage least squares.

The motivation behind integrating machine learning in two-stage least squares is clear: to the extent that researchers can incorporate "better" first-stage predictions, researchers obtain more-precise second-stage estimates. Because

---

[1] Here, we've closely paraphrased Jeffrey Wooldridge's Twitter post of 26 April 2021, where he continues with "Even worse would be if ML becomes a *de facto* requirement for empirical work in cases where its benefits are questionable—or even when ML might be harmful." In terms of recommendation, Mullainathan and Spiess (2017) writes that "Machine learning... revolves around prediction" and "belongs in the part of the toolbox marked $\hat{y}$ rather than in the more familiar $\hat{\beta}$ compartment." The authors then immediately recognize that "the first stage of a linear instrumental variables regression is effectively prediction."

most ML methods are built explicitly for prediction—they typically outperform ordinary-least squares (OLS) at this task—using ML for first-stage predictions seems quite natural.[2] The risks of adopting out-of-the box ML methods for 2SLS-type applications are less clear.

In this paper, we discuss several phenomena that can bias ML-based 2SLS away from its target parameters. Some of these phenomena are implications of the *forbidden regression*, which naïve implementations of ML in 2SLS are likely to lead to—injecting predictions from a nonlinear estimator into the first stage of 2SLS (J. Angrist & Pischke, 2009; J. D. Angrist & Krueger, 2001; Wooldridge, 2010).[3,4] If a linear first stage adequately approximates the

---

[2] This point has been recognized in the literature (J. Angrist & Frandsen, 2020; Belloni, Chernozhukov, & Hansen, 2011; J. Chen, Chen, & Lewis, 2020; Chernozhukov et al., 2018; Singh, Sahani, & Gretton, 2019; **?**; **?**; **?**; **?**) inclusive of new artificial intelligence (AI) methods (Bennett, Kallus, & Schnabel, 2020; Hartford, Lewis, Leyton-Brown, & Taddy, 2017; Liu, Shang, & Cheng, 2020). Further, the *ad hoc* integration of 2SLS and ML is already appearing in applied work across a wide range of fields—estimating labor market impacts of imprisonment (Mueller-Smith, 2015), the effects of racial-composition shocks during the Great Migration (Derenoncourt, 2019), the effect of expropriation on growth (D. L. Chen & Yeh, 2020), the "true" size of China's GDP growth (W. Chen, Chen, Hsieh, & Song, 2019), the inter-generational transmission of health (Bevis & Villa, 2020), and the heterogeneous impacts of family size and parental labor supply (Biewen & Kugler, 2020). In a recent working paper, J. Chen et al. (2020) also recognizes this motivation, suggesting that the traditional OLS-based implementation of 2SLS "leaves on the table some variation provided by the instruments that may improve precision of estimates." *If* one is willing to accept the fairly strong assumption that any function (nonlinear or linear) of valid instruments is itself a valid instrument, J. Chen et al. (2020) provides an interesting solution to some of the challenges involved with including ML methods in 2SLS. We do not make this assumption.

[3] With valid instruments, applying OLS in the first stage of 2SLS produces predictions ($\hat{x}$) that are a linear combinations of the exogenous instruments. Thus, $\hat{x}$ is itself exogenous in the traditional 2SLS procedure. Predictions produced by nonlinear functions are not guaranteed to be orthogonal to their residuals, generating additional bias/inconsistency in second-stage estimates.

[4] Another flavor of the *forbidden regression* involves applying different specifications of controls in the first and second stages. Most out-of-the-box ML methods do not offer a method to ensure that second-stage controls are used for prediction in the ML-based first stage (and *in the correct functional form*). There are *ad hoc* solutions to this problem—writing custom functions that implement the ML algorithm *plus* a linear specification of the controls/fixed effects, or residualizing (i.e., Frisch-Waugh-Lovell). For an example, see the `fixest` package in `R` and its `feNmlm()` function, which is written to efficiently estimate maximum likelihood models with multiple fixed-effect (i.e.large factor variables). This issue is particularly important for situations

underlying data-generating process (DGP), then the exclusion restriction reduces to a simple assumption that the instruments $\mathbf{z}$ are uncorrelated with the endogenous disturbance u. If one introduces nonlinearity into the first stage, then the assumption of "no correlation" must be strengthened to conditional mean-independence between the instruments and disturbance. This (stronger) assumption requires more careful consideration of the structural relationships of $\mathbf{z}, x, y$.

Other issues are less common to "traditional" econometrics but become key to understanding ML-based results. These include:

- **Recovering endogeneity:** If the prediction algorithm is *too* good, then the first-stage predictor may entirely recover the endogenous regressor (including both *good* and *bad* variation). With (*i*) a small set of valid instruments and (*ii*) a linear estimator (e.g.OLS), this scenario is of less concern. As the number of potential instruments increases and the estimator becomes more nonlinear and flexible (a hallmark of many ML methods), we show that this concern becomes real.
- **Exclusion restrictions:** ML methods are not designed to choose exclusion restrictions. If a researcher relies on ML methods to determine a nonlinear functional form, choose instruments, and select first-stage controls in a 2SLS framework, then she ultimately must assume that the algorithm is capable of settling on a valid exclusion restriction—placing a lot of trust in ML to do something it is not typically designed to do. As J. Angrist and Frandsen (2020) point out, nonlinear estimators generate nonlinear combinations of

---

where conditioning on controls/fixed effects is integral to the instruments' exogeneity. Again, ML methods will, in this way, expose researchers to potential pitfalls.

the original instruments and thereby require additional exclusion restrictions *beyond* the original exclusion restriction implied by the linear combination of the instruments. With highly flexible ML methods, the set of exclusion restrictions is nearly infinite—the researcher must either assume that (*i*) the ML algorithm will choose the appropriate exclusion restrictions or (*ii*) *all* possible exclusion restrictions are valid (as the algorithm's choice set is infinite).

- **Amplified bias:** As we show below, the bias of second-stage estimates in 2SLS is inversely related to the variance of the first-stage predictions ($\hat{x}$). Most ML methods reduce variance in the predictions (to reduce out-of-sample prediction performance in the canonical *bias-variance tradeoff*). This variance-reduction strategy leads to inflated bias in second-stage applications—a consideration not typical to OLS-based 2SLS applications.

We show that most ML-rooted solutions that use common ML procedures in the first stage of 2SLS fail to improve upon standard 2SLS (i.e.using OLS in the first stage)—and generate more bias. Two linear estimators—post-Lasso selection and principal component analysis (PCA)—are the exceptions. Post-Lasso and PCA perform as well, or better, than standard OLS-based 2SLS. Perhaps more importantly, we show that highly nonlinear tree-based methods (e.g.random forests and boosted trees) can amplify bias—providing parameter estimates farther from *truth* than naïve OLS regressions that ignore endogeneity. Given sufficient training time, naïve implementations of neural networks in 2SLS can reproduce the original OLS bias—providing little to no advantage over traditional approaches to recovering exogenous identifying variation through 2SLS. [5]

Ultimately we conclude that while ML methods offer many promises for a range of applications, most out-of-the-box ML methods are not well suited for two-stage least squares. Moreover, applying the *wrong* ML method in the first stage can actually generate more bias in parameter estimates than entirely ignoring endogeneity.

In Section 1.3 we formalize the theoretical settings and define the estimators. In Section 1.4 we introduce two data-generating processes—respecting that use cases will likely differ, we detail one that is rather simple in its construction and another that is more complex. In Section 2.6 we present the empirical results for the discussed estimators and DGPs. Finally, in Section 1.6 we conclude.

1.3 Models

1.3.1 The problem.

Applied researchers commonly apply 2SLS to estimate the causal effect of some x on some y in a setting where the exogeneity of x cannot reasonably be assumed. In other words, where

$$y = \beta_0 + \beta_1 x + u \ , \tag{1.1}$$

there is concern over the potential for non-zero covariance between the variable of interest x and the disturbance u when estimating the parameter β_1.

---

[5] This ignores the practical as well - on our resources, the simulation for neural network frequently took several hours to complete which is considerably longer than the time it takes to run a traditional 2sls

Let **z** denote a vector of *instrumental variables*, we express the first stage of a 2SLS estimates x as a function of these instruments:

$$x = f(\mathbf{z}) + \varepsilon \; . \tag{1.2}$$

In its traditional OLS-based implementation, $f(\mathbf{z})$ is linear in **z**.

Defining the predictions from (**??**) as $\hat{x} = f(\mathbf{z})$, the second stage of the 2SLS procedure then regresses the outcome variable y on $\hat{x}$,

$$y = \gamma_0 + \gamma_1 \hat{x} + w \; , \tag{1.3}$$

to achieve an estimate for β_1 in (**??**)—we let $\hat{\gamma}_1$ be this estimate of β_1. If the instruments are valid (i.e., predictive of x and uncorrelated with u) and $\hat{x}$ results from an OLS regression, then $\hat{x}$ will also be exogenous.[6] The second stage of OLS-implemented 2SLS then generates consistent estimates of β_1, interpreted as the causal effect of x on y.

So why adopt ML at all? Applications of 2SLS identify the effect of x on y by extracting only a fraction of the "good" (exogenous) variation in x. The hope for ML-infused 2SLS methods is that researchers can extract more of the good variation in x—nonlinear combinations of the instruments, specifically—while still omitting the bad variation. This desire has likely increased following Lee, McCrary, Moreira, and Porter (2020), which argues that many traditional evaluations of instrumental variables considerably overestimate their significance.

---

[6] We are assuming homogeneous treatment effects, which removes the requirement of monotonicity. For an inspection into monotonicity under heterogenous treatment effects in the machine learned case, see E.3

1.3.2 Estimators. In the analysis below we examine three classes of 2SLS-motivated estimators:

Class 1: 'Traditional' two-stage regression methods: This set of estimators covers the standard two-stage regression estimators in an econometrician's toolbox: two-stage least squares, (unbiased) split-sample IV (J. D. Angrist & Krueger, 1995), the Fuller implementation of limited-information maximum likelihood (LIML) (Anderson & Rubin, 1949; Fuller, 1977), and jackknife IV (JIVE) (J. D. Angrist, Imbens, & Krueger, 1999). These methods overlap in three important ways: they (*i*) employ a two-stage approach (*ii*) whose first stage creates a linear combination of the instruments (*iii*) with no formal variable selection.

Class 2: Machine-curated variable selections in standard 2SLS: This second class augments the standard OLS-based version of 2SLS with variable selection/synthesis. Specifically, these methods feature an additional procedure, *prior to the first stage*, that downselects or combines $\mathbf{z}$ into a more parsimonious set of variables—it is the elements of this more parsimonious expression of $\mathbf{z}$ that then appear in the first stage. The rest of the 2SLS process proceeds as usual (i.e.OLS). Importantly, while these models feature variable selection or synthesis, they also preserve linearity in both stages—without any regularization or penalization. Because these estimates result from linear combinations of $\mathbf{z}$, the original exclusion restriction of $\mathbf{z}$ passes through to the selected/synthesized instruments.

The first of machine-curated methods is the post-Lasso procedure of Belloni et al. (2012) that first estimates the linear relationship between x and $\mathbf{z}$ (a linearized version of **??**) using penalized regression. This penalized regression minimizes the sum of squared error (SSE) *plus* a penalty proportional to the sum of the coefficients' magnitudes. That is, $\lambda \times ||\gamma||$, where γ is the vector of coefficients on the (standardized) instruments and λ is a the shrinkage parameter chosen by the researcher (typically via cross validation). Because each instrument's coefficient-based penalty changes discontinuously when moving from $\gamma_i = 0$, Lasso can be used to select a set of *stronger* instruments (whose coefficients are non-zero). Post-Lasso selects the instruments whose coefficients are non-zero and then estimates standard, OLS-based 2SLS using those selected instruments.[7]

Principal-component analysis (PCA) offers an alternative route to simplifying $\mathbf{z}$ by selecting $\mathbf{z}$'s first k principal components (Pearson, 1901). Thus, as the second machine-curated method we consider, Principal-component analysis (PCA) applied to 2SLS (as in Ng and Bai (2009), and Winkelried and Smith (2011), e.g.) passes this set of principal components into the first stage of standard OLS-based 2SLS. While PCA may reduce the first stage's interpretability, this approach can drastically reduce the number of first-stage instruments while retaining considerable explanatory power.

Class 3: ML-based first stages in 2SLS: Our final class of estimators retains the general two-step framework of 2SLS but replaces the first stage with a variety of cross-validated ML algorithms. We evaluate a meaningful subset of machine-learning methods suitable for regression, including random forest (Breiman, 2001; Ho, 1995), boosted trees (Breiman, 1998; Friedman, 2001, 2002; L. Mason, Baxter, Bartlett, & Frean, 1999), neural networks (Farley & Clark, 1954; McCulloch & Pitts, 1943; Turing, 1948), and Lasso (Santosa & Symes,

---

[7] J. Angrist and Frandsen (2020) notes that this methodology may suffer from potentially unseen pre-test bias. Because our model comes from relatively strong instruments, as with the intuition of Zhao, Witten, and Shojaie (2020), we do not estimate de-biased Lasso models. We therefore allow post-Lasso to serve as a representation of both.

1986; Tibshirani, 1996).[8,9] Notably, most of these algorithms offer considerable flexibility (e.g.nonlinearity in $\mathbf{z}$) and variable selection (to varying degrees). This class offers considerable insights into the merits of off-the-shelf ML methods' for machine-assisted 2SLS.

1.4 Data-generating processes

In order to examine the performance of ML in the predictive stage of 2SLS—in absolute terms and relative to *traditional* options—we employ two general data-generating processes (DGPs). For reasons that will become clear as we describe each, we refer to them as the *low-complexity* case and the *high-complexity* case. While subjective, our intention is to provide bookends of a sort, as the applied researcher rarely knows the extent to which her case is *complex*—particularly in terms of extent of nonlinearity or the efficient number of instruments.

1.4.1 A *low-complexity* case. The motivation for this case is to depict the estimators' performances when the DGP is simple and closely matches the ideal scenario for OLS-based 2SLS: an endogenous regressor that is a linear combination of a relatively small set of strong, valid instruments. For example, we imagine this case appealing to researchers seeking to estimate the causal effect of a variable of interest x_1 on outcome y, i.e.

$$y = \beta_0 + \beta_1 x_1 + \varepsilon_y \,, \tag{1.4}$$

---

[8] For our purposes, the contributions of Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014) (dropout), Ioffe and Szegedy (2015) (batch normalization), and Kingma and Ba (2017) (stochastic optimization) are particularly relevant.

[9] For a nice review of ML methods in applied economics—including Lasso, tree-based methods, and neural networks—please see Storm, Baylis, and Heckelei (2019). For broader and more in-depth coverage, see James, Witten, Hastie, and Tibshirani (2013) and Hastie, Tibshirani, and Friedman (2009).

but facing the challenge—omitted variables, simultaneity, *etc.*—that x_1 is endogenous and $\mathrm{E}\left[\varepsilon_y \middle| x_1\right] \neq 0$ prevents OLS from cleanly identifying β_1 in (**??**). (Note that the causal effect β_1 is common across all individuals—this ensures that differences across estimators are not due to the estimators recovering different local average treatment effects (LATEs).)

In this low-complexity scenario, ML-based 2SLS methods are overkill: neither variable selection, nor nonlinearity are necessary. As our results demonstrate, ML methods can increase bias relative to 2SLS and even endogenous OLS.

Formally, to model a scenario with a single endogenous regressor (x_1) and a small set of valid (and *individually* strong) instruments, we define the DGP as

$$\begin{aligned} \varepsilon_y &= \beta_2 x_2 + \eta \ , \\ x_2 &= 1 + \varepsilon_c \ , \ \text{and} \\ x_1 &= g_x(\mathbf{z}) + \varepsilon_c \ , \end{aligned}$$

drawing special attention to the inclusion of ε_c as the disturbance common to both x_1 (the variable of interest) and x_2 (the *omitted* variable). This common error follows a standard normal distribution; η is distributed uniformly between -1 and 1.

We assume that a set of valid instruments $\mathbf{z}$ exists such that $\mathrm{E}\left[\varepsilon_y \mid \mathbf{z}\right] = 0$ and $\mathrm{E}[x_1 \mid \mathbf{z}] \neq 0$ (we focus on the case where $|\mathbf{z}| = 7$). We also anticipate that the researcher has no beliefs or insights about the functional form of $g_x(\cdot)$, as is likely the case in practice. In the true DGP for this case, $g_x(\mathbf{z}) = \sum_{i=1}^{7} z_i$. That is, $g_x(\cdot)$ is linear.

In particular, we draw the instruments **z** from a multivariate normal distribution centered at zero (i.e.$\mathrm{E}[\mathbf{z}] = \mathbf{0}$) with variance-covariance matrix $\Sigma_{\mathbf{z}}$ where $\mathrm{Cov}(z_i, z_j) = 0.6^{|h-k|}$ (and thus $\mathrm{Var}(z_i) = 1$ for each i). By implication, $x_1 \sim N(0, \text{Grand Sum}(\Sigma_{\mathbf{z}}) + 1)$.

In full, then, the data represents the following system of equations:

$$\begin{aligned}
y &= \beta_0(=1) + \beta_1(=1)x_1 + \beta_2(=1)x_2 + \eta \ , \\
x_2 &= 1 + \varepsilon_s \ , \\
x_1 &= g_x(\mathbf{z}, \varepsilon_s) = \sum_{i=1}^{7} z_i + \varepsilon_s \ .
\end{aligned}$$

Importantly, the specification of the instruments in this DGP produces a very strong first-stage with a relatively large *concentration parameter* Belloni et al. (2012). Put simply, the concentration parameter μ^2 describes the extent to which the weak-instrument problem may arise within a given DGP. A higher value of μ^2 implies that 2SLS, without variable selection, will converge to the true β_1 at relatively small sample sizes.[10] (We discuss this further in the *high-complexity case* section below.) Consequently, the low-complexity case allows us to test how machine-curated first stages perform when there is little to be gained from variable selection/synthesis.

1.4.2 A *high-complexity* case. As our high-complexity case, we follow the DGP developed by Belloni et al. (2012) with two extensions. This DGP allows the researcher to customize instruments' strength, and with *many* instruments.

---

[10] In this “low-complexity" case, $\mu^2 \approx n \times 20.71$, which exceeds the values in Belloni et al. (2012).

Following Belloni et al. (2012), the DGP in our high-complexity case results from

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \varepsilon_y \ , \\
x_1 &= \boldsymbol{\pi}\mathbf{z} + \varepsilon_v \ ,
\end{aligned}
$$

where

$$
\begin{aligned}
(\varepsilon_y, \varepsilon_v) &\sim N\left(0, \begin{bmatrix} \sigma_y^2 & \sigma_y\sigma_v \\ \sigma_v\sigma_y & \sigma_v^2 \end{bmatrix}\right) , \\
\mathbf{z} &= \begin{bmatrix} z_1 & z_2 & \cdots & z_{100} \end{bmatrix} \sim N(\mathbf{0}, \Sigma_z) \ , \\
\Sigma_z[j, j] &= \mathrm{Var}(z_j) = \sigma_j^2 = 1, \ \ \forall j \in \{1, \ldots, 100\} \ , \textit{and} \\
\Sigma_z[j, k] &= \mathrm{Cov}(z_j, z_k) = \mathrm{Cor}(z_j, z_k) = 0.6^{|j-k|}, \ \ \forall (j, k) \in \{1, \ldots, 100\} \ .
\end{aligned}
$$

As before, the researcher's interest is in identifying β_1. However, unlike the earlier DGP, the high-complexity case produces sets of relevant and exogenous instruments that vary in their correlation and individual strength (i.e.π_i).

In defining the "exponential" design of the *first-stage* coefficient vector $\boldsymbol{\pi}$, we follow Belloni et al. (2012): $\boldsymbol{\pi}$ captures a "beta pattern" $\widetilde{\boldsymbol{\pi}} = (0.7^0, 0.7^1, 0.7^2, \ldots, 0.7^{99})$ that is then multiplied by a constant C, i.e., $\boldsymbol{\pi} = C \times \widetilde{\boldsymbol{\pi}}$. The constant C implies a value for the concentration parameter, $\mu^2 = \frac{n\boldsymbol{\pi}'\Sigma_z\boldsymbol{\pi}}{\sigma_v^2}$.[11] Panels **??**–**??** (Figure **??**)[12] illustrates the three beta patterns that we adopt in the 'high-complexity' DGP—generating three subcases of this DGP. As described above, and in greater depth in Hansen, Hausman, and Newey (2008), the concentration

---

[11] For a proof of this statement, see Belloni et al. (2012).

[12] All figures in this document are located in the appendix: see section A for figures referenced in this chapter.

parameter is useful for determining the behavior of IV estimators. Because we are less interested in the case of weak instruments, we use $\mu^2 = 180$, which creates a strong set of instruments as outlined in Belloni et al. (2012).[13]

Belloni et al. (2012) arrange the coefficients $\boldsymbol{\pi}$ in descending order (i.e.$\pi_1 > \pi_2 > \cdots > \pi_{100}$). However, the definition of Σ_z implies that 'proximate' instruments are more correlated than 'distant' instruments—i.e.$\text{Cor}(z_i, z_{i+1}) > \text{Cor}(z_i, z_{i+k})$ for $k > 1$. Thus, the DGP of Belloni et al. (2012) ensures that it is the strongest instruments that are correlated with each other. While this feature may be desirable in many contexts, we will remain agnostic with regard to whether the strongest instruments are most correlated with each other or with other instruments. However, this does require that we consider three sub-cases that each arise from different orderings of the coefficients in $\boldsymbol{\pi}$:

- **Randomly shuffled:** After generating the coefficients, we randomly re-order them to break the relationship between instruments' strengths and their covariance (Σ_z).
- **Descending from** z_1: In this subcase, as in Belloni et al. (2012), $\pi_1 > \pi_2 > \cdots > \pi_{100}$.
- **Descending from** z_{50}: Here we modify Belloni et al. (2012) by defining π_{50} as the largest coefficient: $\pi_{50} > \pi_{51} > \cdots > \pi_{100} > \pi_1 > \pi_2 > \cdots > \pi_{49}$. Because "proximate" instruments are correlated in Σ_z, this subcase implies that the strongest instrument (z_{50}) is very correlated both with the second-strongest instrument (z_{51}) and with the weakest instrument (z_{49}).

---

[13] It is important to select this value thoughtfully. Choosing a μ^2 that is too small will simulate a weak-instruments problem. Choosing a μ^2 that is too large will yield a scenario in which all instruments are "overpoweringly" valid, which reduces the effectiveness of selection or dimension-reduction techniques.

Finally, we define $\sigma_v^2 = \boldsymbol{\pi}'\Sigma_z\boldsymbol{\pi}$ (which forces that $\text{Var}(x_1) = 1$) and $\sigma_y = 1$. In panels **??-??** of Figure **??** we illustrate the cross-instrument correlations implied by Σ_z: in Panel **??** we show a correlation matrix among the 100 instruments, and in Panel **??** we highlight the correlation of z_1 and z_{50} to each of the other 100 instruments. Instruments are strongly correlated with their neighbors and weakly correlated with non-neighbors—limiting the information accessible from any single instrument.

1.5 Results

Now we turn to discussing the results of our simulations. Among the simulations, we will include an "oracle model" that extracts the exogeneous component of x_1 in its entirety (perfectly removing endogeneity) and a simple OLS model (where we entirely ignore endogeneity). While one might expect the oracle and plain OLS models to bookend 2SLS models, our simulations demonstrate that they *do not*. That is, machine learning can lead to outcomes that are even worse than ignoring endogeneity.

In each case, we are interested in the performances of the estimators in terms of their potential biases and the precision of estimates. Recall that these estimators include three broad classes: (*i*) traditional methods (OLS-based 2SLS, split-sample IV, LIML, and jackknife IV), (*ii*) machine-curated 2SLS (variable-selection or -curation via post-Lasso and PCA), and (*iii*) 2SLS applications with ML-powered predictions in their first stages (i.e.replacing first-stage OLS with either Lasso, boosted trees, random forests, or neural networks).

1.5.1 Which hammer?. In Figure A2 we depict the distributions of point estimates ($\hat{\beta}_1$) for a given method in given DGPs—in Panel **??** we illustrate the low-complexity case, and in panels **??–??** we represent the high-complexity cases.[14] We summarize simulations by their means and standard deviations in Table A1.

To those with use cases that resemble our "low-complexity case," the simulation results have a clear takeaway: PCA-based 2SLS and post-Lasso perform well and offer very safe choices.[16] Important for the practitioner: All four nonlinear ML-in-the-first-stage methods (i.e.Lasso, boosted trees, neural networks, and random forests) perform poorly in terms of both bias and variance. In fact, "random-forest infused 2SLS" generates *more bias* in $\hat{\beta}_1$ than the OLS estimator that entirely ignores endogeneity—it is possible for an ML-based 2SLS estimator to *amplify* bias relative to plain OLS.[17]

In the three high-complexity cases in Table A1 (columns *B*–*D*) and in panels **??**–**??** of Figure A2, LIML and Jackknife IV generate very little bias in their estimates of β_1, outperforming 2SLS. Across all three DGPs, 2SLS produces mean estimates roughly 2.3–5.8 percent larger than the true parameter, while the centers of LIML's and JIVE's distributions are within 0.4 percent of the true parameter. Injecting random forests into the first stage, on average, produces more biased estimates than naïve (endogenous) OLS—generating coefficient estimates that are 32–56 percent larger than the true estimates. In short, one can worsen endogeneity issues by using ML-based 2SLS estimators.

1.5.2 Decomposing the bias.

To diagnose the sources of bias from different methods, we show below one can decompose the wedge between β_1 and

---

[14] The target parameter β_1 equals 1 throughout (indicate with a thin dashed line). Each distribution results from 1,000 iterations of the simulation. Table A1 summarizes[15] each of these method-by-DGP combinations (i.e.14 × 4 = 56) with the mean and standard error from each.

[16] LIML also performs well, but with slightly larger variance.

[17] It is worth noting that the Jackknife IV estimator yields *very* high variance in this low-complexity DGP, as do Neural Networks.

$\hat{\beta}_1^{2SLS}$ into three components,

$$\text{Wedge} = \hat{\beta}_1^{2SLS} - \beta_1 = f\Big(\beta_1 \text{Cov}(\hat{x}, e),\ \text{Cov}(\hat{x}, u),\ \frac{1}{\text{Var}(\hat{x})}\Big), \tag{1.5}$$

where f is non-decreasing with respect to each of its arguments. Each component of the wedge offers insights into how first-stage methods differentially produce biases—and delivers helpful intuition regarding the pitfalls that may arise in 2SLS applications that include ML-based first stages.

To see the component parts of the bias drawing $\hat{\beta}_1^{2SLS}$ away from β_1, suppose again that the parameter of interest is β_1—the causal effect of x on y in

$$y = \beta_0 + \beta_1 x + u\ . \tag{1.6}$$

Suppose also that x is endogenous, i.e.Cov$(x,\ u)\ \neq\ 0$. The 2SLS estimate of β_1 comes from estimating

$$y = \beta_0 + \beta_1 \hat{x} + w\ , \tag{1.7}$$

where $\hat{x}$ is the first-stage-based prediction of x from some set of valid instruments $\mathbf{z} = z_1,\ z_2,\ \ldots,\ z_p$.

Because we estimate the second stage in (**??**) via OLS, the estimate for β_1 can be written

$$\hat{\beta}_1{}^{2SLS} = \beta_1 + \frac{\text{Cov}(\hat{x}, w)}{\text{Var}(\hat{x})}\ . \tag{1.8}$$

Using (**??**) and (**??**), we can rewrite w as

$$\begin{aligned} w &= y - (\beta_0 + \beta_1 \hat{x}) \\ &= \beta_0 + \beta_1 x + u - \beta_0 - \beta_1 \hat{x} \\ &= \beta_1 (x - \hat{x}) + u \\ &= \beta_1 e + u \ , \end{aligned} \tag{1.9}$$

where e is the first-stage residual—the difference between x and $\hat{x}$.

Using (**??**) for w, we can decompose the covariance in (**??**) into two components:

$$\begin{aligned} \text{Cov}(\hat{x}, w) &= \text{Cov}(\hat{x}, \beta_1 e + u) \\ &= \beta_1 \text{Cov}(\hat{x}, e) + \text{Cov}(\hat{x}, u) \ . \end{aligned} \tag{1.10}$$

If the first-stage predictions ($\hat{x}$) come from OLS, then $\text{Cov}(\hat{x}, e)$ is mechanically zero. The second term, $\text{Cov}(\hat{x}, u)$, is typically small when $\hat{x}$ comes from a linear-combination of valid instruments.

Finally, substituting (**??**) into (**??**) yields a helpful expression for the 2SLS estimate for β_1, which we can write as

$$\hat{\beta}^{\text{2SLS}} = \beta_1 + \frac{\beta_1 \text{Cov}(\hat{x}, e) + \text{Cov}(\hat{x}, u)}{\text{Var}(\hat{x})} \ . \tag{1.11}$$

OLS guarantees that $\text{Cov}(\hat{x}, e)$ is zero and, with valid instruments, that $\text{Cov}(\hat{x}, u)$ is small. Whether $\text{Var}(\hat{x})$ is “small” is typically of little consequence with OLS (as $\beta_1 \text{Cov}(\hat{x}, e) + \text{Cov}(\hat{x}, u)$ is typically small). However, all three points can generate important issues when we mix ML methods into the first stage of 2SLS.

With ML methods, nothing guarantees that $\text{Cov}(\hat{x}, e)$ is zero or that $\text{Cov}(\hat{x}, u)$ is small. Moreover, many ML methods are constructed to *reduce* the variance of predictions—further amplifying bias. This variance-reduction aspect is particularly relevant for nonlinear methods.

1.5.2.1 The $\beta_1\text{Cov}(\hat{x}, e)$ component. For the term $\beta_1\text{Cov}(\hat{x}, e)$ to differ from zero and generate bias, $\beta_1 \neq 0$ and $\text{Cov}(\hat{x}, e) \neq 0$. We assume that the population-regression coefficient β_1 differs from zero.[18] With this assumption imposed, the term $\beta_1\text{Cov}(\hat{x}, e)$ only generates bias when $\text{Cov}(\hat{x}, e) \neq 0$; β_1 scales the bias and affects its direction.

By construction, OLS produces predictions that are orthogonal to their residuals, i.e.$\text{Cov}(\hat{x}, e) = 0$. This first term is therefore irrelevant when the first stage uses OLS. However, when practitioners adopt other methods in the first stage (e.g.non-linear methods) nothing guarantees first-stage predictions are uncorrelated with their residuals. Put differently, this part of the bias results from using estimators whose predictions correlate with their residuals (rather than resulting from a violation of the exclusion restriction). While it is possible for nonlinear methods to generate $\text{Cov}(\hat{x}, e) = 0$, many do not.

In addition, because $\text{Cov}(\hat{x}, e)$ typically drops out of OLS regression, OLS-based empirical intuition does not help here. One implication of this non-OLS intuition of $\text{Cov}(\hat{x}, e)$ is that the bias generated by it is proportional to the size of the target parameter β_1. Where treatment effects are larger, the bias transmitted through this component is also larger.

---

[18] The case where it exactly equals zero is a measure-zero event that is uninteresting to the researcher.

To understand why some methods produce larger values of $\text{Cov}(\hat{x}, e)$ than other methods, first decompose this covariance into $\text{Cov}(\hat{x}, x)$ and $\text{Var}(\hat{x})$:

$$\text{Cov}(\hat{x}, e) = \text{Cov}(\hat{x},\ x - \hat{x}) = \text{Cov}(\hat{x}, x) - \text{Var}(\hat{x})\ . \tag{1.12}$$

While $\text{Cov}(\hat{x}, e)$ is not generally signable, its central component (i.e.the covariance between $\hat{x}$ and e, $\text{Cov}(\hat{x}, e)$) is bounded between $-\text{Var}(\hat{x})$ and $\text{Cov}(\hat{x}, x)$.[19] In addition, we can sign $\beta_1 \text{Cov}(\hat{x}, e)$ in fairly general subcases:[20]

$$\begin{aligned}
\text{Sign}\Big\{\ \beta_1 \text{Cov}(\hat{x}, e)\ \Big\} &= \text{Sign}\Big\{\ \beta_1 \text{Corr}(\hat{x}, e)\ \Big\} \\
&= \text{Sign}\Big\{\ \beta_1 \sigma_e^{-1} \Big(\text{Corr}(\hat{x}, x)\, \sigma_x - \sigma_{\hat{x}}\Big) \Big\} \\
&= \text{Sign}\Big\{\ \beta_1\ \Big\} \cdot \text{Sign}\Big\{\ \text{Corr}(\hat{x}, x)\, \sigma_x - \sigma_{\hat{x}}\ \Big\} \\
&= \begin{cases}
(+) & \text{if } \beta_1 > 0 \text{ and } \text{Corr}(\hat{x}, x)\, \sigma_x > \sigma_{\hat{x}} \\
(-) & \text{if } \beta_1 > 0 \text{ and } \text{Corr}(\hat{x}, x)\, \sigma_x < \sigma_{\hat{x}} \\
(-) & \text{if } \beta_1 < 0 \text{ and } \text{Corr}(\hat{x}, x)\, \sigma_x > \sigma_{\hat{x}} \\
(+) & \text{if } \beta_1 < 0 \text{ and } \text{Corr}(\hat{x}, x)\, \sigma_x < \sigma_{\hat{x}} \\
0 & \text{if } \beta_1 = 0 \text{ or } \text{Corr}(\hat{x}, x)\, \sigma_x = \sigma_{\hat{x}}\ ,
\end{cases}
\end{aligned} \tag{1.13}$$

where σ_x refers to the standard deviation of x ($\sigma_{\hat{x}}$ and σ_e are defined similarly).

As (**??**) reveals, the sign of $\text{Cov}(\hat{x}, e)$ depends on two quantities: (*i*) the sign of β_1, and (*ii*) the sign of $\text{Corr}(\hat{x}, x)\, \sigma_x - \sigma_{\hat{x}}$. It is difficult to generalize the sign of $\text{Cov}(\hat{x}, e)$ without further assumptions. While one may be tempted to assume $\sigma_x > \sigma_{\hat{x}}$, this assumption is not sufficient for signing $\text{Cov}(\hat{x}, e)$, as it still depends

---

[19] We assume estimates, $\hat{x}$, will have non-negative covariance with the true values, x.

[20] We assume e, x, and $\hat{x}$ have variation and that the predictions $\hat{x}$ positively correlate with the true values x.

upon the magnitude of $\text{Corr}(\hat{x}, x)$.[21] The knife-edge case where $a = 0$ appears unlikely except in cases where either $\beta_1 = 0$ or where $\text{Cov}(\hat{x}, e)$ is mechanically zero (e.g.OLS).

Across the twelve models that we consider in Table A2, only the non-OLS models produce $\text{Cov}(\hat{x}, e) \neq 0$ (it is mechanically zero for OLS-based models)—this is unsurprising. Lasso, neural nets, boosted trees, and random forests all produce positive covariance between $\hat{x}$ and e. In other words, in all of our DGPs, the term $\text{Cov}(\hat{x}, e)$ biases $\hat{\beta}$ upward (positively) whenever it is non-zero.[22] Random forest models generate the largest covariance between $\hat{x}$ and e (and consequently the largest $\text{Cov}(\hat{x}, e)$) in each of the DGPs. Depending upon the DGP, Lasso, neural nets, and boosted trees generate the second-highest covariance. Because our *shallow* subcase of neural nets approximates OLS, its covariance between $\hat{x}$ and e is approximately zero.

One way to ensure that $\text{Cov}(\hat{x}, e) = 0$ for a nonlinear model is to linearize its output—e.g.by using the ML-based prediction $\hat{x}(\mathbf{z})$ as an *instrument* for x, rather than plugging it into the second stage (J. D. Angrist & Krueger, 2001; J. Chen et al., 2020). While this approach forces $\text{Cov}(\hat{x}, e) = 0$, it requires strengthening assumptions on $\text{Cov}(\hat{x}, u)$ (as we discuss in Section 1.6).

More broadly, the component of bias due to covariance between first stage predictions ($\hat{x}$) and their residuals (e)—the $\text{Cov}(\hat{x}, e)$ term—accounts for the vast majority of the bias for Lasso and substantial amounts of the bias in random forests, boosted trees, and neural nets (the exact portion of the bias differs across

---

[21] Further, this assumption is equivalent to making an assumption on $\text{Cov}(\hat{x}, e)$, which means one is essentially assuming the result. That is, $\text{Var}(x) = \text{Var}(\hat{x}) + \text{Var}(e) + 2\,\text{Cov}(\hat{x}, e)$. That said, in every iteration of our simulations, $\text{Var}(x) > \text{Var}(\hat{x})$.

[22] This upward bias is partly due to the true parameter β_1 being positive.

DGPs and iterations). While $\mathrm{Cov}(\hat{x}, e)$ does not account for all of the bias, the non-zero covariance between first-stage predictions and residuals is an important (potentially large) component of the bias of ML-based 2SLS models.

1.5.2.2 *The* $\mathrm{Cov}(\hat{x}, u)$ *component.*

Unlike $\mathrm{Cov}(\hat{x}, e)$, the second component of the wedge between $\hat{\beta}_1^{\mathrm{2SLS}}$ and β_1 can be non-zero for both OLS-based methods and non-OLS models.[23] However, methods that use *non-linear* predictions of x in the first stage (i.e.ML-assisted 2SLS) require special care to produce low-bias estimates of β_1.

This second term, $\mathrm{Cov}(\hat{x}, u)$, is effectively the exclusion restriction, and any 2SLS-inspired estimator can reduce bias in $\hat{\beta_1}$ by ensuring $\mathrm{Cov}(\hat{x}, u)$ is approximately zero. Doing so under a machine learned estimator in-sample requires additional effort, even in the linearized case described above which escapes the ’forbidden regression’ trap. Assuming the instruments **z** are valid, an arbitrary prediction algorithm can maintain $\mathrm{Cov}(\hat{x}, u) \approx 0$ through either of three conditions:

1. **Restrict the algorithm’s choice set:** By restricting the learning algorithm to choosing from a set/class of functions where each individual function satisfies the exclusion restriction, one mechanically ensures the first-stage predictions $\hat{x}$ do not covary with the unobserved disturbance u. For example, when we employ OLS in the first stage of 2SLS, the first-stage regression is chosen from a set of class of linear functions that all include valid exclusion restrictions—linear combinations of the exogenous instruments (all linear combinations of exogenous instruments are themselves exogenous).

---

[23] For example, a mis-specified OLS regression or any 2SLS regression. While 2SLS is a biased estimator ($Cov(\hat{x}, u) \neq 0$)*, its bias increases substantially when instruments are not exogenous* ($Cov(z, u) \neq 0$) *or not relevant* ($Cov(z, x) \approx 0$)*.*

2. **Extend the exclusion restriction:** One may extend the assumption underlying the exclusion restriction into a much stronger assumption. Rather than only assuming all *linear* combinations of the valid instruments are (which is directly implied by instruments' validity), one could assume that **all** functions of the instruments—nonlinear and linear—satisfy the exclusion restriction. Put differently, this condition requires $\text{Cov}(f(\mathbf{z}), u) \approx 0$ for all functions f.

3. **Lean *very* hard on the ML algorithm:** The final option is to simply rely upon the algorithm to find a function that satisfies the exclusion restriction, irrespective of choice set—something akin to closing one's eyes and hoping for the best. While this option makes a rather heroic assumption, as ML algorithms are typically not designed to search for and find valid exclusion restrictions, it is the default scenario. If a practitioner does not enforce condition **1** and does not assume condition **2**, then she is left with **3**—i.e.hoping that the ML methods successfully choose a function that includes a valid exclusion restriction.

In summary, sufficiently flexible learning algorithms can recover endogenous variation in x *only using only valid instruments*. Importantly, many ML training methods explicitly incentivize and enable algorithms to do this.

Column b of Table A2 documents the tendency of flexible first-stage models (e.g.tree methods and neural nets) to recover endogeneity. As the learning algorithms permit more flexibility, $\text{Cov}(\hat{x}, u)$ tends to increase (across all DGPs). This covariance and its associated bias are particularly large for tree-based methods (especially random forests) and neural nets with multiple hidden layers. Notably, in panels b–d of Figure A2, the densities of unrestricted and narrow neural networks are bimodal. As Appendix Figure A6 illustrates, the bimodality results from whether the neural network (i) "chooses" zero hidden layers (the less biased mode) or (ii) goes deeper (learning the endogenous error and generating more bias).[24,25] This covariance between predictions $\hat{x}$ and the unobserved disturbance u accounts for a substantial amount of the bias in nonlinear methods—demonstrating that the previously discussed first component $(1 = \text{Cov}(\hat{x}, e) \neq 0)$ is not the only issue facing these models.

1.5.2.3 The $\frac{1}{\text{Var}(\hat{x})}$ component. While the first two bias components enter additively, the third component scales their sum. Any method that reduces the variance of the first-stage predictions—reduces $\text{Var}(\hat{x})$—mechanically inflates the bias produced by $\beta_1 \text{Cov}(\hat{x}, e) + \text{Cov}(\hat{x}, u)$.

In the case of properly specified, OLS-based, 2SLS, the variance of the predictions hardly affects bias in β_1, since $\text{Cov}(\hat{x}, e) = 0$ and $\text{Cov}(\hat{x}, u) \approx 0$. However, as most ML algorithms implicitly reduce the variance of their predictions to optimally trade between out-of-sample bias and variance. This tradeoff between bias and variance happens *outside of a 2SLS framework*—when practitioners infuse variance-reducing ML methods into 2SLS, the variance reduction actually amplifies bias in the second-stage estimates.

---

[24] This result highlights the importance of allowing neural networks to choose no hidden layers.

[25] Another, related, concern familiar to the ML literature is *overfit*. Overfit models tend to produce larger values of $\text{Cov}(\hat{x}, u)$ than models that have been cross-validated. Though cross-validation is best/standard practice for machine-learning methods in prediction problems, here it retains importance by preventing the algorithms from overfitting the target variable x in the first stage (even when out-of-sample performance is no longer the goal). We use five-fold cross-validation (CV) to tune the hyperparameters for Lasso-, tree-, and neural-net-based methods. Our neural-net cross-validation departs from standard five-fold CV. In Appendix Section E.1.2 we detail our cross-validation process for training neural net. One might further avoid overfit by applying holdout-style methods—only generating predictions for observation i when i is not in the training set. JIVE, split-sample IV, and J. Chen et al. (2020) all feature this additional safeguard. We do not employ these holdout-based methods because our goal in this paper is to simulate the results of a researcher using off-the-shelf ML tools in the first stage of 2SLS.

Taking these insights to the results in Panel B of Table A2, notice that variance reduction can cause methods to perform poorly. For example, Lasso-assisted 2SLS produces the lowest variance $\hat{x}$ in two of the three high-complexity cases—in cases 2 and 3, Lasso-based 2SLS has the highest $1/\text{Var}\hat{x}$-based amplifier of the bias. This high degree of bias amplification generates notable bias in Lasso, relative to many other methods (evident in Figure A2). So while Lasso's $\text{Cov}(\hat{x}, e)$ and $\text{Cov}(\hat{x}, u)$ are less than or equal to those of many other methods, the amplification produced by variance-reduction in $\hat{x}$ ultimately causes Lasso to have substantial bias. Notably, post-Lasso-based 2SLS produces less bias, partly due to the fact that it includes less variance reduction.

Worse yet, tree-based methods substantially reduce variance *and* produce relatively large $\text{Cov}(\hat{x}, e)$ and $\text{Cov}(\hat{x}, u)$ components—resulting in very large bias in their parameter estimates (even larger than naïve OLS).

1.6 Potential solutions

This paper examines the implications of plugging off-the-shelf ML methods into a 2SLS framework. In many cases, injecting ML into the first stage of 2SLS generates substantial bias.

While there are many approaches to combining instrumental-variable intuition and machine learning, they relax the traditional 2SLS structure and require different, generally stronger, identifying assumptions.[26] Among the current options, the closest in spirit to our question of "What are the implications

---

[26] See, for examples, MLSS (J. Chen et al., 2020), DeepIV (Hartford et al., 2017), DeepGMM (Bennett et al., 2020), KIV (Singh et al., 2019), Adversarial Estimation of Riesz Representers (Chernozhukov, Newey, Singh, & Syrgkanis, 2020), Neural Estimation of SEM (Liao et al., 2020), and Non-Parametric IV (Kilbertus, Kusner, & Silva, 2020).

of inserting ML into 2SLS?" is the "machine learning split-sample" (MLSS) estimator proposed by J. Chen et al. (2020).[27]

With two fairly simple expansions of the traditional 2SLS framework, MLSS mitigates *most* biases generated by naïvely plugging ML methods into the first stage. However, as with other more ML-forward methods, the solution is not without the cost of substantially strengthening the exclusion restriction. Specifically, J. Chen et al. (2020) proposes augmenting 2SLS with two simple techniques: restrict ML-based predictions to be explicitly out of sample (using split-sample methods), and use the ML-generated predictions as a "synthetic" instrument that then enters linearly in the first stage.

The idea for out-of-sample (split-sample) ML predictions follows the lead of Jackknife IV and Split Sample IV. By introducing out-of-sample methods to the ML-prediction exercise, J. Chen et al. aims to prevent the ML algorithm from fitting the first-stage errors—shutting down the bias generated by $\text{Cov}(\hat{x}, u)$. The drawback, however, is that this out-of-sample step likely increases variability (as seen in the JIVE results of Figure A2).

The second component of J. Chen et al. involves a "zeroth stage" (i.e., before the first stage), in which the practitioner trains an ML algorithm to predict x using the instruments $\mathbf{z}$.[28] The predictions from this zeroth stage are then used as the instrument within a traditional 2SLS framework. The benefit here is that the resulting linear first stage—linearizing the results of a potentially forbidden regression—guarantees that $\text{Cov}(\hat{x}, e) = 0$, shutting down one avenue

---

[27] J. Angrist and Frandsen (2020) also applies split-sample methods to several ML algorithms (i.e.post-Lasso, random forest)—both in the first stage of 2SLS and while synthesizing instruments in a stage that precedes the first stage.

[28] Note that this zeroth stage is identical to first stages that naïvely insert ML methods into 2SLS.

in which bias enters. Importantly, this method assumes that no learnable function of instruments meaningfully predicts the structural disturbance u. As the ML algorithm's function space grows (trades bias for variance), it can cover all possible functions of the instruments (e.g.most neural networks are universal approximators), which requires strengthening the exclusion restriction $\text{Cov}(\hat{x}, u)$ from any linear function of the instruments to one that includes all possible functions of the instruments. This strengthened exclusion restriction is generally much stronger (and likely more difficult to justify) than the typical identifying assumption assumed in 2SLS applications. As a simple example, most machine learning algorithms excel at detecting *interactions* of predictor variables that produce powerful variation in their predictions.

However, many econometricians fail to consider multi-way interactions in their exclusion restriction set, which ML methods will detect and produce bias in the resulting estimates, even for interactions up to the fourth degree, as seen in A3.

CHAPTER II

REVISITING THE STRATIFIED COST INDEX

2.1 Ch.2 Introduction

For much of modern history, wildfires have been the most common natural disasters, and have been increasing in salience for the greater population. United States suppression costs have increased alarmingly—while no US wildfire season prior to the year 2000 recorded suppression costs of over a billion dollars, 16 of the last 20 fire seasons have exceeded that threshold. This startling and anthropologically recent Marlon et al. (2012) change downplays the severity of the trend, as the last five years of suppression have, on average, cost two point three billion dollars *Suppression costs* (2020). Because of this rapidly rising cost, fire suppression costs and its underlying causes have become increasingly of interest for policy makers. Unfortunately, investigating why one wildfire is more expensive than another has proven to be quite challenging for economists to identify. The assumptions required for modern causal inference methods—which rely on random variation in treatment to produce causally valid estimates, are likely violated in the case of wildfire suppression costs. Natural experiments in the field are difficult to find,[1] as wildfires usually do not last long enough for a the pre-treatment period to be meaningfully removed from the post, preventing difference methods from removing fire-specific effects from any sample. Additionally, relying on random ignitions for causally valid variation is unlikely to be a winning strategy either. The works that do consider wildfire ignitions to be considered to be random events, must carefully trim their dataset, limiting the validity of the conclusions to sometimes as narrow as a single wildfire

---

[1] and experiments are ethically challenging to run

.[2] This means most externally-valid identification of causal factors underlying wildfire suppression costs rely heavily on conditionally as-good-as-random variation in observational settings.

To find causal estimates of parameters in such frameworks, researchers usually turn to techniques such as propensity score matching or conditional outcome models, which attempt to estimate $P(Y|z, X)$ for treatment z and outcome Y by directly controlling for variables in X. The problem with such methods is that they require a pre-existing structural causal model to select X, and rely on two strong assumptions for a continuous treatment variable z [3]:

$$\text{Conditional Independence} : (Y_{counterfactual}) \perp\!\!\!\perp z|X, \ \forall z$$

$$\text{Common Support} : 0 \leq P(z|X) < 1, \forall z$$

When the above assumptions appear to be satisfied, there are still ample risks to using this strategy for identification, as improperly defining the set X can mislead a researcher into interpreting effects of z on Y as causal. In particular, the wildfire management literature interchangably uses property value, count of homes and property presence in measuring the effect of homes on suppression costs. The argument for this modeling choice is usually that; under an optimal wildfire suppression strategy, a central planner should spend resources proportional to the value of the property at risk. Property value, home count and property presence variables, however, are not interchangable

---

[2] This is because the causality for suppression expenditure on many fires is confounded when the burn-paths of the wildfires depend on existing vegetation or longer-term trends in weather, both of which correlate to variables that may influence local policies and fire manager decisions.

[3] See Cunningham (2021) for a full overview

in consequence. In particular, the often cited research in Gude, Jones, Rasker, and Greenwood (2013), who use a home-count variable in their work, build on models developed in Gebert, Calkin, and Yoder (2007), which relies on evidence from conditional outcomes for causal interpretation. This choice for modeling is not of no consequence for policy application, as many of the proposed solutions for adaptation to rising suppression costs, such as restricting new construction in at-risk wildland urban interface communities, have differing impacts on home counts vs. property values vs. property location, thus understanding of the causal structure of fire suppression effort is of utmost importance to guide future policy decisions. Complicating the estimate of the causal effect, property value → suppression costs is a set of high-dimensional, spatially-varying environmental amenity confounders. While past work has used both spatial and point-level wildfire data to estimate how property value influences suppression decisions, estimates using fire boundaries or final burned acres leave a great deal of spatial variation available for potential confounding, and requires controlling for outcomes of fire suppression itself. This research seeks to validate this assumption in the literature, by refining a structural model of wildfire suppression costs that combines findings from the economics hedonic regression literature and wildfire research to create a unified causal model of wildfire suppression costs. With this model, the work then compares the resulting econometric estimates of suppression cost elasticity in regards to those derived by past studies.

Assuming as-good-as-random conditional variation admits a structural causal model of suppression costs. Doing so, however, consists of adjusting for a high-dimensional set of controls that reasonable encompass all of the spatial variation in wildfire and amenity sets, such as topology, fuels and weather and uses machine learning rather than conditioning the data with explicit fire boundaries to reduce the dimension of the features. This allows for identifying the causal effect of increasing property values on suppression costs, while controlling for environmental variables that cause wildfire suppression costs and property amenities to vary endogenously.

To the credit of researchers in ecological and economic sciences, the causes of total escalating expenditures are well developed. A century-old legacy of rapid response suppression and focus on a conservation of natural resources has led to an overabundance of plant life per acre when compared to historically observed vegetation patterns Busenberg (2004). At the same time, human-caused climate change has put pressure on forest systems from another direction. Recent work has estimated that climate change has led to a doubling of the total expected burn area of wildland fire since the 1980s, alongside increases in burn intensity Abatzoglou, Balch, Bradley, and Kolden (2018). Given the ongoing concern, there is a sizable economic literature in estimation of wildfire suppression costs, as seen in Donovan, Noordijk, and Radeloff (2004), Liang, Calkin, Gebert, Venn, and Silverstein (2008), Abt, Prestemon, and Gebert (2009), Preisler, Westerling, Gebert, Munoz-Arriola, and Holmes (2011), Yoder and Gebert (2012), Hand, Thompson, and Calkin (2016), Florec, Thompson, and y Silva (2019) and Baylis and Boomhower (2019). Despite this large body of work, Gebert et al. (2007), which introduced the *Stratified Cost Index* (SCI) remains the reference for fire cost estimation - driven in part by its adoption as a metric used to measure relative cost effectiveness of suppression efforts. The SCI; produced using a simple log-log ols regression estimate with ignition-point-level covariates, estimates suppression costs per acre as an ex-ante problem solved by fire managers.[4] As such, it is often cited by economists performing ongoing research on factors involved in wildfire suppression costs. Using measures of fire environment, values at risk, delay of detection, initial suppression strategy and availability of resources, it was able to explain roughly half of the in-sample variation[5] for logged large-fire per-acre costs across the country. As it was intended primarily as a solution to the forest service's cost forecast problem, many of the included factors are not rooted in understanding their effects from a causal standpoint—rather, they were intended to capture as much variation in the data as possible for use as historic benchmarks. Given there was no forest service metrics for tracking expenditure-to-performance, developing some way of tracking expenditure between divisions and across different fires is critical. However, given recent policy advice from economists and forest service professionals a more nuanced understanding of how fire managers change their assignment of resources in response to higher property values is critical.

The inclusion of 'total housing value'; an artifact of attempting to control for suppression effort spent on preventing loss of life and property,[6] has had a particularly lasting effect on the literature and fire suppression in the intervening years since the study. While count of nearby homes has been demonstrated to

---

[4] Yoder and Gebert (2012) updates this model with a maximum likelihood estimation that accounts for variation in fire acres over time.

[5] R^2 of .44 for Western US, R^2 of .49 for Eastern US. Out of sample predictions had R^2 of .33 for the Western Region and .18 for the Eastern Region. Restricted to Fires of Acreage 100 or greater, and also focused within Forest Service Lands - a caveat that was introduced due to data availability at the time.

[6] Other factors considered but discarded due to poorer fit were nearby population and total count of homes

have a causal effect on fire suppression costs Gude et al. (2013), most economic models of suppression expenditure treat total nearby property value as the input to the fire manager's problem, but empirical work has not been able to disentangle this causal link. Since the publication of Gebert et al. (2007) distributional concerns of suppression expenditures in the continental United States have been raised by the literature. Evidence shows that spending on wildfire suppression disproportionately benefits individuals in the wildland-urban interface (WUI) Baylis and Boomhower (2019), who are on the whole whiter and wealthier than populations in other regions. This has been reinforced in preliminary results from the working paper Wibbenmeyer and Robertson (2021) with nearly universal results for the Western United States, save Washington. One overlooked avenue complicates interpretations of this result—algorithmic bias in the SCI itself, by way of the fire-manager's information set.

Wildfires are extremely complex events, where incident managers are forced to make military-scale decisions with lives in the balance—even so, an incredible amount of consideration is made towards optimizing choices in a careful way. Home value directly or indirectly correlates to fire manager decisions and suppression expenditure in four distinct ways-

- **The "Direct" Path:** In a very real sense, allocating resources to proportionately protect expensive properties is a logical policy objective: if a manager seeks to minimize total dollar losses, spending more money to prevent high-value property loss relative to low-value property loss falls directly in line with an optimal decision rule.

- **The "Amenities" Path:** Property prices in and around the WUI are; more so than for other homes, a product of the portfolio of amenities that homeowners value and are willing to pay for. Amenities in hedonic analyses are usually assumed to be orthogonal to parameters of interest, but this is not the case here. For example, property views generally produce utility for the homeowner and are considered characteristics of the home. Views are functions of surrounding topography, property ownership of neighboring properties, nearby plant growth, and their interaction. Topography plays an important role in directly unrelated expected fire spread, as do 'fuels' which themselves are functions of plant growth. Expected fire spread is directly factored into suppression strategy which directly drives costs. This effect has been well discussed in the property valuation literature.
- **The "Procedural" Path:** Fundamentally, suppression costs are perfectly identifiable from inputs used for the suppression production function. Which inputs are optimal for this production function change, based on the characteristics of the fire. Inputs to this production function have distinct advantages and their utility is differentially impacted by topography and development. Even under conditions where fire managers allocate equal importance to home protection regardless of property value then more expensive inputs may be required to defend an expensive property, depending on where it is located.
- **The "Model Loop" Path:** One unconsidered route for the apparent bias for protection of higher incomes is the SCI itself. If a wildfire's per-acre expenses are 2 standard deviations above historic norms, responsible managers will very likely undergo a thorough review of their work, likely leading to

> many hours authoring government reports that explain any cost overrun. Given property value directly factors into some regional SCI models—any existing policy pathways that promote expenditure on expensive property are reinforced by having the historic expenditures respond to threatened property value.

All property value effects measured by studies so far have not tried to digest and understand the relative magnitudes from these different causal pathways[7] The challenge of doing this is that the specific functional form of the suppression production function is unknown, and the interaction between this function with spatially dynamic wildfire is likely to be extremely complex and non-linear. Existing research on the spatial dynamics of wildfire mostly includes some form of data about the final burned area of the fire either as a control as in Abt et al. (2009) or to dictate a study area, as in Hand et al. (2016). For understanding the causal effect of pre-determined variables such as property values on suppression costs, conditioning on post-treatment variables are bad controls. What is required for interpretable parameters in such a model is a function that takes a set of spatial variables, weighted by an either known or learned event-agnostic kernel to control for indirect effects. This should capture a sufficient degree of variation to meaningfully capture the relationship, as outlined in Marchal, Cumming, and McIntire (2017). Linear regression is not sufficient for such a task without some pre-identified spatial kernel that can combine the requisite number of inputs into a sufficiently small feature set. One solution is to treat the problem as a partially linear regression - that is, that the parameter of

---

[7] Ongoing research in the working paper, Plantinga, Walsh, and Wibbenmeyer (2021), has controlled for local fuel levels, which may partially block the indirect causal pathways.

interest has some objective linear effect on the outcome, but the controls enter in a highly nonlinear way to both the outcome and the causal variable of interest. One solution is to use newly developed machine learning algorithms to weight the spatially explicit data and then utilize double/debiased machine learning to understand the ex-ante effect of property value on suppression expenditures decisions.

To extend understanding of suppression, this work seeks to combine the climate and event-level causal models of wildfire and wildfire suppression expenditures, identify a sufficient conditioning set of variables using methods outlined in Pearl (2010) to block the non-direct causal pathways that may lead to spurious correlation between home prices and suppression costs, using newly developed vision transformers,[8] a computer vision technique that is capable of capturing both sequential and spatial knowledge with minimal inductive bias—allowing us to most completely identify complex nuisance parameters, while leaving the direct causal path unblocked.

The results have both theoretic and policy implications for wildfire suppression. When not conditioning on acres burned, per acre costs of suppression are less responsive to property value alone than past estimates imply and are indistinguishible from zero. This remains true whether or not estimates are conditioned on Western United States.

This work will proceed with a stylized model of suppression cost expenditures building off of work done in Bayham and Yoder (2020) in section 2.2, followed by a semi-parametric econometric model in section 2.3. The work then turns to existing understanding of wildfire spread, expected valuation of

---

[8] First proposed in Dosovitskiy et al. (2021)

damage/suppression and hedonic property modeling in section **??**. Section 2.4 sets up the assumptions for the modeling framework and outlines the machine learning and causal models in full, before moving on in section 2.5 to discuss the data used for the study before finally; in section 2.6, outlining and discussing initial results.

2.2 Model of Cost

Fire managers make decisions in inherently unsure environments, under less than ideal conditions. Using resources owned by a combination of federal agencies and private contractors, they are tasked with protecting resources they do not necessarily personally benefit from with limited or changing guidance in terms of which ought to be prioritized. In this sense, fire managers do not face a budget constraint unlike many other market actors, noted in Calkin, Venn, Wibbenmeyer, and Thompson (2013) though their employing agency must still pay these costs. Instead, there are constraints placed on them from two different sources: first, from the limiting number of resources present in the United States and abroad, and second, through limited monitoring by their cost-sensitive employer, via comparison to historical norms from a regionally-varying metric, the ’stratified cost index’. Such problems are common in the real world when an agent performs relatively independent work that is only monitored by a manager through a loose collection of weakly correlated signals.

To motivate the econometric estimation, this work begins with a stylized model of fire resource assignment to determine total suppression costs, as designed by Bayham and Yoder (2020) and extend it to include endogenous values at risk, as well as smaller fires. In the authors’ model, a set of fire managers communicate needs to a regional fire coordination group who choose allocations of resource bundles over many concurrently burning fires. A fire manager is treated as a loss minimizer making decisions at time t, and where losses are accumulated over two periods, $\{t, t+1\}$. Losses in a given period are represented by a non-parametric function

$$\ell_t(d_t, c_t)$$

(2.1)

This is done because costs spent to protect assets can exceed total avoided losses (Calkin et al. (2013).) Suppression costs are then simply generated by a vector of assigned inputs,

$$c_t = \mathbf{y_t'w} + m(\mathbf{y_t}, \mathbf{p_t})$$

(2.2)

Where $\mathbf{y_t}$ represents a $\{N \times 1\}$ binary vector of all resources and $\mathbf{w}$ represents a corresponding $\{N \times 1\}$ vector of pre-negotiated contract prices or wages, m captures operating costs generated by these resources, for instance, travel time from point of storage to fire front. $\mathbf{p}$ is a interconnected set of all environmental and geographic features, $\mathbf{p_t}$ are those nearby features (and their connections) relevant to the fire at the time of response and impacts operating costs by changing feasible staging locations, available water sources or other geographic assets that make fire resources more or less costly to use. These costs are used to change outcomes in a future damage function, d_{t+1},

$$d_{t+1} = d(\mathbf{v_t}(\phi_\mathbf{h}, s_t(\mathbf{p_t}) + \mathbf{e_t^v}), \mathbf{y_t}, s_{t+1}(\mathbf{p_{t+1}}, s_t(\mathbf{p_t}), \mathbf{y_t}) + \varepsilon^s_{t+1}) + \varepsilon^d_{t+1}$$

(2.3)

Where damage is a monotonically decreasing function of $\mathbf{y_t}$, an increasing or decreasing function of both fire shape, s_{t+1} and $\mathbf{v_t}(\phi_h, s_t(\mathbf{p_t}) + \mathbf{e_t^v})$, which represents a vector of potentially affected values, which itself is a function of nearby environmental amenities capitalized into the home ,[9] conditions at the ignition location $s_t(\mathbf{p_t})$ and market based shocks $\mathbf{e_t^v}$. This change to values at risk from Bayham and Yoder (2020) is done for two reasons: first, because the goal is to allow for endogenous values at risk, and second to extend the model to smaller fires—as small fires often bring beneficial changes to ecosystems and can reduce fire risk in the long run. The shape function is a function of ignition conditions and location, s_t, which itself comes from environmental conditions at the ignition location p_t. Both $\mathbf{p_t}$ and $\mathbf{p_{t+1}}$ are vectors of ϕ_t(fire front$_t$), and ω_t, where $\{\phi_t, \phi_{t+1}\}$ form an incomplete atlas of $\mathbf{p}$ at the location of the fire front in periods t and t+1, and a set of time-varying environmental factors like weather, in ω_t and ω_{t+1}. Fire shapes in future periods is uncertain (even with partially unobserved p_{t+1}), and so resulting damage from this uncertainty is captured in a mean zero noise variable,

---

[9] represented by a function $\phi_h(x_1, \ldots, x_n) \quad \to \quad \mathbb{R}$ defined by points on the set p: to function as a smooth surface of country-wide environmental variables

ε^{s}_{t+1}, and uncertainty in the overall damage function from resource variability is captured in ε^{d}_{t+1}.

The vector of at-risk values contains any homes at risk with varying property values, other private structures, public structures, infrastructure, natural resources such as watersheds or ecological values such as wildlife habitat. Fire suppression resources can protect these assets through a few mechanisms—one, by preventing the fire shape from reaching these assets, or two by applying effort to protecting individual resources. Environmental conditions at the front of the fire, p_{t+1} can have a strong impact on effectiveness of selected resources, where humidity can reduce the spread of fire Rothermel (1972) and steep/inaccessible terrain can make engines and firefighter efforts on foot substantially more challenging as noted in Katuwal, Calkin, and Hand (2016) and Butry, Gumpertz, and Genton (2008).

From here, with these changes to the prior work's response function, the analysis can follow in the footsteps of Bayham and Yoder (2020).

Fire managers minimize the loss function over two periods, subject to available resources at time t, $\mathbf{Y_t}$

$$\min_{y_t \geq 0}\{\ell_t(d_t, c_t) + E_t\{\ell_{t+1}(d_{t+1}, c_{t+1})\} : \ \ s.t\ \mathbf{Y_t} \geq \mathbf{y_t}\}$$

(2.4)

This makes the first order conditions of this problem become

$$\mathbf{FOC}_{\mathbf{y_t}} \equiv \frac{\partial \ell_t}{\partial c_t}\frac{\partial c_t}{\partial \mathbf{y_t}}\mathbf{w} + \frac{\partial \ell_t}{\partial c_t}(\frac{\partial c_t}{\partial m_t}\frac{\partial m_t}{\partial \mathbf{y_t}}) + E_t\left[\frac{\partial \ell_{t+1}}{\partial d_{t+1}} \quad \frac{\partial d_{t+1}}{\partial \mathbf{y_t}} + \frac{\partial d_{t+1}}{\partial s_{t+1}}\frac{\partial s_{t+1}}{\partial \mathbf{y_t}}\Bigg)\right] - \boldsymbol{\lambda}_\mathbf{t} = 0$$

(2.5)

$$\mathbf{y_t} \geq \mathbf{0},\ \mathbf{Y_t} \geq \mathbf{y_t},\ \mathbf{FOC}'_{\mathbf{y_t}}\mathbf{y_t} = \mathbf{0},\ \lambda_{n,t}\left[Y_{n,t} - y_{n,t}\right] = 0,\ 0 \leq \{y_{n,t},\ Y_{n,t}\} \leq 1 \tag{2.6}$$

where $\lambda_{n,t}$ is the Lagrangian multiplier for the specific resource n, $\boldsymbol{\lambda}_\mathbf{t}$ is an $\{N \times 1\}$ vector of such multipliers. [10] These first order conditions consist of four interpretable terms: the first represents is positive, and represents increasing cost of assigning a resource to a given fire due to increasing expenditures on contract fees and the second represents increasing expenditures due to practical considerations of mobilizing the resource - either due to their home-base or due to support costs (food, housing, fuel etc.) The third and fourth terms represent expectations in time t of damage reduction occurring in period $t + 1$, the third being the direct effect of suppression resources on damage (a direct protection measure) and the fourth is the avoided damage due to containment of the shape in period t+1: s_{t+1}. Importantly, the third and fourth terms may be positive or negative, given fire is sometimes ecologically beneficial and lowers future fire risk. This vector of first order conditions defines, as in the prior work, an equilibrium

---

[10] The discrete problem for langrangian optimization can be made into a continuous optimization following Wah and Wu (1999).

$\{\mathbf{y_t^*}(\mathbf{Y_t}; \mathbf{v_t}, \mathbf{p_t}, \mathbf{p_{t+1}}, \mathbf{w}), \lambda_t^*(\mathbf{Y_t}; \mathbf{v_t}, \mathbf{p_t}, \mathbf{p_{t+1}}, \mathbf{w})\}$ can be defined for each wildfire.[11] This includes all scenarios, and includes endogenous values at risk, $\mathbf{v_t}$ directly.

Observed costs after a fire, then, can be constructed by plugging in the definitions of optimal resource usage ($\mathbf{y_t^*}$) derived from **??** into **??** to get,

$$c_t^* = \mathbf{y_t^*}\mathbf{w} + m(\mathbf{y_t^*}, \mathbf{p_t}) \tag{2.7}$$

Where $\mathbf{y_t^*}$ is a vector of functions, where each depends on local fire conditions, wages, values at risk, and also fire conditions and values at risk for other fires burning at time t .[12]

Understanding the cost response of fire suppression costs to a specific value in $\mathbf{v}$ like property values is represented by

$$\frac{dc_t^*}{dv_{t,h}} = \frac{d\mathbf{y_t^*}'}{dv_{t,h}}\mathbf{w} + \frac{\partial m'}{\partial \mathbf{y_t^*}}\frac{d\mathbf{y_t^*}}{dv_{t,h}} \tag{2.8}$$

Equation **??** gives us an intuitive interpretation of the coefficient on a regression of costs on property values - the optimal assignment of resources as it responds to housing values times the wages, plus the responsiveness of support costs to those optimal choice of resources. However, we know y_t responds to changes in p_t and p_{t+1}, and $v_{t,h}$ is a function of p_h and p_t. This overlap in functional inputs will result in bias in any unadjusted econometric estimates of suppression cost elasticity. This also allows us to understand, in context of endogenous values at risk, past estimates of the contribution of housing value on suppression costs

---

[11] Bayham and Yoder (2020) use planning periods as t, but given many resources remain assigned to the fire, we can relax this to a 2-period per fire model.

[12] This comes from a regional fire manager who responds to several fires requesting resources and using the relative shadow prices $\lambda_{n,i,t}$ where i tracks fires.

using point data, as in Gebert et al. (2007) and using shape/total acreage, as in Liang et al. (2008) or Hand et al. (2016).

2.3 Econometric Model

Both property values and fire suppression are intensely intertwined with nearby natural features, and thus any unconditional estimate of the price elasticity of suppression to property values is likely to be biased. One way of understanding this dependence is by examining the simplest linear regression model of suppression costs on summed nearby property values.[13]

$$cost_i = \beta_0 + \beta_1 \sum_{n=1}^{N} propVal_i + \epsilon_i, \text{ for fire } i$$

(2.9)

Estimating this model, however, is problematic using OLS.

$$cost_i = \beta_0 + \beta_1 \sum_{n=1}^{N} (\phi_{n,h} + s_t(\mathbf{p_t}) + v_{n,t}) + \epsilon_i \tag{2.10}$$

Using this regression, we can find the estimated $\hat{\beta}_1$

$$\hat{\beta}_1 = \left[(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})'(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})\right]^{-1} (\phi_h + s_t(\mathbf{p_t}) + v_{n,t})' c_t \tag{2.11}$$

---

[13] For illustrative purposes, we will discuss this problem in terms of linear parameterization for each function - in practice, these functions may not be linear, and then we can extend to the nonlinear case.

$$\hat{\beta}_1 = [(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})'(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})]^{-1} (\phi_h + s_t(\mathbf{p_t}) + v_{n,t})'(\sum_j w_j(y^*_{j,t}) + \sum_j (\delta_j(y^*_{j,t}) + \gamma_j \mathbf{p_t})) \tag{2.12}$$

In the most straightforward case, we can produce the usual $\hat{\beta}_1 = \beta_1 + bias$ decomposition.

$$\beta_1 = \frac{(\phi_h + s_t(\mathbf{p_t}) + v_{n,t}) \sum_j (\frac{w_j + \delta_j}{\xi_j})(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})}{[(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})'(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})]} \tag{2.13}$$

$$bias\ in\ \hat{\beta}_1 = \frac{\left(\sum_j w_j(y^*_{j,t}) + \delta_j(y^*_{j,t}) + \gamma_j \mathbf{p_t} - \sum_j (\frac{w_j + \delta_j}{\xi_j})(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})\right)((\phi_h + s_t(\mathbf{p_t}) + v_{n,t}))}{[(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})'(\phi_h + s_t(\mathbf{p_t}) + v_{n,t})]} \tag{2.14}$$

Where $\xi_{j,i}$ is the inverse share of all weighted values at risk represented by private property. Assuming no bias in our estimator, assumes most problematically that environmental conditions at the point of ignition ($\mathbf{p_t}$) are uncorrelated with the ignition location ($s_t(\mathbf{p_t})$). The very earliest works attempting to predict wildfire costs understood this, and tried to adjust for $\mathbf{p_t}$ directly—as was done in Gebert et al. (2007) (the Stratified Cost Index). When implemented correctly, this changes equation **??** substantially.

$$bias\ in\ \hat{\beta}_1 | \mathbf{p_t} = \frac{\left(\sum_j w_j(y^*_{j,t}) + \delta_j(y^*_{j,t}) - \sum_j (\frac{w_j + \delta_j}{\xi_j})(\phi_h + v_{n,t})\right)((\phi_h + v_{n,t}))}{[(\phi_h + v_{n,t})'(\phi_h + v_{n,t})]} \tag{2.15}$$

This completely removes bias due to correlation between inclusion of homes from changing ignition location and environmental conditions at the point of ignition. Though such an approach goes a long ways towards a causal estimate of β_1, as is outlined in section 2.2[14] additional bias may arise from

---

[14] and recognized in Gebert et al. (2007)

optimal resource choices $\mathbf{y_t^*}$. Using the conditional functional form of $\mathbf{y_t^*}$,[15] this requires a strong assumption to drive equations **??** and **??** to 0,

$$w_j * E_t(y_j^*(\{\phi_{t+1}, \omega_{t+1}\}, v_{k,t})(\phi_h + v_{n,t})|\mathbf{p_t}) = 0, \forall\ j,\ k \neq n \tag{2.16}$$

In words, this assumption implicitly encompasses two logical leaps: first, that the environmental conditions at the head of a fire are uncorrelated with the environmental conditions that are capitalized into a private property ($cov_{\mathbf{p_t}}(\phi_{t+1}, \phi_h) = 0$), and second, that the ratio of costs assigned to different values at risk are independent from one another for any resource j. One way to avoid this problem is to control for observed factors of $\mathbf{p_{t+1}}$, most commonly, this is done by using observed environmental/topological conditions contained by the perimeter of the fire alongside observed weather conditions, as seen in Donovan et al. (2004) and Hand et al. (2016). This

This procedure, however, indirectly conditions data on the final state of the fire, s_{t+1}. As was outlined in equation **??**, this has the effect of conditioning on $\mathbf{p_{t+1}}$ and $\mathbf{p_t}$ as intended, but also conditions on utilized resource's effect on fire growth, $\mathbf{y_t^*}$. Such a conditioning statement does not matter when the parameter of interest is an optimal prediction, but, as observed in all three referenced studies doing so finds much smaller effects from property value on suppression costs.[16]

$$\beta_1 = \mathbf{w}\frac{\partial \mathbf{y_t^*}}{\partial v_{t,h}} + \frac{dm}{dv_{t,h}} \tag{2.17}$$

---

[15] $\mathbf{y_t^*}|_{\mathbf{p_t}}(\mathbf{Y_t}; \mathbf{v_t}, \mathbf{p_{t+1}}, \mathbf{w})$

[16] Though in Hand et al. (2016) they internally find the opposite- that point ignition results are smaller than spatially aligned results. This may be due to the dataset in use - which spans 2000-2008 and thus induces measurement error from homes valued pre and post great recession.

If using final fire boundaries and acreage as weights does not sufficiently condition on spatially varying data, the question then becomes, how should one estimate this causal effect and is any specification valid? Imagine the optimal estimator for the query - what is the elasticity of suppression costs in response to changes in property values, ceteris paribus? Ideally, we would like to control for $\{\mathbf{p_t}, \mathbf{p_{t+1}}\}$, but without using observed fire boundaries that are a function of the outcome we are hoping to measure. The standard econometric approach would be to look for an observable policy experiment that changes the property values of certain homes in a way that is orthogonal to the physical factors of fire suppression. However, such a policy experiment would also require two wildfires occurring at similar intensities during the same period of time that threatened both sets of homes. This makes many of the policy tools for natural experiments available to economists unlikely to be viable.

Alternatively - an econometrician could draw a sufficiently large boundary around the point of ignition, and use all relevant environmental variables within it to control for both the environmental conditions at the point of ignition and at all potential future fire-front locations, ie $\mathbf{p_b} \supseteq \{\mathbf{p_t}, \mathbf{p_{t+1}}\}$ but not condition on the future fire manager information set I_{t+1}. Doing so would result in an estimate of β_{cond},

$$\beta_{cond} = \mathbf{w}\frac{\partial \mathbf{E}(\mathbf{y_t^*}|p_b)}{\partial E(v_{t,h}|p_b)} + \frac{dm}{dE(v_{t,h}|p_b)} \tag{2.18}$$

Conditioning on this set of variables would neither condition on an outcome, nor would it require assuming exogeneity of property values and fire suppression conditions. Problematically, physical characteristics of a fire are extremely high dimensional, thus without having some method of weighting spatial variation (as point-level statistics do, or averaging within the burned area does) in nearby environmental factors, some degree of feature engineering or selection is required to produce an estimate. One such approach for doing exactly this, in the partial linear regression framework[17] is developed in Chetverikov et al. (2016), which permits machine learning methods to estimate valid controls in high dimensional settings for causal queries in a doubly robust way.

However, to meet sufficiency requirements, we must assume that not all of the variables in $\mathbf{p_b}$, $\mathbf{v_t}$, ω_t are relevant for understanding the causal query in question, and must use some consistent[18] algorithm to learn a control that would remove bias due to confounding. Doing so requires understanding how spatial variables can lead to changes in suppression expenditures. The next section will outline theoretical understanding of wildfire spread, expected costs and expected net value change and end with a discussion of results from the environmental economic hedonic literature.

2.4 Models of Wildfire Suppression and Hedonic Valuation

In order to produce a comprehensive system in which to conduct a causal analysis, it is important to understand institutional knowledge around each subcomponent of suppression - simulated physical models of spread provide an understanding of a fire-manager's expected damages they will try to avoid, which can be used alongside economic-based models of suppression effort. This is done to identify the minimal adjustment set for a causal analysis.

2.4.1 Wildfire spread.

---

[17] first described in Robinson (1988)

[18] in the mean square error rate sense

Wildfires are inherently unpredictable events, and take place in an evolving and complex environment where suppression strategies must be adopted prior to full knowledge of damages (or benefits) that may occur as a result. This means that in order to understand suppression decisions of fire managers, it is necessary to understand how expectations of wildland fire spread evolve and the resulting changes to assets in its perimeter, absent of suppression. Fire risk, like other amenities, is closely associated

Despite the uncertainty underlying wildfire events themselves, wildfire spread, conditional on what is burning is a fairly well understood problem. The trouble comes in simulating wildfire events which often cross heterogeneous fuel types and occur under rapidly changing real-world atmospheric conditions. Though these models are well defined on a small scale, validating a molecular-level simulated physics model scaled to a full fire system is nearly impossible to validate, leaving how to expand detailed small-scale understandings to larger-scaled phenomenon of interest. Understanding the physical processes that underpin models of fire spread, along with how those models are operationalized in the field provide insight into the forecasting problem fire managers face. This requires some background on both physical and computational models of fire spread. [19]

Existing models can be categorized into one of two categories, those that rely primarily on a series of vector calculations, at the cost of more complicated two dimensional dynamics,[20] and those that take inspiration from, but forego some of the formal physical processes in exchange for two dimensional spread.

---

[19] This analysis does not seek to reinvent the model of fire spread, but organize the existing field of knowledge. As such, it includes this section to motivate the future econometric analysis.

[20] models of this type make assumptions about width of fires

Whichever category of model is being used, the underlying dynamics of the spread are dictated by the well-known Rothermel (1972) model of fire spread, which was expanded upon soon after by Rothermel (1983). The goal is to simulate how fire spreads in a plane, with changing wind vectors and slopes on a small scale to hopefully be able to scale said model up to larger wildfire size events.

To illustrate this model, imagine simulating such a problem in absence of wind, moisture and slope, in a uniform fuel bed (ie, a patch of dried grass in a laboratory.) Dropping a match in the middle of the grass patch will lead to the fire burning the fuel (grass) at the center of the patch and spreading evenly to the outside edges at some rate. To find that rate, it's reasonable to measure the rate of spread as a ratio of the intensity of the reaction (how hot the fire is burning) over some function of density of fuel (as the fire will consume proximal fuels first before spreading to neighboring fuels) and ambient heat. Where I_R is the intensity of the burn, ρ_b is the density of the fuel, absent of moisture in $\frac{kg}{m^3}$, Q_{ig} is ambient heat, and ε is a scaleless 'effective heating number' - this spread is simply $\frac{I_R}{\rho_b \varepsilon_{ig}}$.

What makes this problem more complex is that fire spread is directly impacted by the slope on which the reaction is occurring and wind speed. This complicates the problem because both slope and windspeed are 'orientation sensitive', that is, given a spread direction of interest, they have a differing effect on the observed rate of spread.[21] In addition, many fires burn in tree-stands meaning a fire that begins by burning only underbrush can escalate in intensity and eventually what is known as a 'crown fire', where the tree canopy ignites in

---

[21] Rothermel's surface fire spread rate, which takes into account these factors, given a spread direction, is very similar to the equation written above. $R_{rothermel}$ can be calculated by using the following $R_{rothermel} = \frac{I_R \xi (1+\Theta_w+\Theta_s)}{\rho_b \varepsilon Q_{ig}}$ where ξ is the propagating flux ratio, and Θ_w and Θ_s are windspeed and slope respectively.

a tree stand. Crown fires complicate analysis of fire spread for three reasons - one, they occur at a different elevation and implicit slope than surface fires, and therefore spread at different rates two, fires burning further from the ground produce much higher rates of ’spotting’ – burning debris transported by wind and rising heat, that can seed ’spot fires’ in non-neighboring regions Albini (1979), and lastly, how crown fires have been modeled and understood is still limited, more research needs to be done to understand how different tree species and combinations of ground fuels change flame dynamics.

2.4.2 "Damages" from Wildfire. Understanding wildfire damage, as with understanding wildfire itself, is challenging. Even the term ’damages’ is misleading – wildfire can be, and often is, beneficial. This was an early learning of the US forest service when, in 1960, it was discovered that not a single giant sequoia tree had begun to grow in California since the turn of the century as a direct result of hyper-aggressive early fire suppression strategies. This has led to a changing of posture from one of damage-prevention to one of maximization of net-value-change, or NVC, conditional on ignition. The description of this change, and how it is included in policy to inform fire procedures is best described in Scott, Thompson, and Calkin (2013). This change of focus to NVC has led to increased uptake of the practice of ’prescribed fire’, where experienced fire managers intentionally light fires to clear fuel and repair ecosystems that are out of balance.

The complication this adds to analysis is that net value change is not only a function of binary burned and unburned cells, but rather a complex weighting of conditional damage plus benefit at every burn intensity level. Damage and benefit, however, do not just come from local burn sources – crown fires that produce spotting can lead to damage to WUI property from a considerable distance. They also can deposit higher densities of ash in local watersheds, poisoning communities despite being nowhere near a flame.

This means a fire-manager is required to assess the conditions of a fire, how likely the fire will enter more or less severe intensity states and how allowing a fire to burn in new locations may increase the expected likelihood of negative NVC in both adjacent and non-adjacent locations.

As discussed in Scott et al. (2013), and first directly estimated in Dillon (2020), wildfire NVC risk is modeled as a discrete probability distribution over damages to specific assets, as weighted by local decision makers within a decision-making body. Where w_g are locally determined weights of importance, BP is burn probability, NVC is net value change, RI_g is some metric of relative importance, RE is a metric of relative extent,[22] i indexes locations, and j indexes values at risk, risk is defined as -

$$Risk \equiv E(w_g NVC) = \sum_j \sum_i (BP_i * NVC_{ij} * \frac{RI_g}{RE}) \tag{2.19}$$

However, fire managers rarely respond to the unconditional risk function during a suppression action. They interact instead with a version of the above function, conditional on specific time and ignition. Grouping these initial conditions into variable ξ, the conditional version of the above looks like -

$$E(Risk|\xi) \equiv E(w_g NVC|\xi) = \int_\xi \sum_j \sum_i (BP_{i,\xi} * NVC_{i,j,\xi} * \frac{RI_{g,\xi}}{RE_{i,\xi}})$$

---

[22] This is a unit-agnostic spatial measure constituting the area the value at risk occupies - this is done to prevent overweighting certain amenities because they occupy a larger area than others, see Scott et al. (2013) for more information. In this work, the units for this measure will be grid cells, ie, 90 m^2

Where $BP = P(i|\xi)$, $NVC = P(NVC_{i,j}|\xi, i)$

This minor change allows for a reformulation of burn probability in cell c, to

$$BP_i \equiv \prod_{c=1}^{c=C} P(BP_c|pa(BP_c, \xi), \xi)$$

.

This reformulation allows for a connection between the damages from wildfire and spread. Spread of fire fronts can be modeled as a natural process that potentially spreads between cells in a Moore neighborhood on a grid, as shown in figure **??**, with burn probability driven by $R_{rothermel}$.[23] Damage arises from the fire intensity, and thus expected NVC from an action that lowers the burn probability of cell c is

$$\Delta BP_c * w_g NVC_c * \frac{RI_g}{RE} + E(Risk_{-c}|\Delta BP_c)$$

Where risk is defined as seen in equation **??**. Clearly, the expected NVC of any fire suppression action must depend in part on variables that dictate fire spread, local priorities as defined by the management region and the set of values at risk during an incident. Thus, optimal suppression effort must also depend on a combination of the variables driving expected NVC, alongside variables that

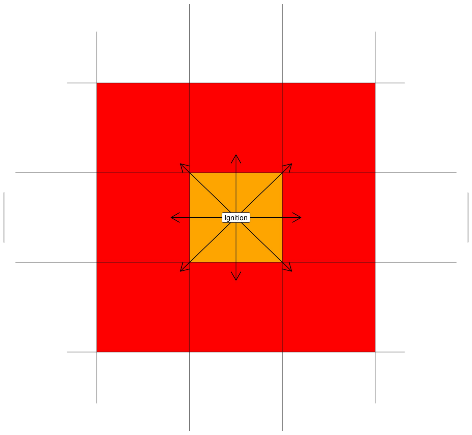


Figure 1. Visualization of the Moore neighborhood

Cells in orange are burning, those in red are potentially at risk of igniting in period 2

---

[23] This notably ignores spotting- but spotting can be incorporated by including canopy information in the cell and creating a more complex interconnection between cells

change the cost of that action, such as contracted price for the resource and how many hours are required to achieve that outcome, which is a function of both the type of resource being used and accessibility of the cell.

2.4.3 Property Valuation and Environmental Amenities.

'Hedonic' valuation is a technique to estimate how product 'quality' is valued by consumers (in a revealed preference way) in any good, and is estimated by examining how individual contributions of specific characteristics of some complex good contribute to the whole of price. In economics, this approach was first popularized as a mechanism to measure of changes in quality for automobiles in Griliches (1961), but soon was adapted to be used to estimate factors in hedonic cost models in Kain and Quigley (1970). The set up for such a model is modeled as a regression between values or monthly rent on individual characteristics of the home $\mathbf{x}_i$, including location, size, school quality etc. as well as changes in value arising over time, t. Generally, this method attempts to identify utility changes from marginal changes in some element $x_k \in \mathbf{x}$ through a nonparametric function, $V_t(\mathbf{x}, t)$. Most commonly, the assumptions that go along with such an estimation strategy come from one of two methodologies, either a 'correct adjustment set', ie, assuming the economist has chosen and included the correct set of controls $\mathbf{x}$, or from a first-differences strategy, where the assumption is that differences over the time horizon of some natural experiment do not lead to systematic differences in $\mathbf{x}$. The usage of these types of models to estimate values for non-market environmental assets has been prolific and ongoing, for instance, home views are valued in Rodriguez and Sirmans (1994), forested landscapes , as in Poudyal, Hodges, Fenderson, and Tarkington (2010), and general home siting (Heyman, Law, and Berghauser Pont (2019)). Given the popularity of the methodology for natural asset valuation, and substantial significance of these results, property values are clearly influenced by many of the natural factors that may differentially affect fire suppression costs.

As an example, elevation, slope and aspect play a key role in determining viewshed for a given location, and similarly drives other types of visibility graphs, Turner, Doxa, O'Sullivan, and Penn (2001), which itself is capitalized into a property as well as playing a role in fire spread. More relevantly, fire scars within a viewshed have been shown to lead to lower property values, McCoy and Walsh (2018). Collectively, these types of environmental assets associated with a home are referred to as 'environmental amenities' and can lead to changes in property values. The legacy of this literature however leaves understanding what variables contribute to property values appear to be a catch-all for anything that can potentially vary in space. This feature of property price makes the problem of understanding what role it plays in complex economic systems highly burdensome on the econometrician to either have a compelling argument for the exclusion of any feature, or by placing a heavy burden on some natural experiment's assumption of 'as good as random' assignment.

What is often overlooked is the non-parametric relationship between objective 'quality' in an omniscient sense and 'price.' An early attempt to address the non-parametric nature is performed in a paper C. Mason and Quigley (1996), which lays out how curvature of the utility curve for housing quality drastically changes hedonic property valuation estimates, and uses a GAM to attempt to understand the shape of this curve for each estimated component. If anything is clear from the literature overall, it is that the functional mapping from characteristics to price is anything but simple.

The takeaway here is that to estimate how property prices directly contribute to some outside valuation in absence of a well-proposed experiment, it is very important that any components a consumer hypothetically may be willing to pay for must be considered in the design.

2.5 Unbiasedness through Twin Vision Transformers

There has been an increase of interest in using ML to extract direct effects in cases of complex causal systems. In particular, causal AI has begun exploring in depth the usage of Doubly Robust, Double Machine Learning (DR/DML), first described in Chernozhukov et al. (2018). Applications of the method have been expanded in recent years, and allow for robust inference on causal effects from discrete or continuous treatment effects Colangelo and Lee (2021) using nearly any machine learning algorithm so long as it converges to the estimand, as outlined in Chernozhukov, Newey, and Singh (2021). In the case of suppression costs, this meta-learning method admits opportunities to use computer vision techniques to learn extremely high-dimensional functions of controls in a doubly-robust fashion, with the goal of weighting characteristics at important locations to control for along some collection of potential fire spread paths, weighted by probability of those fire paths occurring. The double robustness provides an extra benefit, as it solves simultaneously for the treatment intensity score and conditional outcome model, defined as

$$e(x) = P(property\ values|X = x, ignition) \tag{2.20}$$

$$\mu(fire\ costs, x, property\ values) = E(fire\ costs|X = x, property\ values, ignition) \tag{2.21}$$

Where **??** is the treatment intensity score model and **??** is the conditional outcome model and X is a vector of controls. If either model can be expected to converge, than the doubly robust estimator is expected to converge.

As was outlined in section **??**, fire spreads at a rate dictated by local and adjacent fuels, weather and soil conditions along with fuel moisture. Fire suppression choices on the other hand are a response to threatened assets, fire intensity at the front of the fire and cost of available resources. The data that is available and used by fire managers to make suppression decisions is exceptionally dense and high-dimensional. This makes estimating fire suppression costs difficult to estimate without conditioning the data on either the burned region or some other kernel weighting scheme. In fact, most other studies either implicitly condition on the kernel represented by the fire boundary or by fire-relevant conditions found at the point of ignition. In truth, the kernel that maps the time and spatially-varying environmental factors to a causally valid and sufficiently rich prediction of suppression costs is highly fire and thus data-dependent. Further, how the forest service simulates fire spread in real-world applications using the FlamMap application is well understood, and is laid out sufficiently in Finney (2006).[24]

D/DML, in this case, functions in a 'partially linear regression' framework with a set of spatially-varying controls, $X_{30km,i}$, continuous treatment $Property\ Values$ and outcome $PerAcreSuppression$ (costs). There are two equations in the partial linear regression system:

---

[24] In general, fire simulation is analyzed using static environmental conditions with burn paths being generated following a minimum travel time (MTT) fire spread model, which is incorporated into the FlamMap simulation following Finney (2002). These minimum travel times can be used to label Isochrome fire boundaries, the are within which is calculated and reported as total acreage. Thus, functions of fuels, wind speed and direction as well as digital elevation model information contain all of the variation required to learn the rules to generate conditional suppression costs.

$$PerAcreSuppression_i = \theta Property\ Value_{i,20km} + f(X_{i,30km}) + v_i \tag{2.22}$$

$$Property\ Value_{i,20km} = g(X_{i,30km}) + \varepsilon_i \tag{2.23}$$

This procedure admits any machine learning procedure to estimate the set of nonlinearities in the system

$$\eta_0 = \{f(X_{i,30km}), g(X_{i,30_km})\} \tag{2.24}$$

Which will recover the correct parameter θ in a score system, ψ whose Gateaux derivative vanishes when evaluated at the true parameter θ, ie,

$$\partial_{\eta,g}\psi(X_{i,30km}, fs_i, pv_{i,20km};\ \theta, \hat{\eta})[\eta_0 - \hat{\eta}] = 0 \tag{2.25}$$

This condition will be satisfied so long as $\hat{\eta} = \{\hat{g}, \hat{f}\}$ is well specified and that $\mathcal{R}_g, \mathcal{R}_f$, the mean square error convergence rates for the chosen learning algorithms are sufficiently fast.[25]

However, in order to believe that the function of interest will produce estimates that converge to η_0, it is necessary to outline how a vision transformer, proposed and tested in Dosovitskiy et al. (2021), uses inputs to produce predictions.

---

[25] X. Chen and White (1999) has asymptotic results for neural networks, guaranteeing sufficiently fast (better than $n^{-\frac{1}{4}}$) convergence if the goal is direct estimation of the conditional mean parameterized by a neural network, which is more than sufficient for DDML. For certain subclasses of problem, this rate can be considerably faster. For more information, see Farrell, Liang, and Misra (2021).

2.5.1 Transformer. Transformers are a special form of neural network originally proposed in Vaswani et al. (2017) to translate between languages in a natural language framework, and have since been applied successfully in many of the large-scale language generation models. Their success is in part due to the ability to learn latent long-range context for words in text to successfully produce a good translation. They do this by treating positional information as a weak structure rather than a strict one, allowing the model to utilize ordering where important and ignoring it when it is not. Traditional time series models impose strong priors on the structure of the data- observations in $t + 1$ are generally 'more related' to observations in t and $t + 2$[26] Instead, the model uses the inputs themselves, words represented by large learned vector representations, to learn whether ordering is important or not and in what way.[27]

They achieve this by utilizing a so-called attention framework. Attention is a simple yet powerful way to learn data-dependent interconnections. It's useful to this work for the reader to understand the attention mechanism from an intuitive standpoint. In the classic encoder, each word-vector representation is weighted into three separate equal-length vectors with distinct uses: a query, a key and a value. Each word uses its 'key' vector to essentially advertise what content it contains, and a 'query' vector to search for keys (words in the text) that have content that informs the meaning of the word producing the query. This is done by simply performing a matrix multiplication of keys and queries and then converting

---

[26] Of course, this structure can be defined more loosely, but in every case, very distant observations are less related to the current period than more proximal ones

[27] The words in the sentence "The child threw a rock at New York diner" have vastly different meanings than "Rock diner threw a child at New York", but "The child threw a rock at diner in New York" are roughly equivalent in meaning to the first.

the resulting product into a probability distribution via the softmax function.[28] Then, from that advertised key-query match weight, a sum of values weighted by the distribution from the query key product is passed into a standard position-wise feedforward layer. Positional information can be included either as a separate dimension of the vector inputs, or included as some transformation of the vector allowing the model to utilize order when it is helpful and ignore it when it is not.

Vision transformers use this strategy instead to learn data-dependent long-distance dependencies in an image important to classification, by dividing an image into equal-sized patches and using those patches in place of word vectors. This approach is a departure from most computer vision frameworks that use convolution neural networks, which require multiple layers of overlapping filters to learn useful image-wide patterns in the data. This theoretically allows the algorithm in a single layer to learn more distant relationships than those available to a single convolution layer. This property of attention makes it very useful in wildfire applications.

Models of wildfire spread, like the ones used by the forest service, are latent sequences of currently-ignited cells, where cells at time t are determined by what cells are actively burning in time $t - 1$, environmental conditions, and type of fuel present in the cell. The issue is that a fire front at time t can be simultaneously in very distant locations. This aspect can lead to varied expected day-level-expenditures on fire suppression, even given identical cell-level characteristics. If a fire manager is expected to protect location a at time t and location b at time $t + 10$, the cost outcome to do so will be much different than

---

[28] The softmax function, where $\mathbf{z}$ is a vector of inputs, produces a vector σ where individual entries are equal to $\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_i e^{z_i}}$

the case where a fire manager must protect location a and b simultaneously over a 10 day period.

2.5.2 Combining Minimum Travel Time, NVC and Transformer Decoders.

Transformers are one of many suitable neural models capable of adjusting for spatial confounders in wildfire suppression costs.[29] Using the graphical framework of Bronstein, Bruna, Cohen, and Veličković (2021), The attention mechanism can be thought of as learning an adjacency graph $\mathcal{G}(\mathbf{v}, \mathbf{e})$, which propagates 'signals' ρ composed from vertex data x_i along edges e_i to all other nodes v_{-i} following the function

$$\mathbf{h}_u = \varphi\left(x_u, \sum_{v \in \mathcal{V}} a(x_u, x_v)\rho(x_v)\right) \tag{2.26}$$

Where $a(x_u, x_v) \in [0, 1]$. This, importantly, is a permutation invariant function, meaning, for permutation operator $\mathbf{P}$

$$\varphi\left(x_{1u}, \sum_{v \in \mathcal{V}} a(x_{1u}, x_v)\rho(x_v)\right) = \varphi\left(x_{2u}, \sum_{v \in \mathcal{V}} a(x_{2u}, x_v)\rho(x_v)\right), \forall x_{1u} \in \mathbf{X} \text{ and } \forall x_{2u} \in \mathbf{PX} \tag{2.27}$$

In other words, to the extent that positional information is important in X, it must somehow be represented explicitly within the features of the node vectors x_u.

Logged, per-acre-costs are particularly suited to be accumulated in this way by combining fire spread with $E(NVC)$ for a given cell. Fire spread models

---

[29] A sufficiently deep convolutional neural network may also be able to perform well on this task, but would potentially risk losing long-range dependencies. Additionally, a message-passing GCN with nodes aligning to points from a fire spread model and connections aligning to time steps would also be able to control for some of the variation, but would require additional engineering to allow for differential fire spread paths.

are spatial graph frameworks, where each node is ’stateful’, taking on a value of ’ignited’ or ’not yet ignited’. From some initial set of ignited nodes and given burn rate properties, time of fire arrival can be estimated for every node connected to the initial set.[30] This creates a graph whose connections are determined by node data, specifically fuel model inputs, wind speed, wind direction, slope and aspect which can be combined to produce Rothermel’s fire spread rate r for every neighboring cell, as described in **??**. This creates data-driven risk linkages between nodes that occur through canopy-canopy and ground fire spread at the same time. However, local information is required to determine if a canopy fire will initiate - which is driven by the canopy base height as well as canopy density and height. Furthermore, in the presence of a crown (or canopy) fire, long-range connections on this unobserved ignition graph can be drawn that come from ’spotting’ behavior, or when wind blows embers over long distances and start new ignitions far from the fire front. Importantly, the connections on both graphs are entirely built based on node-characteristics and relative positional information.

Thus, the $E(NVC)$ from a given action can be calculated by traversing, conditional on ignition at cell i, ’future’ nodes along some expected fire spread graph while summing across $E(NVC)$. As was outlined in section 2.3.2, NVC is a function of fire intensity, resources present, local governance’s weighting of each resource, the probability of any given burn intensity and those factors’ interaction. Thus, positional information of homes in a wildland urban interface alongside fuels, fuel conditions and any other Values at Risk (VAR) are necessary controls for a fire cost model. This location-level $E(NVC)$ can be compared to expected suppression cost in the cell, which is a function of accessibility, terrain,

---

[30] See Finney (2002) for more details

available resources and weather conditions as was described in section 2.2. If $E(NVC) > cost$, then a fire manager will likely undertake the action, incurring a per-acre-cost.[31]

Thus, the minimum travel time graph structure can be modeled in a neural network via the weighted graph structure in **??**, where a link is instantiated as $a(x_{ignited\ node}, x_{target\ node})$, dependent on relative elevation between the two nodes, fuel models and weather conditions, while effort required to suppress the fire's pathway is modeled by $\rho(x_v)$. This graph structure is however not fixed - as a fire manager may suppress one fire front

Importantly, $E(NVC)$ can be viewed as an accumulation of elements, each of which is a product of the probability of ignition and local values—each functions of local cell conditions, accumulated along a graph linked by spread dynamics. The unconditional, expected cost of suppression at a given location is specified by a function of information present within the local cell, as well as accessibility to that cell. In this way, expected total per-acre costs should be calculable through a time-invariant adjacency graph where unobservable future changes in weather or conditions enter as changes in node probabilities that weight expected benefit of suppression, with connection presence determined by the 'frozen' values in **X**.

2.5.3 The adjustment model.

To adjust for all other covariates that drive changes in cost, two separate stacked decoders that loosely resembles Compact Convolution Transformers laid out in Hassani et al. (2021) is built to learn f and g from the beginning of this section. Beginning with a raster image, sized 1001x1001x34 (see fig C1 for the input used for the CZU complex fire

---

[31] These per-acre accumulations can be negative even in densely developed areas if the expected growth of the fire is sufficiently large because $log(\frac{cost}{acre}) = log(cost) - log(acre)$

from August 2020[32]) the model must transform the input into a sequence. The model achieves this by condensing information into tokens by using a single-layer convolution,[33] reducing the height and width of the image, but increasing the number of channels, to a final size of $500 \times 500 \times 110$.

This image is then patched into a sequence of 2500 patches, as seen in figure C2, where each patch is of size 10 × 10 × 110, as visualized in figure C3 and then each is flattened to a feature vector of size 11000 × 1. These vectors are collapsed to a smaller 2870 features. Three attention modules, called 'heads', attend to these patch-level features using the key, query and value mechanism described earlier, where each head is of size $3 * 490$, for an inner-dimension of size 1470. [34] Within these heads, positional information is injected into the queries and keys using the two dimensional rotary positional embedding technique described in Su, Lu, Pan, Wen, and Liu (2021). 5 transformer-encoder blocks are then stacked[35] and passed to a linear activation that condenses the layer from 2870 to 574 (20% of the original dimension) where tabular non-spatial data are concatenated to this vector, including regional coordination center the fire manager reports to, resources currently deployed elsewhere and month of year. This is passed through a fully connected layer 10% of the prior layer, followed by a

---

[32] All figures in this document are contained in the appendix: for figures relevant to this chapter see C

[33] This convolution layer has a 6 × 6 kernel, with a stride of 1, and same padding to create a new image with 110 channels. This new latent image is then passed through an adaptive max pooling layer that produces a 500x500x110 latent image for use by the transformer.

[34] Odd numbers in the shapes of internal dimensions are primarily a function of compute resource limitations, particularly of graphical memory, which was limited in this case to 24 GB.

[35] I exclude the residual connections and layer normalization in this description for brevity, but they are included in the model

SeLU activation, followed by a fully connected layer which is then passed to a final linear activation.[36]

To prevent the model overfitting, several transformations are performed. Mix-up regularization, where inputs and outputs in each batch are occasionally replaced with the convex combination of two inputs and the label replaced with the corresponding output, with mixing probability α generated according to an alpha distribution as described in Zhang, Cisse, Dauphin, and Lopez-Paz (2018). The maximum mixing level α is set to be very low (10%), as perfect mixes between inputs is likely to include corrupt information. This is the primary mechanism to prevent overfit, but coarse dropout and shift/rotate transformations are also applied to the training image pipeline. To prevent overfitting on the tabular data, the first five epochs are trained on raster images only, with tabular data replaced with randomly generated inputs.

Batch sizes are forced to be small due to the size of the inputs and the limited graphical memory, so learning is done over 4 subsets of 5 in mini-batches of size 20, using the pytorch implementation of the AdamW optimization algorithm; engineered initially in Loshchilov and Hutter (2019), to guide updates in the weights. This algorithm allows for 'true' L2 regularization, which is important for good behavior in a DDML estimate. A relatively slow learning rate of 3.0×10^{-5} is used, with a weight-decay value of 2.5×10^{-6}, and modulated throughout training by a cosine annealing with warm restarts learning rate scheduler, where restarts occur every 2000 steps, and the model is trained in total for 155 epochs, where 5 of those epochs are raster-only.

---

[36] This makes the model multi-modal.

2.5.4 'Causal' Model. Verifying whether the engineering 'makes sense' or not requires a documentation of assumptions. In general, this work takes the approach of past works investigating the link of property value to suppression cost—conditional independence. This is not to say that conditional independence necessarily is satisfied in this case, but rather as a mechanism to validate past estimates of the elasticity parameter.[37] To fully select covariates to adjust for, a comprehensive non-parametric causal structural model is built using studies of wildfire spread done by the forest service, validated links describing how amenities are affected by risk-associated variables and lastly, how suppression resources are allocated to wildfires in the contiguous United States.

Achieving a causal estimate of our outcome amounts to combining all of the models discussed, and identifying identified factors that may lead to non-causal association. To keep them organized, they have been written out in table E2.[38] A graphical representation of this table is drawn, using the dagitty tool in fig C6. This causal model's pathways are well verified by the existing literature, and as a whole represent a small step towards relaxing the typical assumption of exogenous amenities and also include an added 'algorithmic bias' pathway.

Returning to the introduction, the work laid out five potential associative pathways that may influence the association between property value and suppression costs. In the given causal model, they appear in separate pathways.

---

[37] The core idea of this causal identification strategy being that either the conditional outcome from treatment is causal, or that the treatment level assignment (treatment intensity model) is causal.

[38] All tables in this document are contained in the appendix: for tables relevant to this chapter see D

- *Direct Path:* This represents the 'direct effect' of property value on suppression costs. This can be seen in the pathway $Property\ Values \rightarrow VAR \rightarrow Strategic\ Concerns \rightarrow Resource\ Assignments \rightarrow Cost\ Per\ Acre$
- *The "Amenities" Path:* This represents the association (potential) between how amenities increase property values and simultaneously may be affected by fuels, fire risk, etc.. This can be seen in the pathway $Property\ Values \leftarrow Amenities \leftarrow Fuels \rightarrow Strategic\ Concerns \rightarrow Resource\ Assignments \rightarrow Cost\ Per\ Acre$. It also is represented by the path traveling through VAR: $Property\ Values \leftarrow Amenities \leftarrow Infrastructure/Public\ Values \rightarrow VAR \rightarrow Strategic\ Concerns \rightarrow Resource\ Assignments \rightarrow Cost\ Per\ Acre$
- *The Procedural Path:* This represents how homeowners may choose locations where it is more costly to defend because defense costliness is related to amenities. This is represented by the pathway $Property\ Values \leftarrow Amenities \leftarrow Infrastructure \rightarrow Safety \rightarrow Strategic\ Concerns \rightarrow Resource\ Assignments \rightarrow Cost\ Per\ Acre$
- *The "Model Loop" Path:* This path represents how using property values may distort fire managers' objectives. This path is represented by $Property\ Values \rightarrow SCI \rightarrow Budget \rightarrow Resource\ Assignments \rightarrow Cost\ Per\ Acre$

The most important way portion of this exercise is to identify a sufficient control set to pass to the model from the prior subsection. In full, adjustment for weather inputs (that may affect fire spread), fuels (local to property and relevant for fires), Infrastructure that is protected/provides protection, such as cell towers and roads, public resources, such as national parks/forest service land, and property locations/population density. To fully describe these causal factors, a thorough and sufficient dataset to describe this minimal adjustment set is outlined in the next section.

2.6 Data

The data for this study is partly spatial and partly tabular in nature, and is gathered from several public-facing datasets. Dynamic panel data of resource assignment to fires and complexes comes from a public store of updates from IROC and local fire managers to IRWIN (Integrated Reporting of Wildland-Fire Information) used to calculate total costs and number of burned acres, as well as find the central location of the fire Wamack, Green, and Stringer (n.d.). This data is updated every minute, though the values change only when fire managers submit new information. That is, the frequency of changes is not constant, and only some data is shown at a time. In addition, strategic information on assets at risk as well as a forest service calculated accessibility layer that approximates maximum movement speed in the United States come from the WFDSS (Wildfire Decision Support System) WFDSS (n.d.). Weather data is gathered from both the NARR/NCEP dataset - Mesinger et al. (2006), which is a gridded dataset of meteorologic data with a resolution of 36x36 km and the CPC, which provides information on drought conditions in an area for a given month from the same source. Early fire location information comes from a combination of latitude and longitude reported coordinates from the IRWIN database, but also regionally located using the VIIRS fire satellite product, which consists of a 375 meter resolution fire detection, alongside a measure of estimated ’intensity’ of burn in the given cell, called Fire Radiative Power (FRP.) VIIRS Level 1 Processing Group At Ocean SIPS (2017). Fuel models, elevation (and associated DEM characteristics), Vegetation Cover are all sourced from the LANDFIRE 2019 remapping of the 2016 data - inputs used actively to assess management choices in wildfire decision scenarios used by fire managers and wildfire simulations to understand and predict spread LANDFIRE (2019). Data on development come from the National Land Cover Database, using the ’impermeable’ measure, specifically that that has been linked to development Wickham, Stehman, Sorenson, Gass, and Dewitz (2021). Lastly, a set of raster layers with a resolution of 270x270, representing forest service models of probabilistic fire risk, generated from repeated simulated fire seasons in 2016 are included as a proxy for expected risk in a given location Short et al. (n.d.). These inputs include conditional flame lengths, or the expected height of a flame, conditional on being ignited, and seasonal burn probabilities from 2016. To limit the number of included layers, a single raster layer created from the PAD shapefiles created in Prior-Magee et al. (2020) is used to understand which areas are under environmental protection. Lastly, a 100m resolution raster layer of disaggregated 2017 ACS US population linked to Microsoft’s open source building dataset is created in Huang, Wang, Li, and Ning (2020), and can be downloaded from Huang (2020). This allows for a much finer grained understanding of where individuals live in the region around the fire, without directly controlling for property values or directly adjusting for individual homes. For an example of these data, see figure C1, which shows a raster input for the CZU complex fire in 2020.

All fires reported through IRWIN occurring in the Contiguous United States are included, so long as they grew to a size of greater than one acre, have a reported final cost of greater than ten dollars. The LANDFIRE, WFDSS, NARR, NLCD, Huang and Short raster layers are stacked to create a single CONUS strategic consideration set. From this large raster, a uniform $30 \times 30\ km$[39] study area; unconditional on fire size, is assigned to each wildfire centered around the point of ignition. This raster image is a $1001 \times 1001 \times 34$ matrix, and is fed to the transformer model discussed in section 2.4.

Data are split into 10 folds which each consist of roughly equally sized non-overlapping training and validation/testing sets for both tabular and raster data, linked by a common ID. In total, information about 1750 wildfires is included, making training samples for each fold 1575 in size with 175 observations held out for validation and analysis in the D/DML procedure. Each observation is linked by a unique character string from the IRWIN database, labeled 'Irwin ID' and to a 34 layer raster and tabular, non-spatial data. Raster data can be seen in table E1, and tabular data used by the transformer are regional indicators [40] (GACC), along with binary indicators for Initial Fire Strategy, month of year and the total number of personnel assigned to other fires.

From the WFDSS, several other metrics are used in the linear regression model that are unneccessary for the transformer but bring the linear model in line with past estimation, namely, distance to Inventoried Roadless Areas, National Recreation Areas, Class I Airsheds, Wilderness Designated Areas and Census Designated Places. In addition to the distance metrics, binary indicators for fires beginning inside of (I = 1) or outside of these regions (I = 0) are included.

For the transformer model, all inputs are normalized using the common min-max normalization technique to be consistent with traditional computer

---

[39] plus 30 meters for a center cell

[40] As these metrics enter with more than one layer of weights, this work implicitly includes interactions of these with the raster data

vision inputs, leaving them to range from 0-1.[41] To maintain coefficient consistency across estimation methods, label variables, namely per-acre costs and total property value are first logged, and then normalized to $N(0,1)$, by first subtracting the training sample outcome mean from the value and dividing by estimated training outcome standard deviation. This allows for a quick transformation to recover predictions of logged per acre fire costs to allow for a comparison of linear regression estimates of the conditional effect of interest to the transformer's predictions of those same estimates, without forcing the transformer to perform in an environment for which it is not optimized.

A full set of descriptive statistics can be seen in table: E1. Additionally, full correlation tables utilizing τ non-parametric correlations are reported in C11. Total fire suppression expenditure, broken down by management region and split out by individual fire are available to be seen in C13i.

2.7 Results

2.7.1 Difference in estimated ATE of Property Value on Suppression Costs. Performing the analysis, the results are fairly clear, the 'causal' effect of property value on suppression costs per acre as described by the SCI and following works is much smaller when using DDML to adjust for spatially varying covariates. The estimate derived in this study implies a .06% increase in per acre suppression costs for every 1% increase in property values within twenty kilometers, versus the prior estimate of .11%. The point ignition values, using covariates derived from Gebert et al. (2007), are significantly higher as well, with an estimate of the ACE

---

[41] For rasters with known maximum and minimum, those layer-wide values are used, else estimated minimum and maximum, generated through random sampling of points from the full image, are used

equal to a .16% increase in suppression costs for every 1% increase in property values.[42] Results can be seen in full in table E3.

Another benefit is that the DDML model, out of sample, significantly improves fit over the in-sample reports from both this studies' R^2 and past estimates' R^2 at a respectable .859, meaning roughly 86% of the variation in wildfire per-acre expenditure is explained by the conditional outcome model and property value. Though R^2 is sensitive to over-inclusion of variables, it is not when using out of sample R^2, implying these models are indeed learning the majority of the causal variation from the included variables.

However, we can visualize the mis-specification bias in the model by examining the salience maps from the conditional outcome model and the treatment intensity model. Though salience maps are not robust to adversarial disturbance of the inputs, they do provide a good visualization as to how the model is using spatial inputs. Using the CZU complex fire as an example, we can see how each submodel learns variation to adjust for in $\hat{f}$ and $\hat{g}$. This fire resulted from a series of lightning strikes in the Western region of the south-San Francisco Bay Area, in 2020. Many high-value homes were at risk, with a large city-center in Cupertino to the Northeast. Winds were initially blowing from the ocean in a North-east direction. In figure C7, there are several regions the model 'pays attention to' that overlap in a way that is not easily controlled for by simply including an average of the burned area/using ignition location controls. Looking at the inputs in figure C1 and comparing them to the activation maps seen in C7, the model is in both cases using inputs near small communities, as well as areas with exceptionally low accessibility. This indicates that property values, and likely

---

[42] This is higher than estimates found in Gebert et al. (2007), but not significantly so, as .11% falls within the confidence interval for this estimate

amenities in particular serve as good controls for nearby strategic concerns and fire risk.

These point-estimates are also significantly lower than historic point estimates from Gebert et al. (2007), which are roughly .11%.[43] Results conditional on being in the Western United States (common in the literature), despite a significantly smaller sample size are smaller (though not significantly so) than comparable point-ignition level estimates. It must be noted, that most analyses implicitly condition on fire size by only including fires over a given acreage- generally restricting observations to those fires over 100 acres. Though, given the causal model, that potentially makes these estimates non-causal, to get a valid comparison between past estimates of fire suppression costs per acre and these, a conditional prediction should be evaluated.

2.7.2 Model Comparisons. Despite good out of sample performance on the relatively simple predictive task of per-acre suppression cost, there is a need for caution before adoption in the transformer-based SCI. Overall, there are three main reasons for concern - **first** ease of adoption, **second** lack of quantile predictions and **third** only having a single model across multiple agencies.

Importantly, fire managers over many years of experience with the WFDSS and its cost predictions are simply used to working with the outputs of the original stratified cost index—and its shortcomings. Given its usage for 13 years in the forest service, the model outputs of the SCI cost estimate are very interpretable to individuals who plan for fire actions. A fire manager or accountant can understand when an environmental variable outside of the SCI's consideration set will

---

[43] In these early results, Western Region results do not necessarily reject the results from Gebert et al. (2007), but as more predictions are produced I expect the confidence intervals to shrink considerably.

likely bias a cost estimate produced, either downward or upward, and thus can articulate why the fire I manage may need more or less resources than the model suggests. The ViT estimate; while having access to all spatial variation, may or may not adequately condition on what the fire manager in question thinks is important, and thus may lead to fire managers asking for more or less resources than is efficient given it is less interpretable. Corrections for this could be made by producing a salience map, but even then they are not fully able to capture the discord between what the local expert thinks is important and what the model does.

Second, current models, in addition to a traditional mse-based OLS prediction produce quantile estimates (usually 25th and 75th percentiles) of per acre costs based on estimates derived from extreme coefficient bounds. This allows them to produce a model-consistent range of possible per-acre costs a fire manager may face, and gives the manager considerable flexibility in requesting resources. Doing this with the ViT-based model in its current form is not possible because pixel-level coefficients do not come equipped with standard errors and estimates produced therefore do not represent a distribution, rather, a single point estimate of expected cost. This prevents fire managers from understanding what an acceptable expense ’overage’ or ’deficiency’ is in context of these estimates. A further model for deployment should likely utilize *quantile loss*; originally laid out in Koenker and Hallock (2001), or a more modern version like that seen in Ben-Or, Kolomenkin, and Shabat (2020) to similarly produce ranges of estimates for per-acre costs.

Lastly, prediction power is only one of many concerns of a model in use for fire suppression resource assignment. In practice, fire management in the

US occurs across a complex tapestry of agencies, each with their own concerns and priorities that should be weighted in the model. The SCI directly conditions on these by building separate models for each agency and subdivisions within those agencies.[44] While the neural model of fire costs should be able to similarly condition its predictions, it is not guaranteed to, and may underserve certain groups and agencies in favor of 'better-fit' elsewhere. A proper implementation of this for public use should produce separate models for each agency and region either by conditioning data to each area, or by finetuning some core 'backbone' model on agency-specific data.

2.7.2.1 A note on implied efficiency. A reader who is versed in classical microeconomics may question what implications the lack of correlation with property values, when surely market value of property should serve as a good signal for fire suppression effort. Answering the question of 'optimal suppression' thoughtfully is very complex and is being investigated currently.[45] To identify what is optimal in a suppression context requires understanding not only the full non-market value of resources at risk, but also how those non-market values contribute in turn to market-based values of property.

The author therfore conditions the statement that the machine learning model implies summed property value is not a factor in suppression choices by fire managers; to: summed property value, *conditional on all environmental amenities and property location* does not appear to have an effect on per-acre suppression costs *on average*.

---

[44] For instance, Eastern vs. Western United States GACC

[45] see Plantinga et al. (2021) and ongoing, unpublished work by economist Josh Olsen for more research on how property value factors into optimal suppression

This is a much different statement than the one that implies property value has no importance in fire suppression effort, rather, that it features minimally on top of property presence/density and environmental amenities.

2.7.3 Evaluating the Conditional ATEs. Examining the trend of $P(\text{Suppression Cost}|X, Acres)$ as acres increase in figure C10, I find similar results to those identified previously in the literature: the effect conditional on small fires is noisy and high and the role of property value appears to decrease as acreage increases. Mechanically, this makes sense, as our outcome is a decreasing artificial function of acres- $ln(\frac{\text{Suppression Cost}}{\text{Total Acreage}})$. One point of interest is to examine the conditional estimates of property value to look for any evidence of the fourth causal pathway - ie, the model-loop path. Exploiting the fact that cost-monitoring is only available once a monitored fire exceeds 300 acres, we gain access to a reasonable test for the existence and sign of such a path by using logic modified from the familiar regression discontinuity design (RDD).

For fire i, the estimated effect of moving from class D to class E on manager responsiveness to nearby property value under a standard non-parametric RDD specification can be represented by examining a binary indicator D for treatment, which toggles based on whether or not the final fire acreage falls above or below the 300 acre threshold. The target estimand in potential outcomes notation then is $E[Y_1 - Y_0|Acres = 300, \mathbf{X}]$ where Y_1 is per acre costs when the fire is assigned (at random) class E (the treatment), Y_0 is per-acre-costs when the fire is assigned class D.

Our problem reduces to a relatively simple case when our running variable; 'acres', is a continuous variable that satisfies:

$$(2.28)$$

$$Y_i = P(\alpha + \tau D + \beta_1 (X_{i,1} - c) + \beta_2 D(X_{i,1} - c) + \varepsilon_i | X_{i,2})$$

$$D = 1 \iff X_{i,1} > c,\ X_{i,1} \equiv \text{acres},\ X_{i,2} \equiv log(\text{total property value}),\ c \equiv 300 \text{ acres} \tag{2.29}$$

Where τ represents the quantity of interest, and can be interpreted as the percent change However, our so-called ’running variable’; acres, is likely a potential outcome of suppression effort which we indirectly measure through per-acre cost.

However, the effect appears to decrease quickly up to fires with total acreages earning a classification of ’D’, after which the estimated "causal" effects increase slightly and stabilize. This cutoff is more meaningful than it appears however, as WFDSS forecasts of total suppression costs that all fire managers have access to are only provided to managers overseeing fires larger than class D (fires greater than 300 acres.) The difference between class D fires and fires of size class C and larger is statistically significant (as can be seen in figure C9), providing some evidence that the ’SCI’ causal pathway is in fact meaningful, and appears to exert some upward pressure on per acre suppression expenditures. Estimates of total property value on per-acre-costs when restricting our sample to fires > 100 acres in size are .048, with an upper bound on the 95% confidence interval of approximately 0.10. This rejects the hypothesis that the coefficients from the nonlinear system are identical to the coefficients in past studies or those in the replicated OLS system.

Further evidence of the 'monitored' effect comes in the form of correlation between actual estimated final costs and total property value. Estimates produced by the approach described in section 2.4 are less than $\frac{1}{4}$ as correlated to total nearby property value as published estimates of final costs.[46] Even though the work did not set out to produce a less biased estimator, using causal foundations for the estimate appears to have produced that result.

The estimated value for θ—logged property value's affect on logged per-acre suppression costs, also varies across management regions, which can be seen in C5. Only the coordination regions covering California (ONCC and OSCC), the Southwestern United States (SWCC) and the Rocky Mountains (RMCC) have coefficients that are distinguishable from 0. However, these regions also represent four of the top five[47] largest coordination regions by suppression spending for the 2020 and 2021 seasons, as is seen in figure C13i.

In general, though the conditional effects do not alone validate the model, they do seem to provide evidence in support of assumptions made in the structural causal model, even those that are not included in the underlying learning model.

2.8 Conclusion

Accurately estimating fire suppression costs for fire managers has been a stated goal of the forest service for 15 years, and much work has been done trying to produce good causal estimates. These estimates of the ACE, without carefully controlling for spatial variation in the covariates, appear to be positively biased. This result is quite important for the forecasting community, however,

---

[46] Neither estimate is strongly correlated, however-WFDSS produces a forecast that has a coefficient of .05 and transformed estimates (from logged per-acre to total-suppression of the predictions produced by this paper produce a coefficient of .009

[47] The Pacific Northwest (NWCC) spent more on wildfires in this time frame than the Southwest region (SWCC)

as forecasters appear to have introduced an algorithmic bias for fires occurring near masses of expensive property. There does appear to be some attempt to de-correlate the measure of estimated cost from property values, as estimated final costs are only very weakly correlated with nearby property value, but there still appears to be some distortionary effect that occurs once fires are eligible for monitoring.

Though property value has served as a very good proxy for nearby risk to structures, it is not necessarily a good covariate to include when trying to produce good predictions, and one can produce extremely good results without relying on those values. The cost of including the value of privately owned land in a model is that Using computer vision techniques developed in the last two years, this work can produce more accurate and less stilted towards property value estimates of final cost - potentially alleviating an increasing concern of unequal cost assignment in fire suppression expenditure.

There is much room for future work-on the algorithmic front, optimizations in network architecture to fully make use of the wildfire communities knowledge base and restrict the solution space of the model to something more restrictive than a transformer's may boost performance. Additionally, there is much room to decrease the noise on predictions and potentially find more accurate estimates of conditional effects.

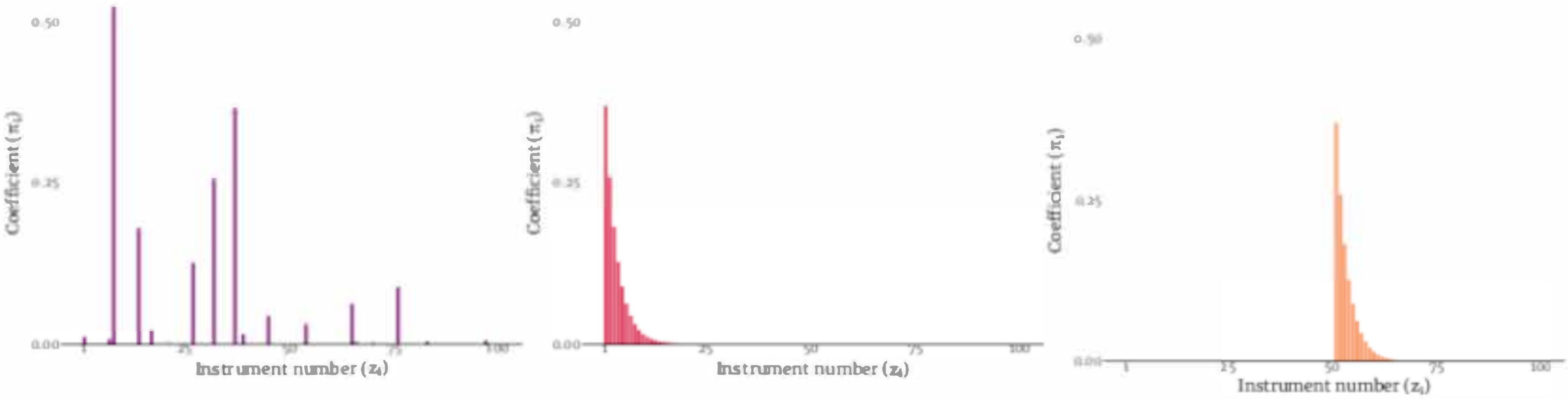
APPENDIX A

CH 1 FIGURE APPENDIX

Figure A1. **The "high-complexity"** DGP: 100 instruments of varying strength

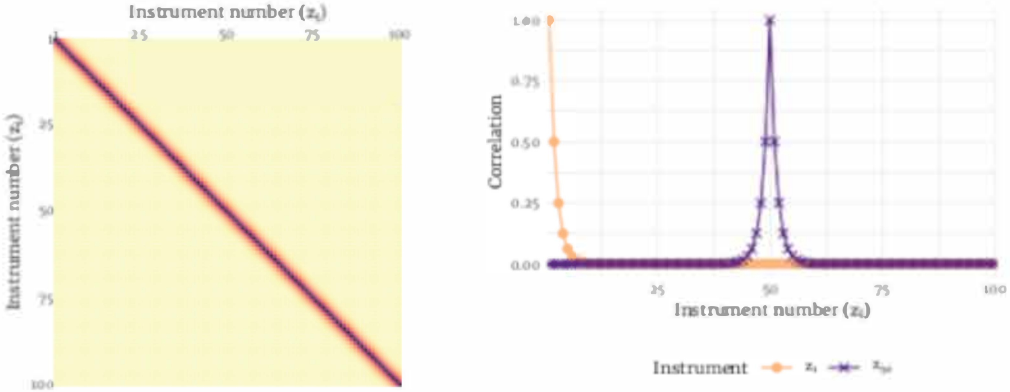
Order of instruments' coefficients (π)

(i) *Subcase 1:* 'Shuffled' coefficients (ii) *Subcase 2:* Coefficients decline from z_1 (iii) *Subcase 3:* Coefficients decline from z_{50}



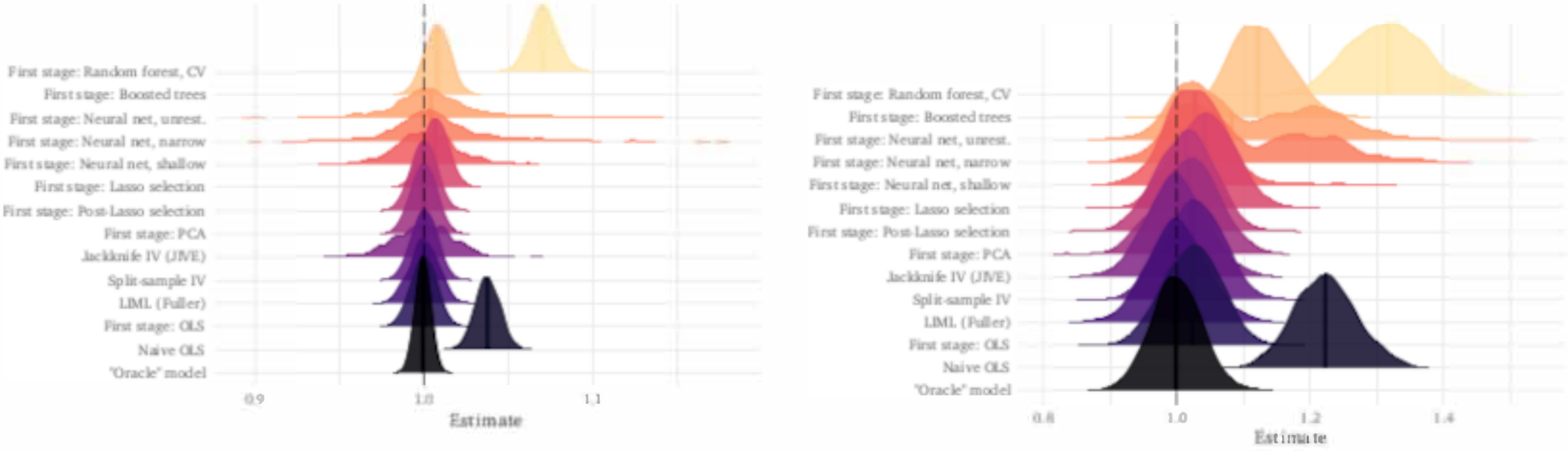
Instruments' correlation (Σ_z)

(iv) Correlation between the 100 instruments (v) Correlation of z_1 and z_{50} with other instruments

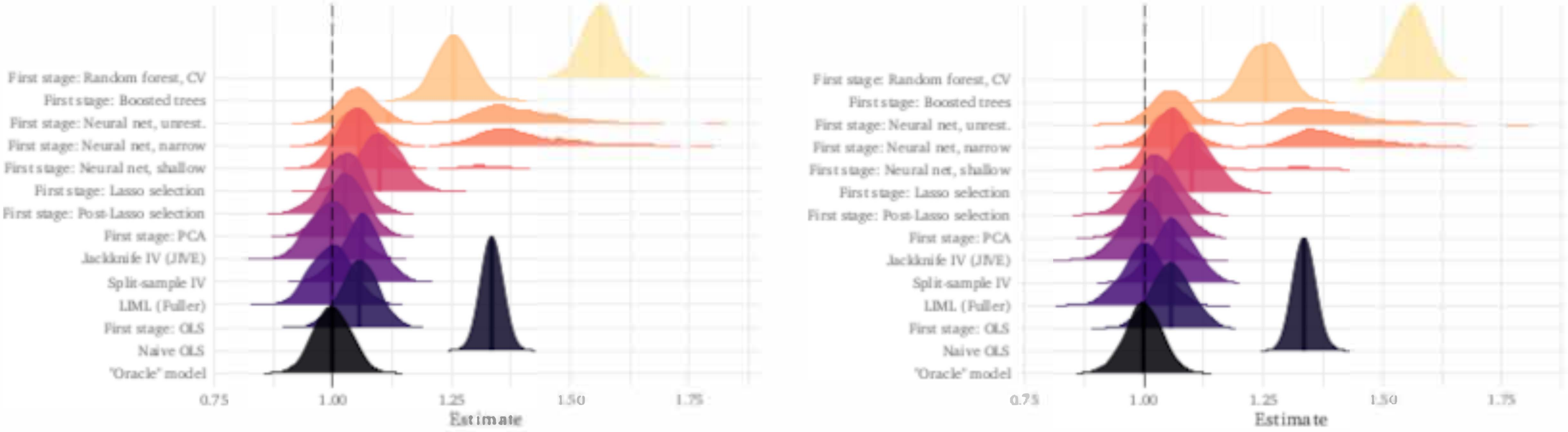


The top panels—A1i, A1ii, and A1iii—illustrate the instruments' coefficients (π) in the DGP for the first stage of each of the three subcases (i.e.how the instruments z relate to x in each subcase).

Figure A2. Main results—$\hat{\beta}$ distributions across competing two-stage methods


(i) ***Low-complexity* case**: 7 strong instruments

(ii) ***High-complexity* case 1**: 100 instruments; coefficients 'shuffled'


(iii) ***High-complexity* case 2**: 100 instruments; strength decreases from z_1

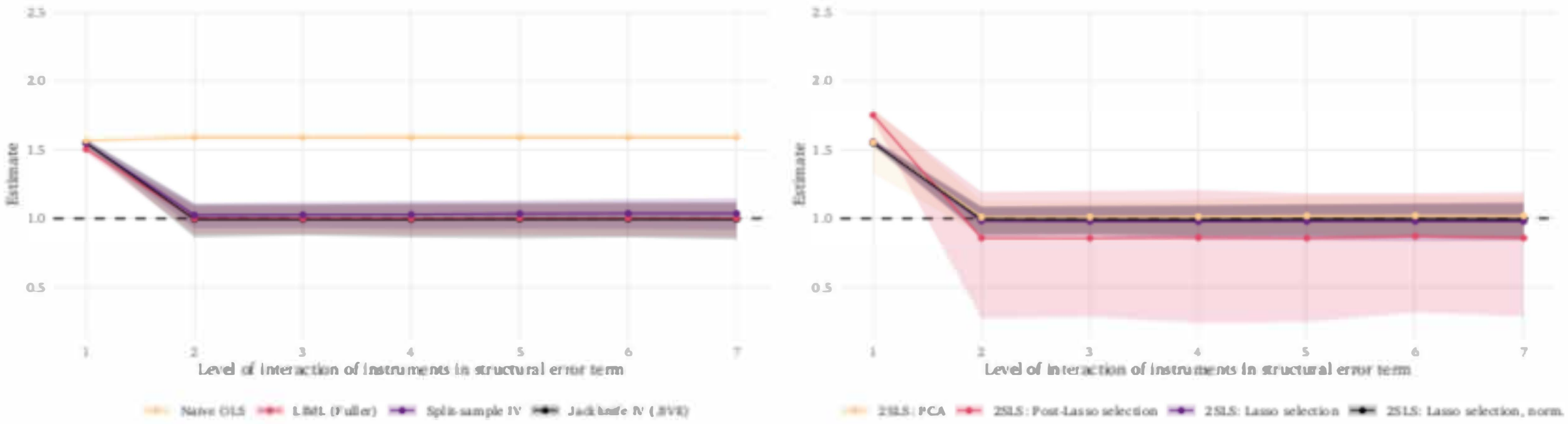
(iv) ***High-complexity* case 3**: 100 instruments; strength decreases from z_{50}

Each individual distribution represents the density of estimates from 1,000 iterations of simulation. The true value of the target parameter β is 1 (the dashed vertical line). The DGP underlying subfigure A2i uses 7 strong and exogenous instruments. For subfigures A2ii, A2iii, and A2iv, the models have access to 100 exogenous instruments of varying strengths.

Figure A3. **Exclusion-restriction violations** via higher-order interactions among instruments ('low-complexity case' of 7 strong instruments)

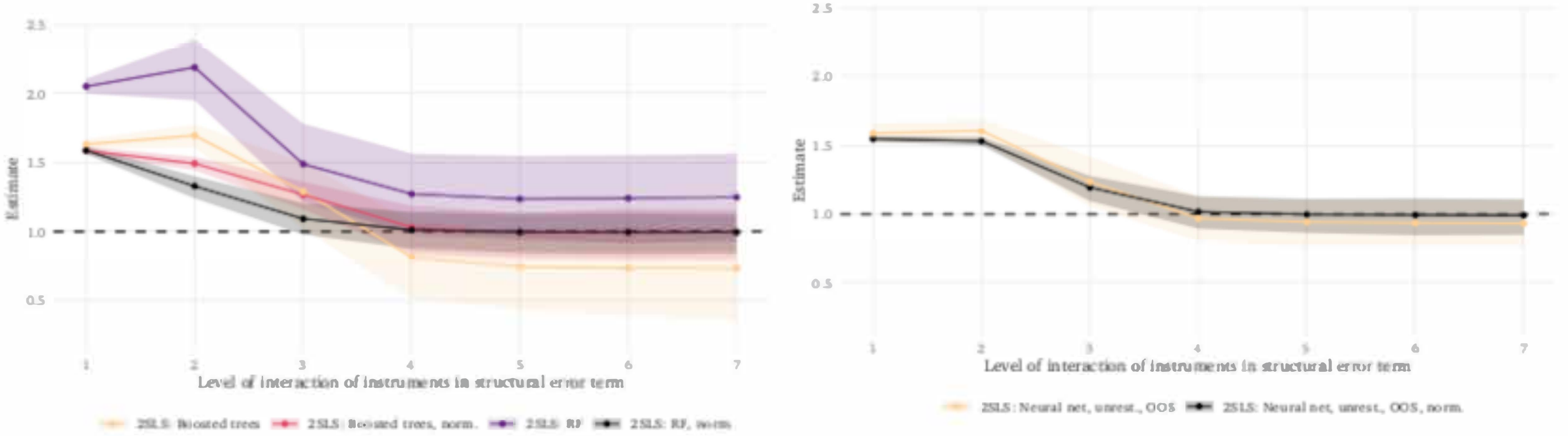
(i) **'Standard' linear estimators:** OLS, LIML, SSIV, and JIVE

(ii) **'Selection' methods**: Lasso, post-Lasso, and PCA



(iii) **Trees**: Random forests and boosted trees—with and without normalization

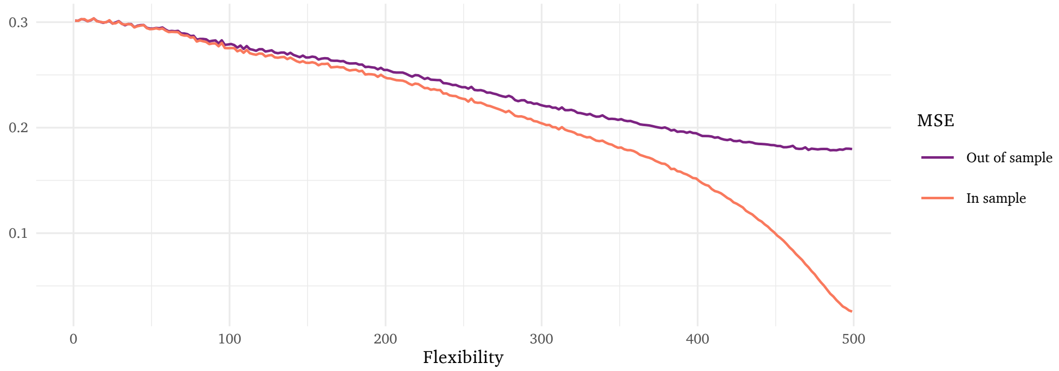
(iv) **Neural networks**: With and without normalization



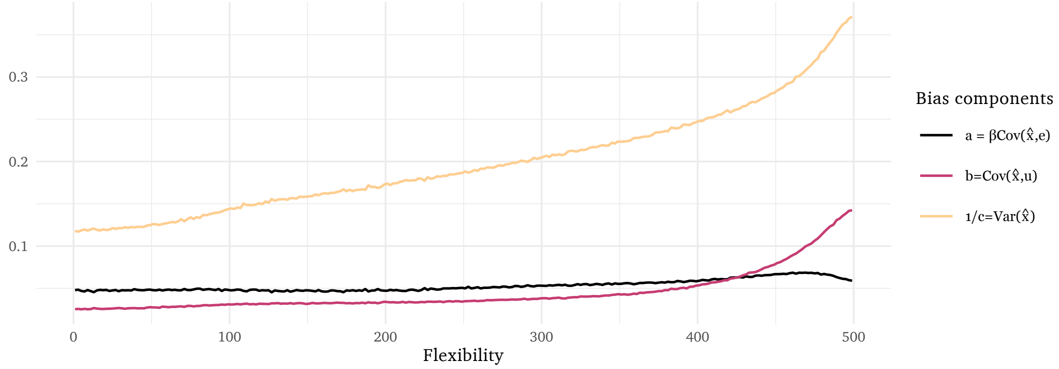
The DGPs underlying these figures add a k-term interaction between the instruments to the structural error. *E.g.*, when the x-axis equals 3, we add the interaction $x_1 \times x_2 \times x_3$ to the error; $k = 1$ is a linear exclusion-restriction violation. Interactions with $k > 1$ do not violate the exclusion restriction of linear methods but do violate the exclusion restriction for methods that require a higher-order/expanded exclusion restriction (i.e.conditional independence).

Figure A4. **Predictions *vs.* estimation:** Comparing cross-validated prediction performance with bias in random-forest-based 2SLS

(i) In- and out-of-sample MSE for predictions of x



(ii) The three 'bias components' in estimating β



(iii) Bias in $\hat{\beta}$ as a function of model flexibility ($\beta_1 = 1$)

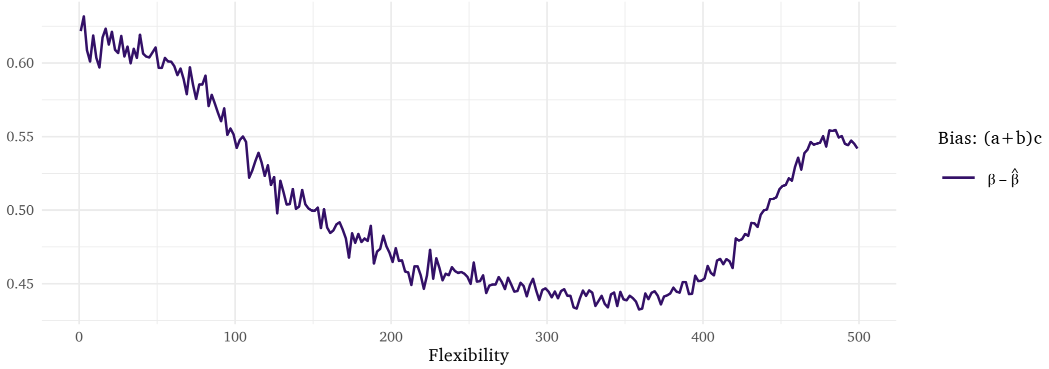
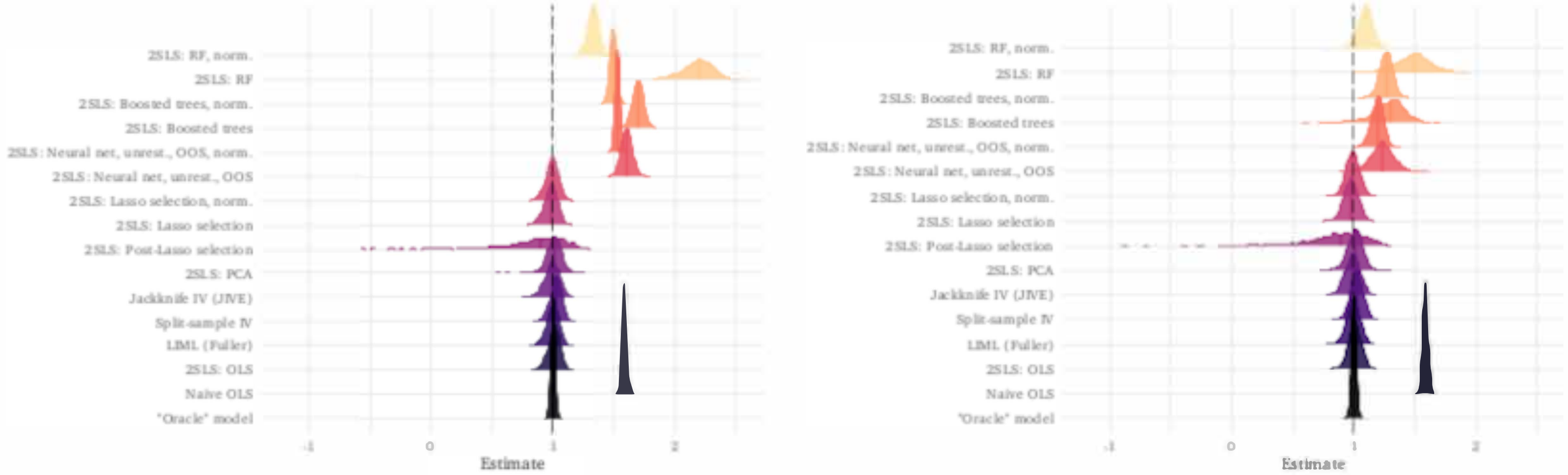
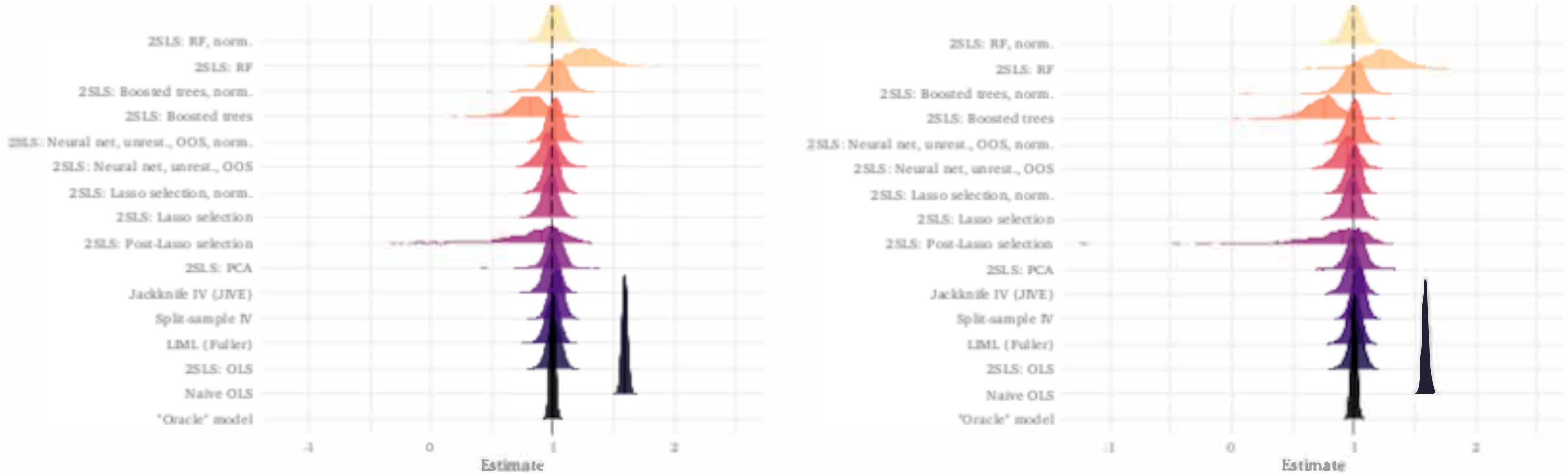


Figure A5. **Distributions of estimates** with "exclusion-restriction violations" from *k*-term interactions ('low-complexity case)


(i) **Two-term 'exclusion-restriction violation':** $x_1 \times x_2$

(ii) **Three-term 'exclusion-restriction violation':** $x_1 \times x_2 \times x_3$

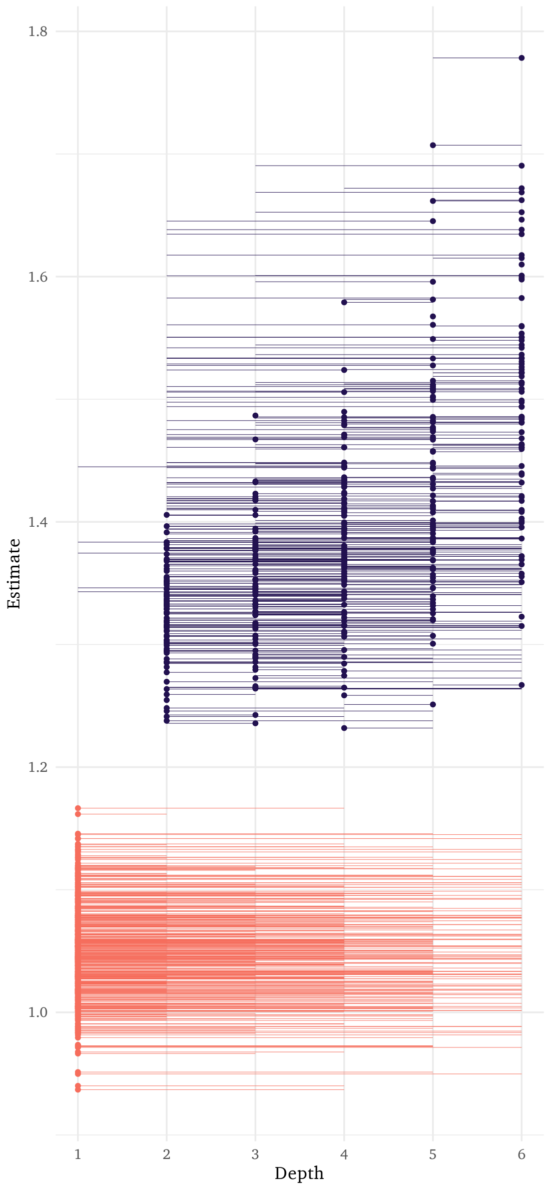

(iii) **Four-term 'exclusion-restriction violation':** $x_1 \times x_2 \times x_3 \times x_4$

(iv) **Five-term 'exclusion-restriction violation':** $x_1 \times x_2 \times x_3 \times x_4 \times x_5$

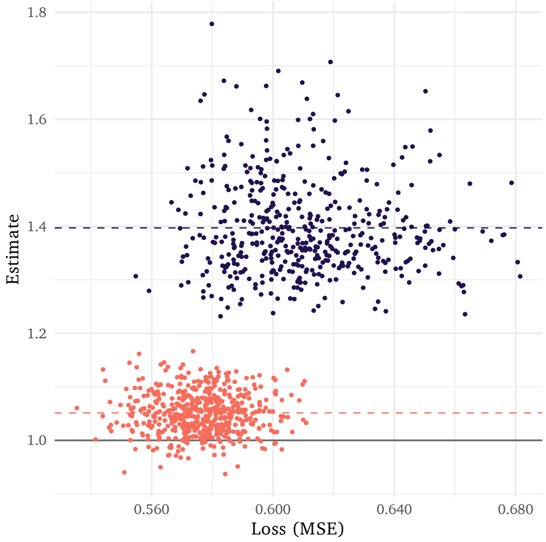
This figure illustrates the densities of the estimates portrayed in Figure A3. Post-lasso's high-variance in this experiment come from using the so-called plug-in lambda value that prevents overselection of poor-performing instruments.

Figure A6. Explaining unrestricted/narrow neural networks' bimodal distributions of $\hat{\beta}$

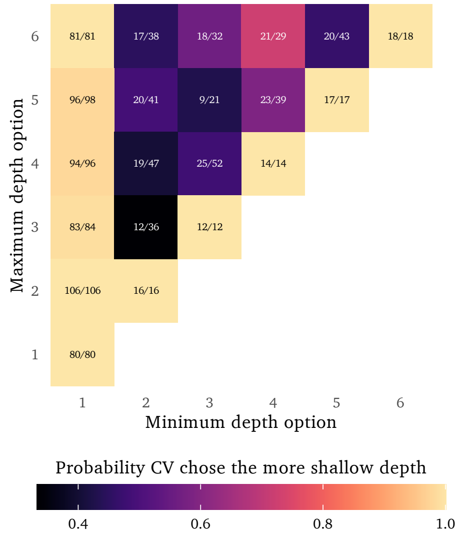
(i) Comparing bias in $\hat{\beta}$: Approximately linear (no hidden layers) *vs.* 'deeper' neural networks



(ii) Comparing bias in $\hat{\beta}$ and out-of-sample loss: Approximately linear *vs.* 'deeper' models



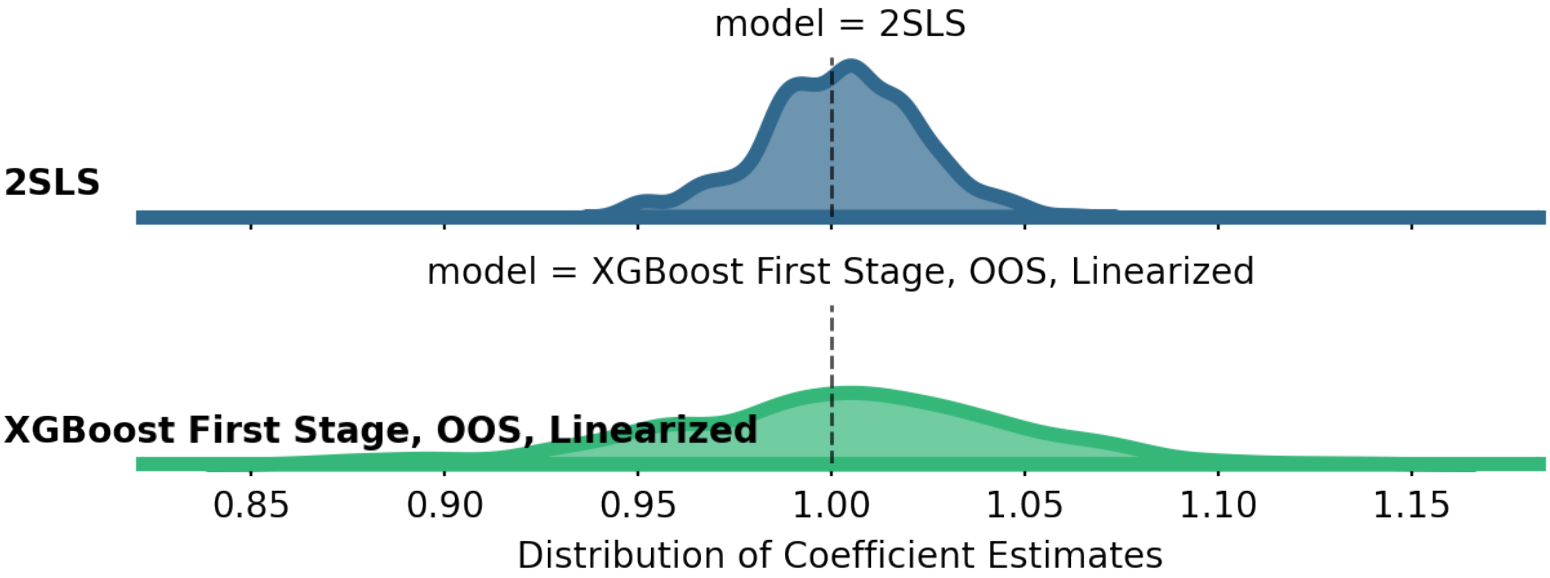
(iii) Cross validation's likelihood of choosing more shallow models when choosing between two options



Maximum depth option	1	2	3	4	5	6
6	81/81	17/38	18/32	21/29	20/43	18/18
5	96/98	20/41	9/21	23/39	17/17	
4	94/96	19/47	25/52	14/14		
3	83/84	12/36	12/12			
2	106/106	16/16				
1	80/80					

The y axis of Panel **a** depicts the second-stage estimate $\hat{\beta}$, and the x axis represents the depths of the cross-validated neural networks. "Depth 1" implies no hidden layers—directly linking the input and output (approximating linear regression). Horizontal line segments in **a** connect the two possible depths that the model chose between. The solid dot marks the chosen depth (by cross validation).

Figure A7. **Distributions of estimates** under heterogenous treatment effects



This figure illustrates the densities of the estimates from 500 simulations of the scenario outlined in E.3. Only 2SLS is used in this case as a comparison against an XGboost-based first stage, cross fit on two folds and linearized as described in J. Chen et al. (2020). 7 instruments are used, with a $\Gamma(.5, 4)$ distributed coefficient shared between all 7.

APPENDIX B
CH 1 TABLE APPENDIX

Table A1. Simulation results: Mean and standard deviation for methods and DGPs

	Low-complexity case	*High-complexity* cases		
	7 strong instruments	100 *mixed* instruments		
	(A)	(B)	(C)	(D)
Naive OLS	1.038 (0.007)	1.335 (0.020)	1.334 (0.019)	1.223 (0.046)
First stage: OLS	1.000 (0.007)	1.058 (0.040)	1.056 (0.039)	1.023 (0.042)
LIML (Fuller)	1.000 (0.008)	1.000 (0.044)	0.998 (0.043)	1.000 (0.043)
Split-sample IV	1.001 (0.007)	1.062 (0.041)	1.060 (0.040)	1.025 (0.043)
Jackknife IV (JIVE)	1.000 (0.017)	0.998 (0.044)	0.996 (0.044)	1.000 (0.044)
First stage: PCA	1.000 (0.007)	1.032 (0.042)	1.026 (0.041)	1.016 (0.045)
First stage: Post-Lasso selection	1.000 (0.007)	1.026 (0.044)	1.023 (0.042)	1.013 (0.042)
First stage: Lasso	1.007 (0.007)	1.100 (0.045)	1.098 (0.045)	1.042 (0.045)
First stage: Neural net	1.008 (0.029)	1.215 (0.180)	1.209 (0.176)	1.110 (0.105)
First stage: Neural net, shallow	1.002 (0.018)	1.069 (0.066)	1.065 (0.067)	1.030 (0.049)
First stage: Neural net, narrow	1.008 (0.027)	1.213 (0.183)	1.210 (0.185)	1.100 (0.103)
First stage: Boosted trees	1.008 (0.007)	1.254 (0.041)	1.255 (0.039)	1.121 (0.047)
First stage: Random forest, CV	1.071 (0.008)	1.562 (0.033)	1.563 (0.034)	1.316 (0.058)

This table provides the means and standard deviations of the distributions illustrated in Figure A2. Each **column** contains a separate DGP: (a) contains the *low-complexity* DGP with 7 (equally) strong instruments; (b)–(d) contain the three *high-complexity* cases with 100 instruments of mixed strengths. **Rows** differ by estimator. For each DGP-estimator combination, we summarize the estimates for the parameter of interest (β) across 1,000 iterations using a mean and standard deviation (the standard deviation is in parentheses).

Table A2. Simulation results: Decomposing bias components (means from simulation)

		Bias components							
	$(a+b)c$	a	b	c					
Estimator	Bias	Cov($\hat{x}, e$)	Cov($\hat{x}, u$)	1/Var($\hat{x}$)	Var($\hat{x}$)	Var(x)	Cov($x, \hat{x}$)	Corr($x, \hat{x}$)	Cov(x, u)
Panel A DGP: *Low-complexity* case									
Naive OLS	0.04	0	1.01	0.04	26.41	26.41	26.41	1	1.01
First stage: OLS	0	0	0.01	0.04	25.41	26.41	25.41	0.98	1.01
Split-sample IV	0	0	0.02	0.04	25.42	26.41	25.42	0.98	1.01
Jackknife IV (JIVE)	0	0	0.01	0.05	20.76	21.74	20.76	0.98	1.01
First stage: PCA	0	0	0.01	0.04	25.41	26.41	25.41	0.98	1.01
First stage: Post-Lasso selection	0	0	0.01	0.04	25.41	26.41	25.41	0.98	1.01
First stage: Lasso selection	0.01	0.17	0.01	0.04	25.08	26.41	25.25	0.98	1.01
First stage: Neural net	0.01	0.1	0.05	0.05	20.52	21.67	20.61	0.98	0.99
First stage: Neural net, narrow	0.01	0.09	0.04	0.05	20.52	21.67	20.61	0.98	0.99
First stage: Neural net, shallow	0	0	0.03	0.05	20.71	21.67	20.71	0.98	0.99
First stage: Boosted trees	0.01	0.01	0.2	0.04	25.51	26.41	25.52	0.98	1.01
First stage: Random forest, CV	0.07	1.08	0.62	0.04	24.01	26.41	25.09	1	1.01
Panel B DGP: *High-complexity* case 1									
Naive OLS	0.22	0	0.15	1.56	0.64	0.64	0.64	1	0.15
First stage: OLS	0.02	0	0.01	1.77	0.56	0.64	0.56	0.94	0.15
Split-sample IV	0.03	0	0.01	1.79	0.56	0.64	0.56	0.93	0.15
Jackknife IV (JIVE)	0	0	0	1.77	0.57	0.65	0.57	0.94	0.15
First stage: PCA	0.02	0	0.01	2.02	0.5	0.64	0.5	0.88	0.15
First stage: Post-Lasso selection	0.01	0	0.01	1.79	0.56	0.64	0.56	0.93	0.15
First stage: Lasso selection	0.04	0.02	0	1.92	0.52	0.64	0.54	0.93	0.15
First stage: Neural net, shallow	0.03	0	0.02	1.77	0.57	0.64	0.57	0.94	0.15
First stage: Neural net, narrow	0.1	0.01	0.05	1.75	0.57	0.64	0.58	0.96	0.15
First stage: Neural net	0.11	0.01	0.06	1.74	0.58	0.64	0.58	0.96	0.15
First stage: Boosted trees	0.12	0.03	0.04	1.91	0.52	0.65	0.55	0.95	0.15
First stage: Random forest, CV	0.32	0.06	0.09	2.04	0.49	0.65	0.56	0.99	0.15
Panel C DGP: *High-complexity* case 2									
Naive OLS	0.33	0	0.35	0.95	1.06	1.06	1.06	1	0.35
First stage: OLS	0.06	0	0.03	1.66	0.6	1.06	0.6	0.76	0.35
Split-sample IV	0.06	0	0.03	1.76	0.57	1.06	0.57	0.73	0.35
Jackknife IV (JIVE)	0	0	0	1.65	0.61	1.06	0.61	0.76	0.35
First stage: PCA	0.03	0	0.02	1.75	0.57	1.06	0.57	0.74	0.35
First stage: Post-Lasso selection	0.02	0	0.01	1.75	0.57	1.06	0.57	0.74	0.35
First stage: Lasso selection	0.1	0.04	0.01	2.11	0.48	1.06	0.52	0.73	0.35
First stage: Neural net	0.21	0.03	0.13	1.49	0.7	1.06	0.73	0.84	0.35
First stage: Neural net, narrow	0.21	0.04	0.12	1.52	0.68	1.06	0.71	0.84	0.35
First stage: Neural net, shallow	0.06	0	0.04	1.63	0.62	1.06	0.62	0.76	0.35
First stage: Boosted trees	0.25	0.07	0.07	1.83	0.55	1.06	0.62	0.81	0.36
First stage: Random forest, CV	0.56	0.16	0.22	1.51	0.66	1.06	0.82	0.98	0.35
Panel D DGP: *High-complexity* case 3									
Naive OLS	0.34	0	0.35	0.95	1.06	1.06	1.06	1	0.35
First stage: OLS	0.06	0	0.04	1.66	0.61	1.06	0.61	0.76	0.35
Split-sample IV	0.06	0	0.04	1.76	0.57	1.06	0.57	0.73	0.35
Jackknife IV (JIVE)	0	0	0	1.64	0.61	1.06	0.61	0.76	0.36
First stage: PCA	0.03	0	0.02	1.76	0.57	1.06	0.57	0.73	0.35
First stage: Post-Lasso selection	0.03	0	0.02	1.74	0.58	1.06	0.58	0.74	0.35
First stage: Lasso selection	0.1	0.04	0.01	2.1	0.48	1.06	0.52	0.73	0.35
First stage: Neural net	0.22	0.03	0.13	1.49	0.69	1.06	0.73	0.84	0.36
First stage: Neural net, narrow	0.21	0.04	0.12	1.53	0.67	1.06	0.71	0.84	0.36
First stage: Neural net, shallow	0.07	0	0.05	1.63	0.62	1.06	0.62	0.76	0.36
First stage: Boosted trees	0.25	0.07	0.07	1.83	0.55	1.06	0.62	0.81	0.36
First stage: Random forest, CV	0.56	0.16	0.22	1.51	0.66	1.06	0.82	0.98	0.35

A cell's value provides the given statistic's mean (**column**) in 1,000 iterations of the given combination of DGP (**Panel**) and estimator (**row**). We omit LIML as it is not a two-stage method and thus does not produce $\hat{x}$.

APPENDIX C
CH 2 FIGURE APPENDIX

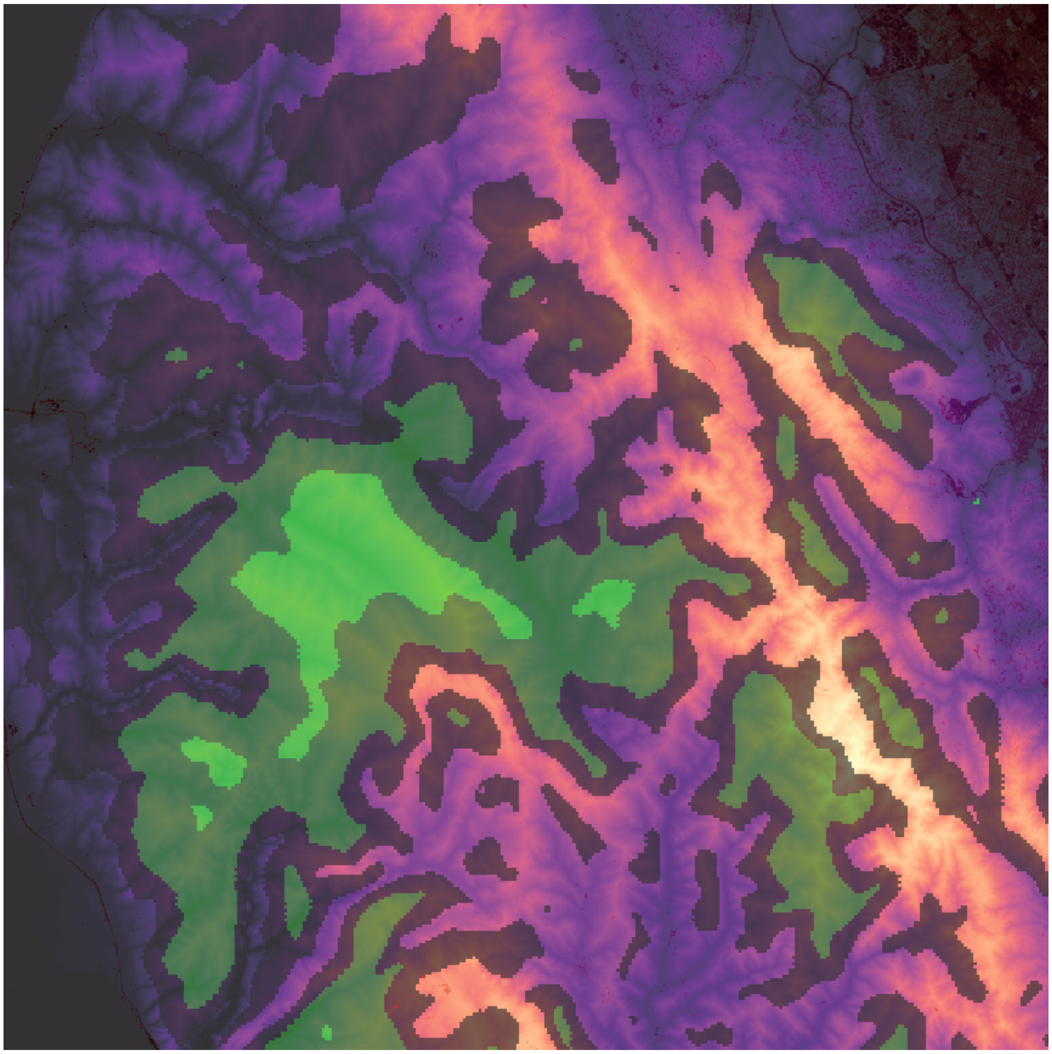


Figure C1. Example Spatial Input for Transformer: CZU Complex Fire 30 × 30 *km* An example input (OOS) that is used for a ViT model. This input is centered at the reported ignition location, though early (pre-response) VIIRS detections are included as a separate raster layer to help correct for incorrect ignition locations. It is not possible to visualize all rasters, so the image is a sample of layers available to the model. In particular, the viridis scale coloring is a hillshade DEM model using a 30x30 elevation raster, the red color layer is the NLCD Impermeable measure representing development, and the greens (black - light green) represent travel times for resources in the area where light green is the slowest speed and no shading represents as-fast-as-highway travel.

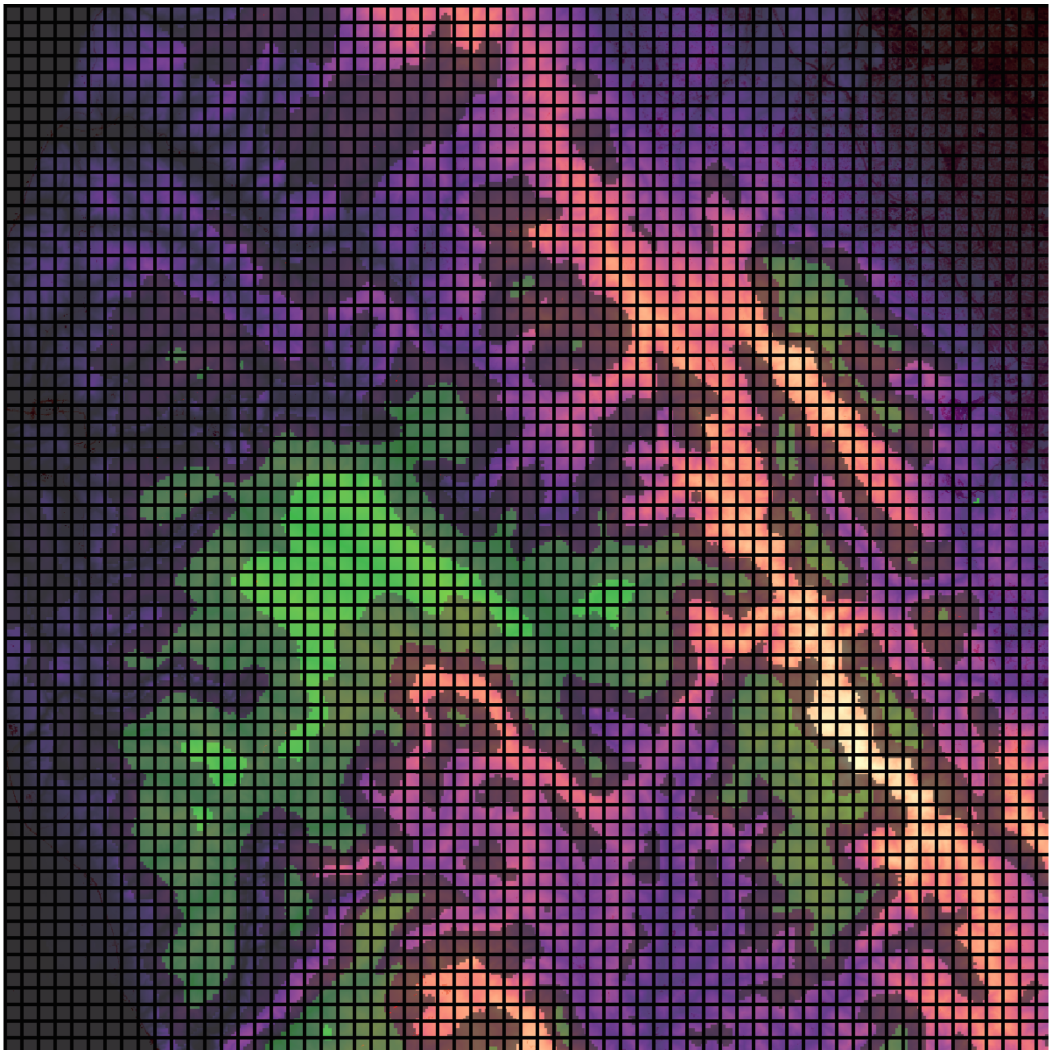


Figure C2. Example Spatial Input for Transformer, with Cut Lines
An example input (OOS) that is used for a ViT model. Lines drawn are where a single ’token’s’ bounds lie, with each square roughly corresponding to a single observation in some latent sequence. By weighting values in these rectangles and using relative placement of the token, the model is able to learn context-inclusive information about risk to structures and suppressability within a given location. This input is centered at the reported ignition location, though early (pre-response) VIIRS detections are included as a separate raster layer to help correct for incorrect ignition locations. Not all layers can be effectively visualized, so this is a sample of layers avaialble to the model. In particular, the viridis scale is a hillshade DEM model using a 30x30 elevation raster, the red color layer is the NLCD Impermeable measure representing development, and the greens (black - light green) represent travel times for resources in the area where light green is the slowest speed and no shading represents as-fast-as-highway travel.

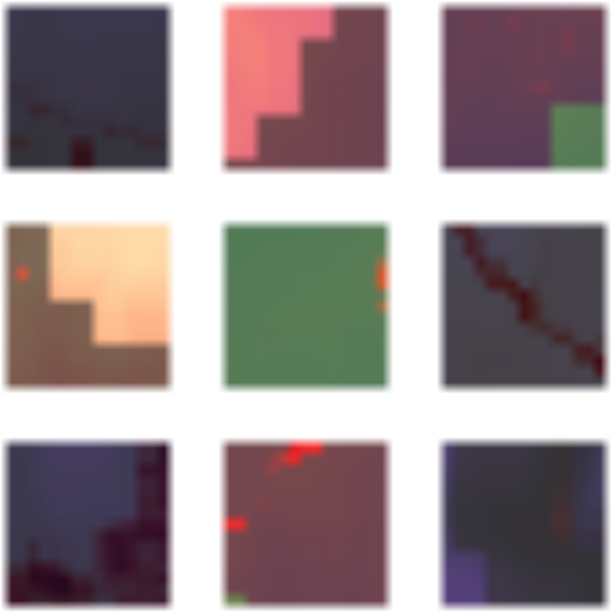


Figure C3. Example Single Patch Inputs 3×3 *km*
A single token's range of focus. This input is 9 separate sections of image from figure C2 that cover a 3x3 km area, represent a single token's focus area, and can be found spread around the image in that figure. Not all layers can be effectively visualized, so this is a sample of layers available to the model. In particular, the viridis scale is a hillshade DEM model using a 30x30 elevation raster, the red color layer is the NLCD Impermeable measure representing development, and the greens (black - light green) represent travel times for resources in the area where light green is the slowest speed and no shading represents as-fast-as-highway travel.

Figure C4. **ViT Fit** Fire Model

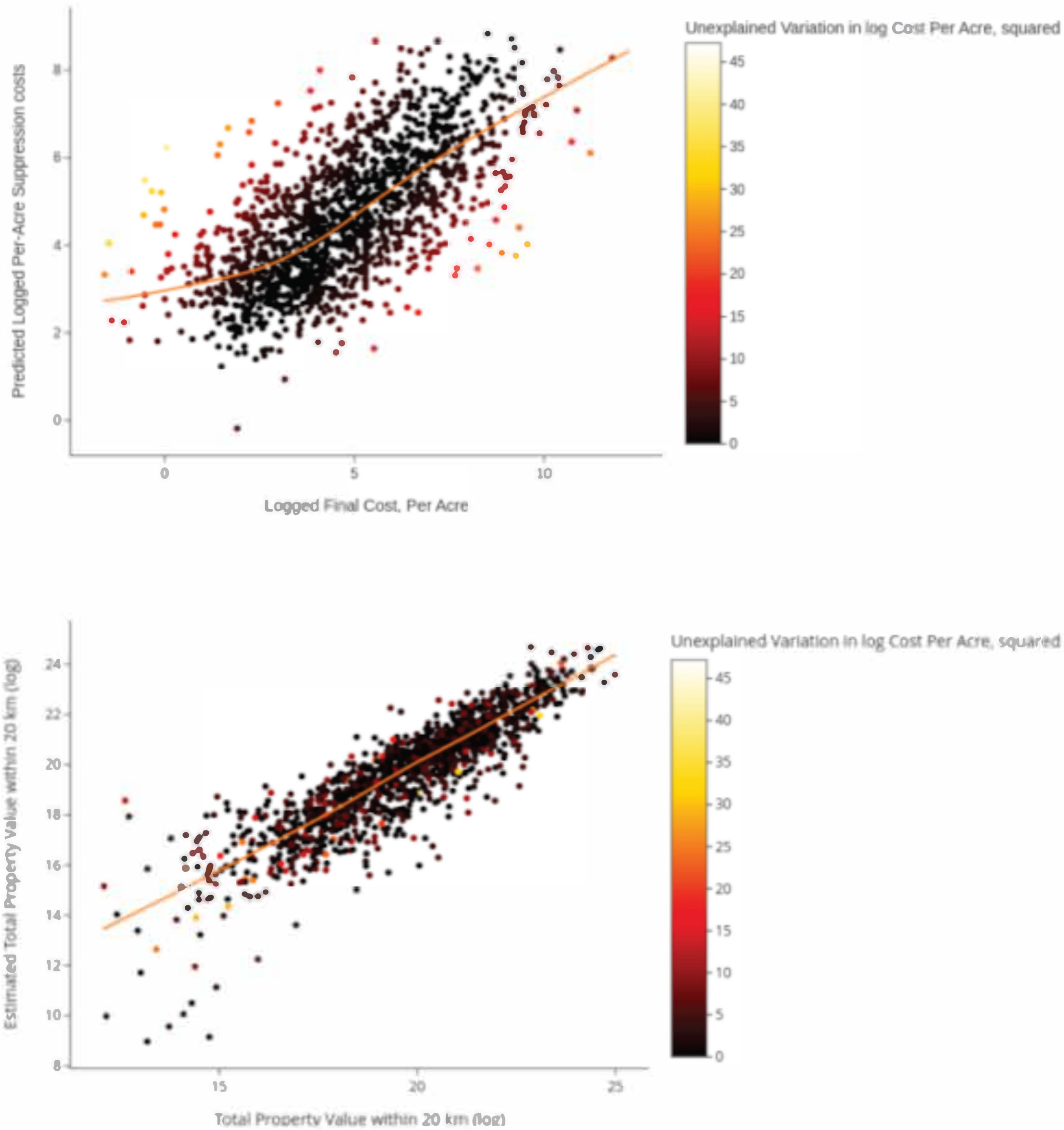
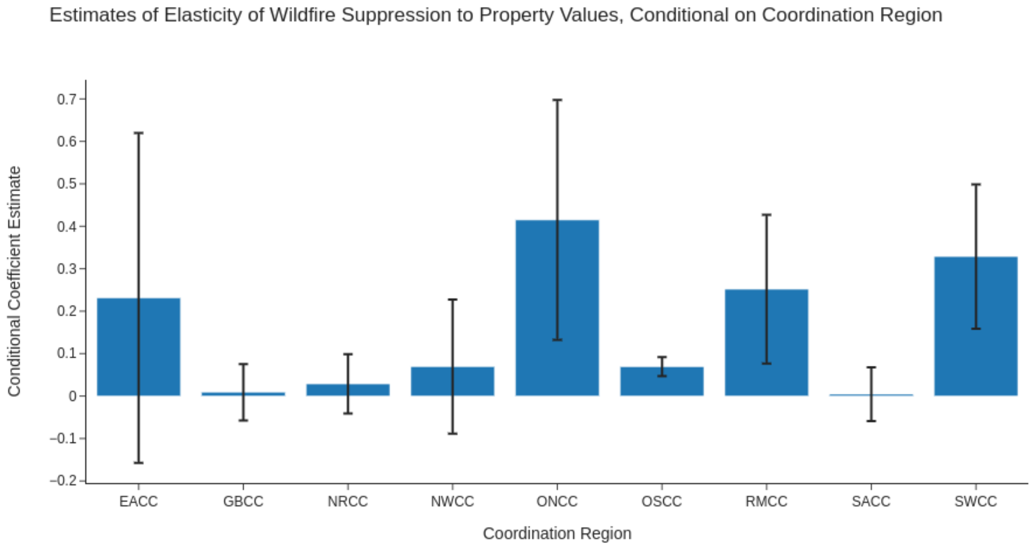


Figure C5. **Regional Estimates of**
θ : Conditional Estimates of Elasticity with Respect to Property Value



Top: Shown are predictions derived from the ViT model of per-acre suppression cost against true value of per acre suppression cost. **Bottom:**Estimates of δ_g in nonlinear system, where g represents what coordination region the fire occurred in.

Figure C6. SCM: Property Value’s Effect on Fire Suppression

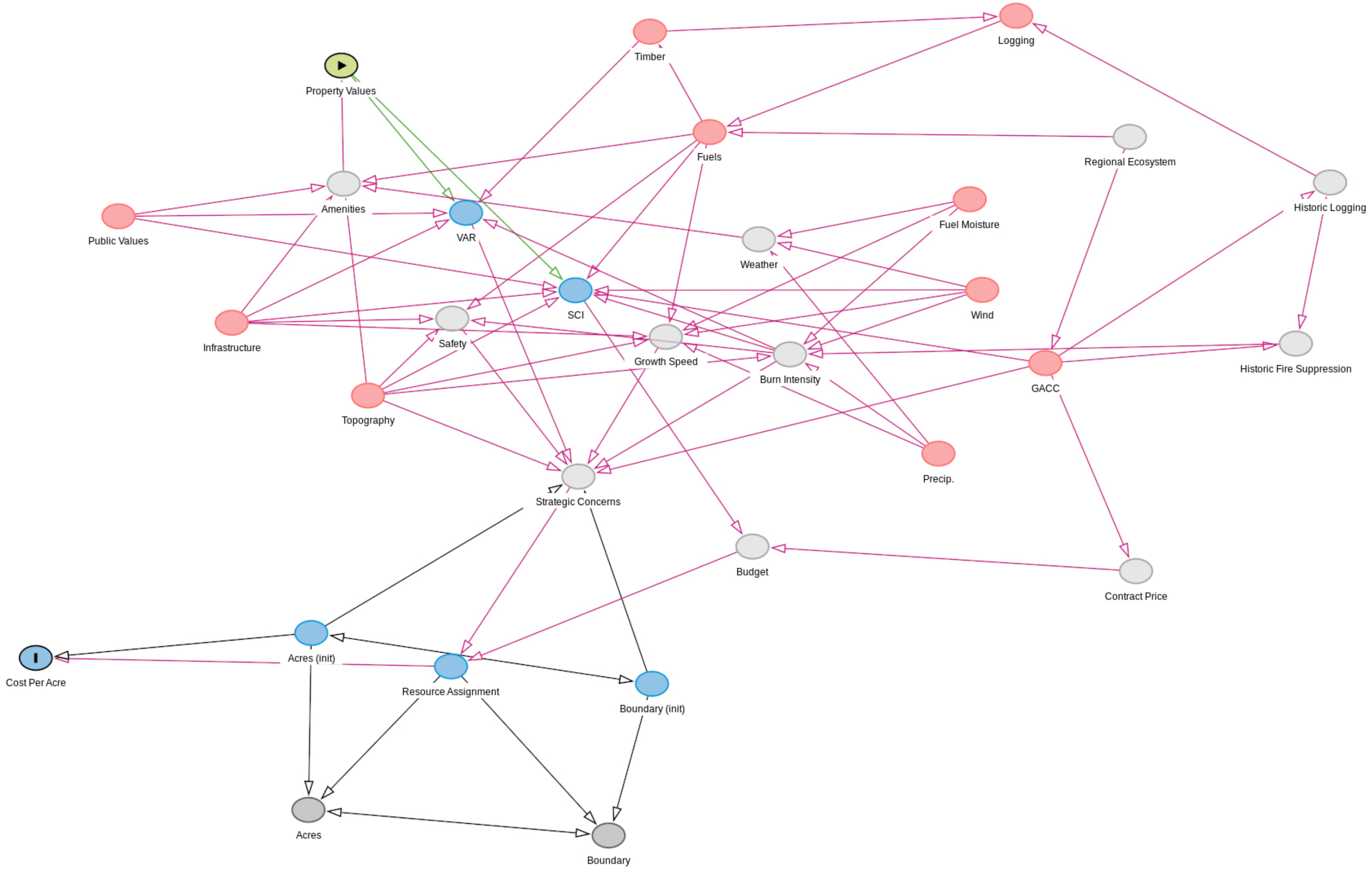
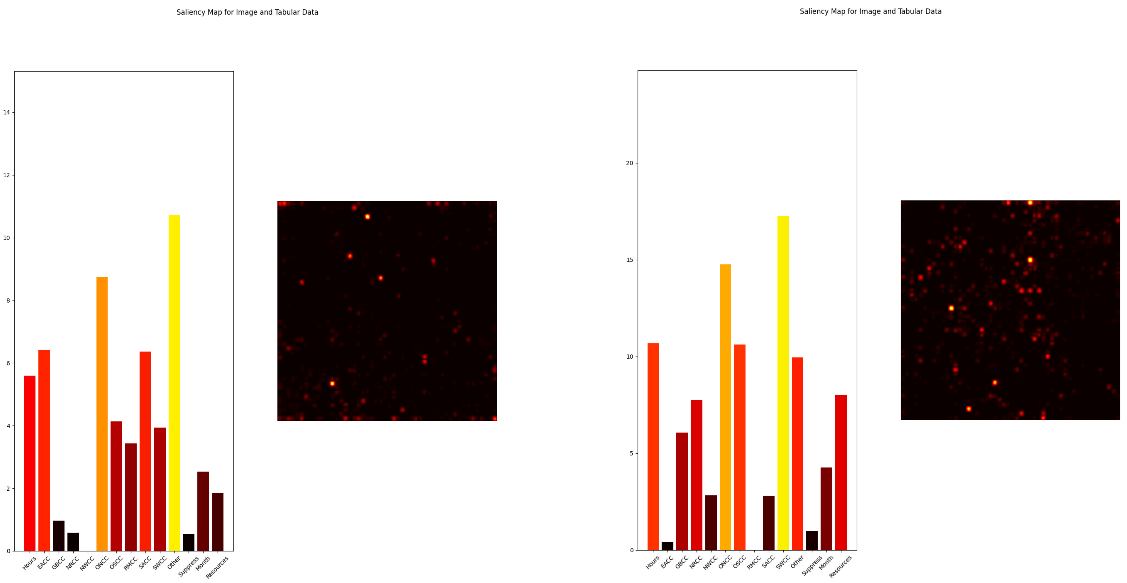
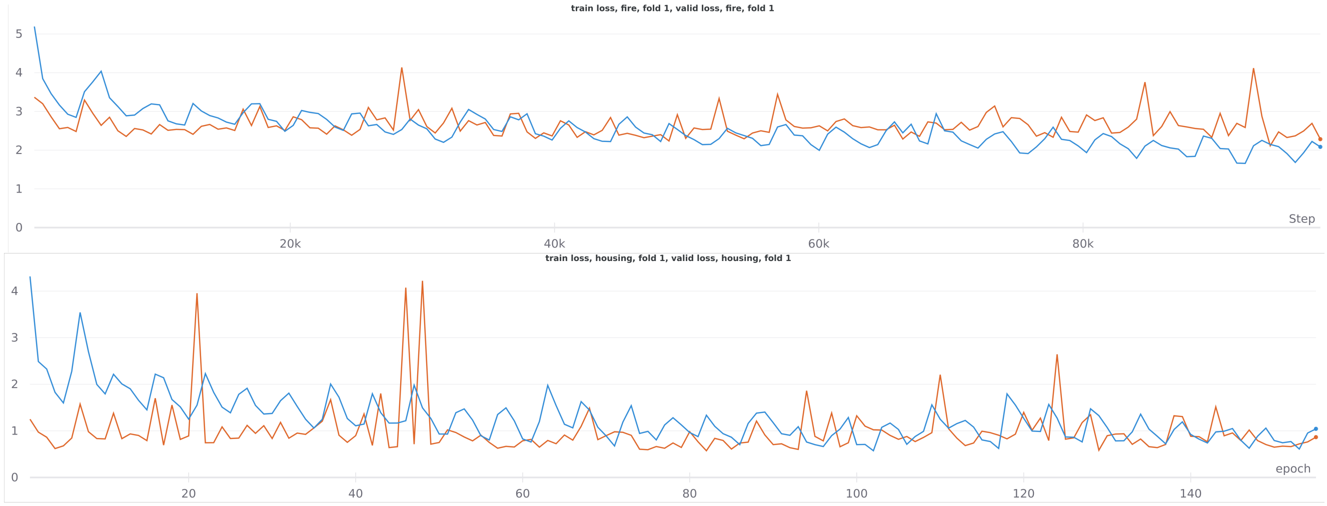


Figure C7. SCM: Property Value's Effect on Fire Suppression



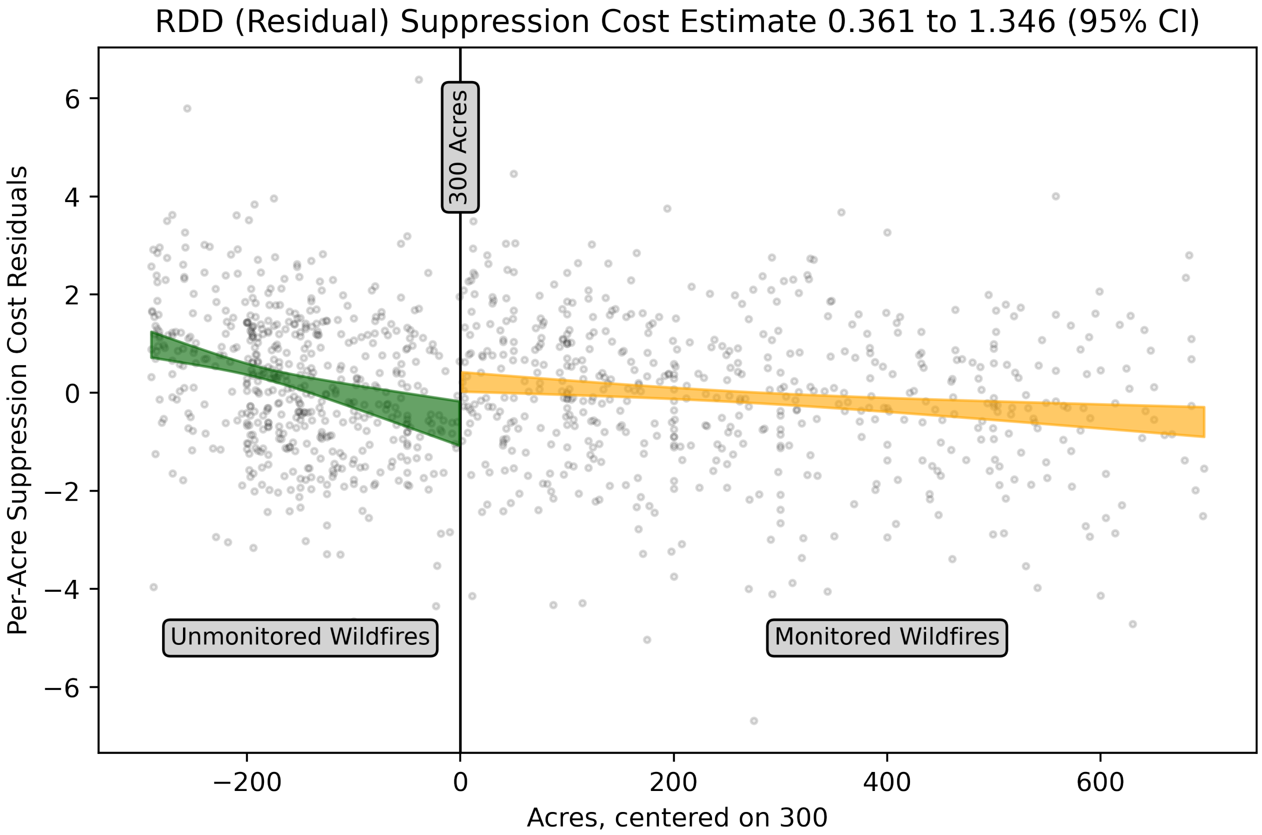
Out of sample 'saliency' maps for each model. Saliency maps identify which pixels receive the largest share of total "attention" across all 34 input dimensions. Brighter colors represent pixels where attention is paid, whereas black pixels represents relatively less attention being paid. Bar charts are gradients for the tabular inputs, and can be seen as a 'feature' importance, independent of the actual values of those features, but conditional on the values of the images. IE: IF this fire was seen in the southwest, region (GACC: SWCC), predictions for costs would be substantially different than if this fire were observed in the Northwest Region, NWCC. Left: property values transformer model, Right: fire suppression costs per acre transformer model.

Figure C8. Train/Validation Loss for Twin Transformers



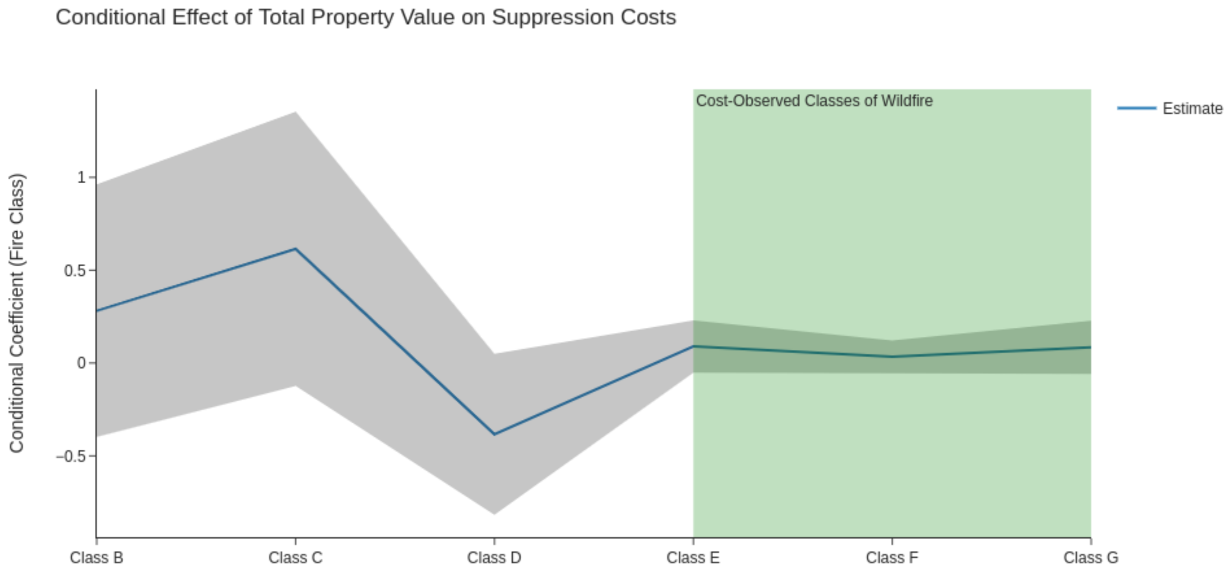
Training/Validation loss for Each Transformer, as seen in fold 1, across 155 epochs. Top: data on fire suppression costs, bottom: data on total property value. The blue line tracks in-sample, training loss while the orange line tracks out-of-sample validation loss

Figure C9. Effects of Property Value on Suppression Costs, Conditioned on Monitoring



Estimates of δ_c in nonlinear system, where c represents whether the wildfire qualifies for cost-monitoring in WFDSS, as determined by class (and therefore size) of fire. Given the estimates themselves are conditional on acres, these cannot be thought of as truly 'causal', but given the assumption of conditional ignorability, this represents a significant effect.

Figure C10. Conditional Effects of Property Value on Fires of Differing Sizes

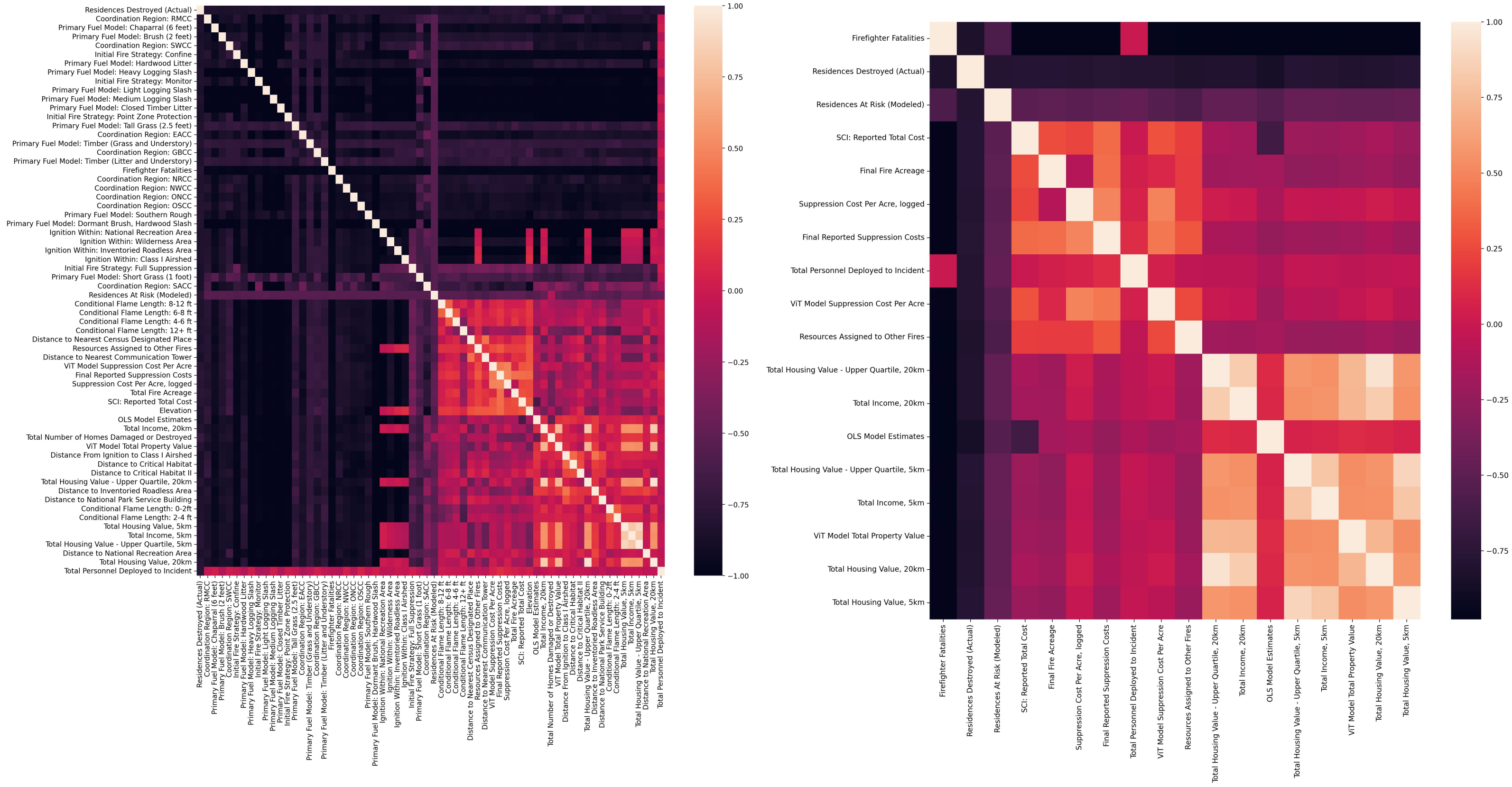


Estimates of δ_c in nonlinear system, where c represents wildfire size, by class. Given the estimates themselves are conditional on acres, these cannot be thought of as truly ’causal’. Green rectangle highlights the classes of wildfires that are eligible for cost monitoring

Figure C11. τ **Correlation A:** Non-Parametric Correlations Between Predictor Variables and Economic Variables of Interest

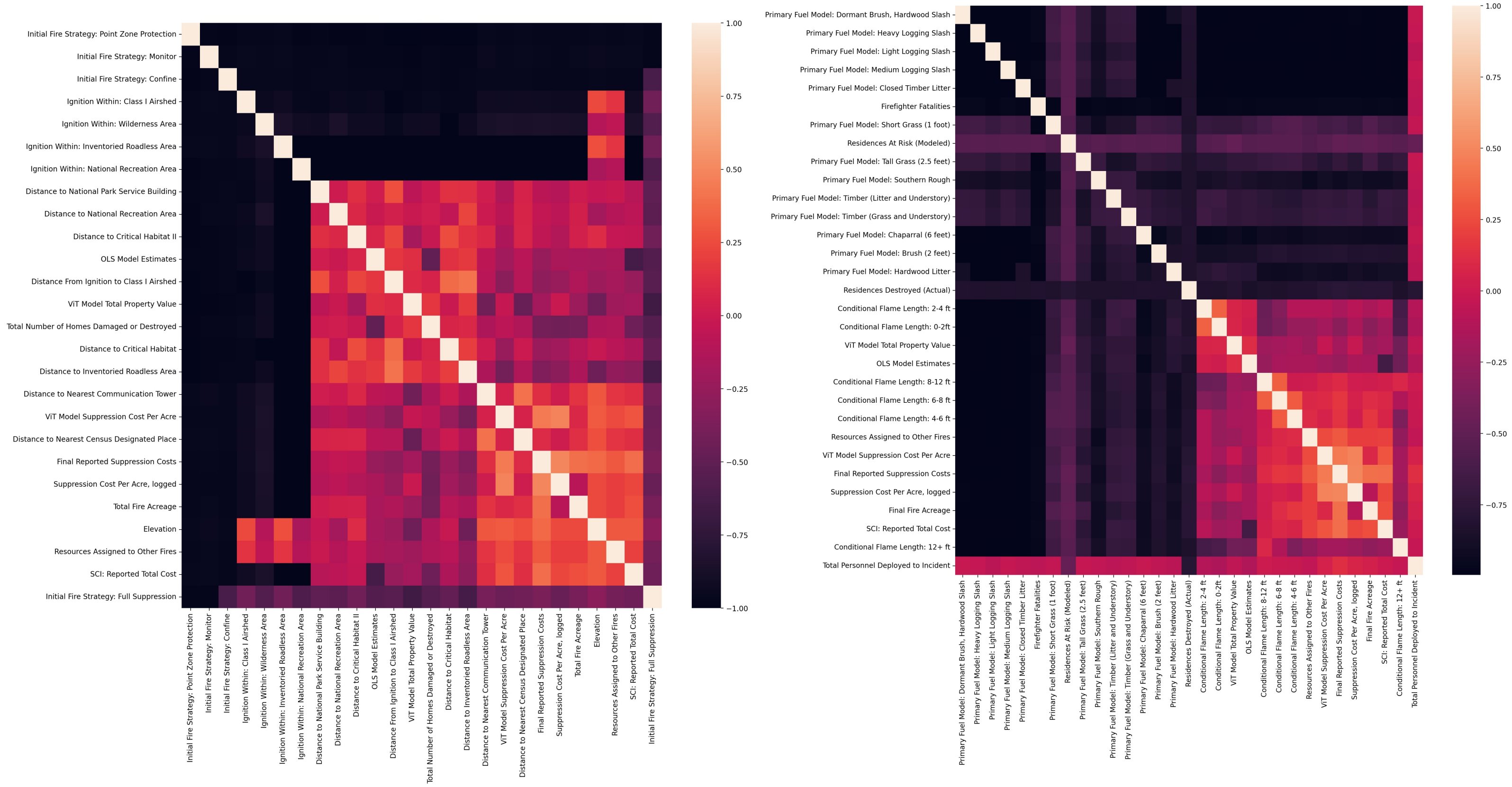
(i) **Correlation Map - Full Variable Set** Point of Ignition

(ii) **Correlation Map - Property Value/Income Variables** Point of Ignition



This figure shows the Kendall τ correlation values for differing subsets of variables and predictions. Included are variables that are represented in traditional SCI cost estimates, such as property value but also covariates of interest such as total income within 5 and 20 kilometers.

Figure C12. τ **Correlation B:** Non-Parametric Correlations Between Predictor Variables

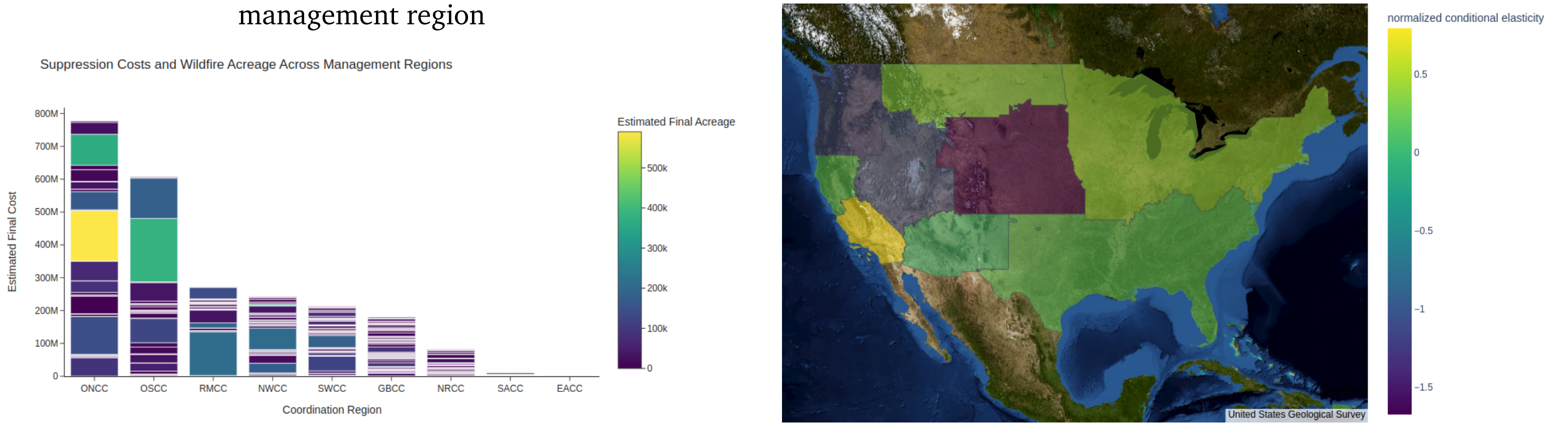


(i) **Correlation Map - Locational Information** Point of Ignition

(ii) **Correlation Map - Fuel Model Variables** Point of Ignition

This figure shows the Kendall τ correlation values for predictor variables used in wildfire cost and spread prediction and the ViT cost predictions

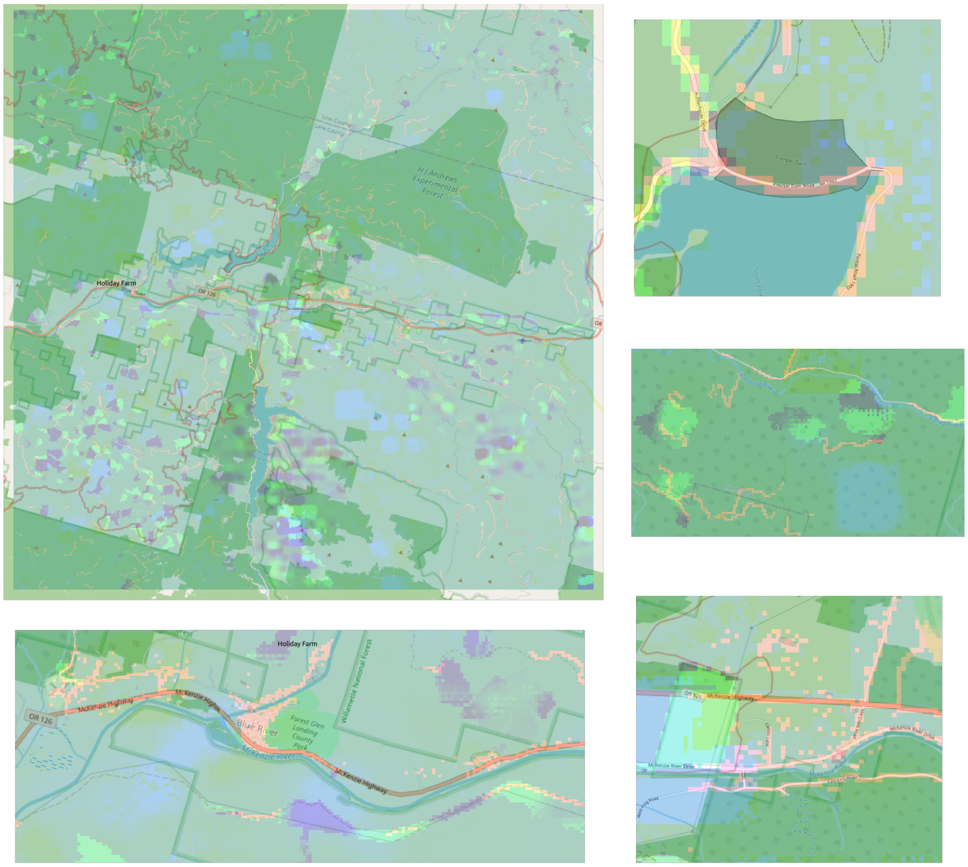
Figure C13. **Regional Statistics:** Coordination-Center Level Figures



(i) **Regional Expenditures** broken out by incident and management region

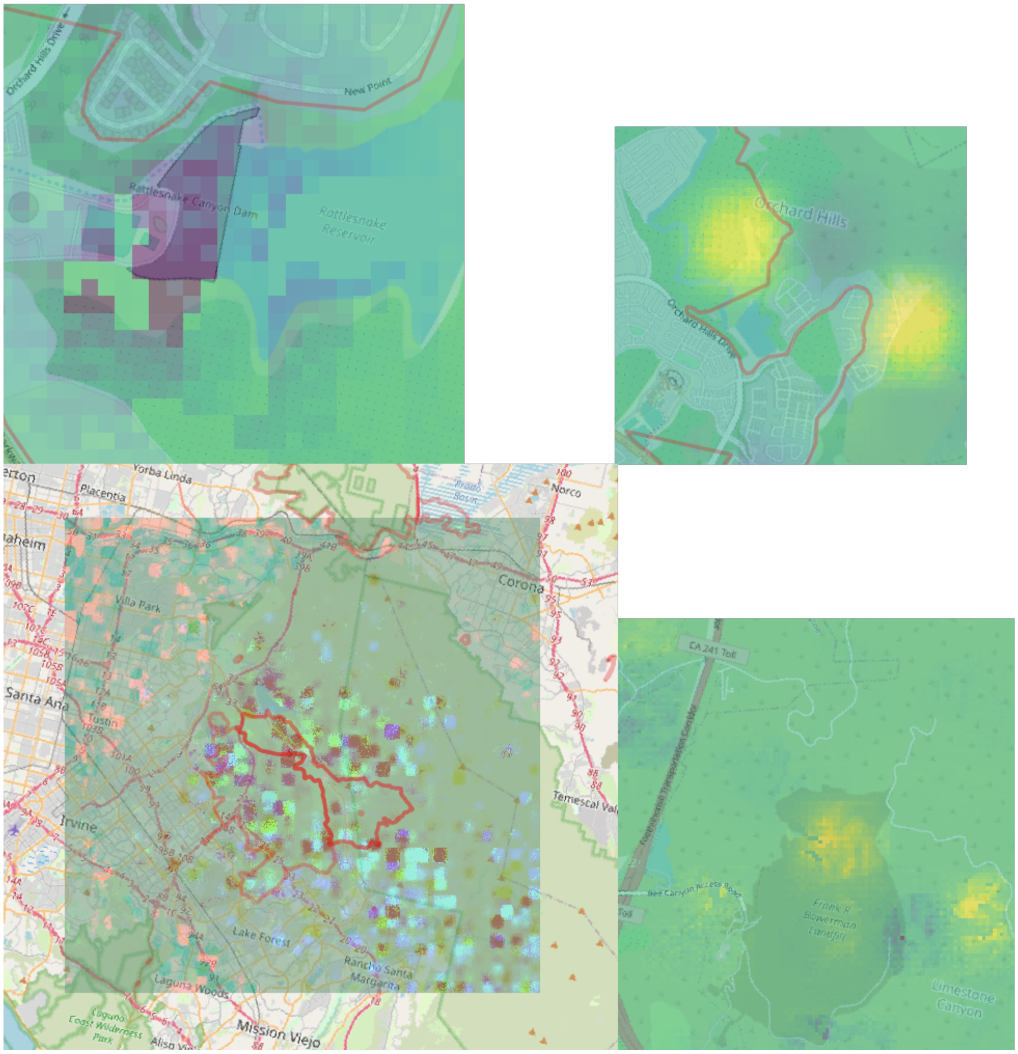
(ii) **Conditional Estimates of** θ

Figure C14. **Integrated Gradients:** Holiday Farm Fire



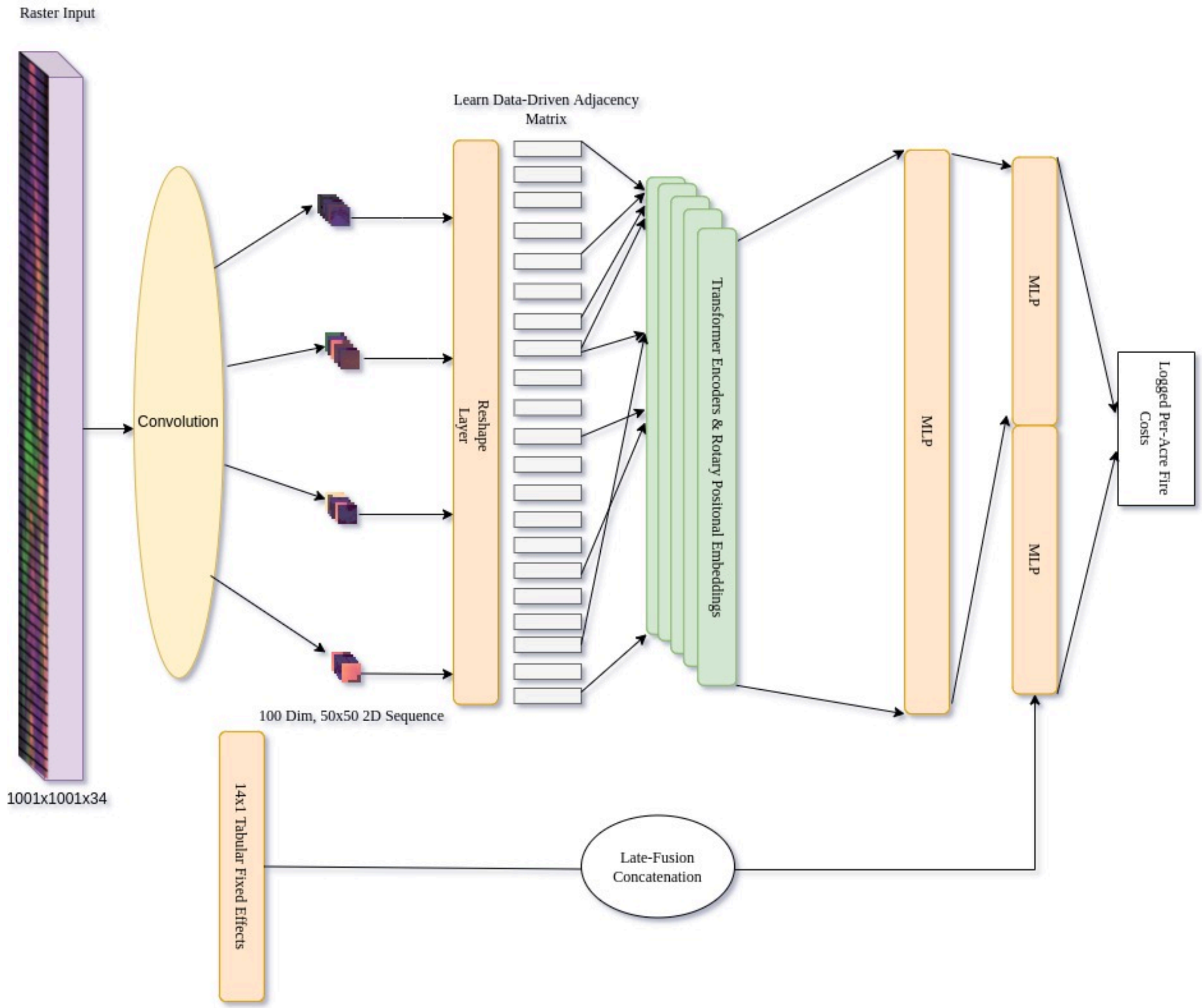
All images are compared against the baseline that consists of the ’average‘ image. **top left:** Full attribution image, focusing on elevation (blue), NLCD Imperviousness measure (red), and forest/fuel disturbance (green) **top right:** Cougar Dam gets highlighted, as well as roads near the resevoir, but dam itself is not. **bottom left:** Focus on Blue River, OR. **mid right:** recent logged patches to the north of Blue River are highlighted when within the burn perimeter. **bottom right:** Point of ignition - power station nearby is highlighted, as is the rising elevation in the direction away from the wind, but not on the east side of ignition.

Figure C15. **Integrated Gradients:** Silverado Wildfire



All images are compared against the baseline that consists of the ’average‘ image. **bottom left:** Full attribuition image, with focuses on NLCD Imperviousness (red), burn probability (green) and aspect (blue). **top left:** Rattlesnake Canyon Dam receiving an abnormally high local downweighting of Aspect - an identifier of man-made structure being a sudden change from 0 (no facing) to East Facing. **top left:** High burn probability gets highlighted when adjacent to likely suppression resouces, in this case, the fire reservoir. **bottom right:** Model uses aspect to identify the hill - a landfill where harmful gasses can often be ignited and complicate suppression efforts.

Figure C16. Compact Convolutional Multimodal Transformer


Raster Input
1001x1001x34
Convolution
100 Dim, 50x50 2D Sequence
Reshape Layer
Learn Data-Driven Adjacency Matrix
Transformer Encoders & Rotary Positonal Embeddings
MLP
MLP
MLP
Logged Per-Acre Fire Costs
14x1 Tabular Fixed Effects
Late-Fusion Concatenation

Diagrammatic illustration of the modified vision transformer.

APPENDIX D
CH 2 TABLE APPENDIX

Table E1. Data and Descriptive Statistics

Variable Name	Mean	Maximum	Minimum	Std.Dev	Source
Raster Data$^{\clubsuit}$	-	-	-	-	-
Aspect, (degrees)	103.76	359	-1	117.61	Landfire
Elevation (meters)	757	14505	-282	728.59	Landfire
Slope (percent)	.1531	1	0	.1876	Calculated
Distance From Ignition, Normed 0-1	-	1	0	-	IRWIN database (Calculated)
Vegetation Departure (percent)	63.54	100	0	32.08	Landfire
Existing Vegetation Cover (discrete)	179.38	399	11	111.15	Landfire
Forest Canopy Height (meters)	58.86	510	0	92.28	Landfire
Forest Canopy Bulk Density ($\frac{100kg}{m^3}$)	2.55	45	0	5.76	Landfire
Forest Canopy Base Height (meters)	3.70	100	0	9.49	Landfire
Forest Canopy Cover (percentage points)	18.71	100	0	29.13	Landfire
Fuel Vegetation Cover (binned)	93.45	172	11	34.05	Landfire
Fuel Vegetation Height (meters)	397.75	651	11	233.03	Landfire
Fuel Disturbance (percent)	13.45	633	0	75.83	Landfire

Continued on next page

Table E1 – continued from previous page

Variable Name	Mean	Maximum	Minimum	Std.Dev	Source
Scott and Burgan Fire Behavior Fuel Model (40 classes)	125.88	204	91	34.95	Landfire
2019 Impervious Surface Conterminous United States	.939	127	0	7.03	NLCD
Population Grid (2017)	8.30	9902	0	3.41	Huang Population Product
Communication Towers (rasterized binary mask)	.133	1	0	.3405	WFDSS
Response Time (Categorical)	2.82	7	0	1.21	WFDSS
Privately Owned Land	.9	1	0	.0459	WFDSS
Conditional Flame Length, Class 1-6 (percent)♡	-	1	0	-	Short et al. Dataset (2020)
Unconditional Burn Probability	.006	1	0	.0087	Short et al. Dataset (2020)
Protected Areas Mask	.035	1	0	.0099	PAD
Fire Radiative Power (FRP), kW	.1644	1317.6	0	5.103	VIIRS (NASA)

Continued on next page

Table E1 – continued from previous page

Variable Name	Mean	Maximum	Minimum	Std.Dev	Source
Climate/Weather Rasters[♣]	-	-	-	-	-
Soil Moisture (ml)	273.7	688.70	0	174.7	NOAA/ Gridmet
Daily Max Temperature (Celsius)	27.53	43.26	-7.61	7.735	NOAA/ Gridmet
Monthly Moisture Anomaly	-29.28	157	-251	39.62	National Weather Service (CPC)
Precipitation Rate (Past Week, $\frac{86{,}400m}{day}$)	<.0001	.0012	0	.0001	NOAA/ Gridmet
10m Wind Speed (m/s)	.0321	.1304	0	.0190	NOAA/ Gridmet
10m Wind Direction (normalized degrees)	.460	1	0	.3125	NOAA/ Gridmet
Mean vapor pressure deficit	1.53	9.83	0	0.09	Gridmet
Fuel Moisture (100 hr)	17.61	.28	33.2	21.9	Gridmet
Energy Release Component (BTU)	50.836	0	131.85	24	Gridmet
Resource/Tabular Data	-	-	-	-	-

Continued on next page

Table E1 – continued from previous page

Variable Name	Mean	Maximum	Minimum	Std.Dev	Source
Burned Acres	5982.3	589,835	1.2	29,794.7	IRWIN Database
Logged Final Cost $ln(\frac{cost}{acre})$	4.722	12.23	-1.576	2.151	IRWIN Database (Calculated)
Final Cost (Total) (1000s)	2,012.7	193,000.0	.17625	10,893.4	IRWIN Database
GACC (Regional Indicators)♡	-	1	0	-	IRWIN Database
Point of Ignition	-	-	-	-	IRWIN Database
Logged Total Property value, 20 km	19.58	25.00	0.00	2.69	Census CPS
Total Property value, 20 km (1000s)	1,879,436.8	71,734,696	0.00	4,844,174.5	Census CPS
Logged Total Property value, 5 km	13.28	21.44	0.00	6.38	Census CPS
Total Property value, 5 km (1000s)	55,888.4	2,049,296.7	0.00	176,732.5	Census CPS
Total Personnel Assigned to Other Fires at time of Ignition	2806.64	8921	3.0	2817.4	IRWIN Database (Calculated)
Initial Fire Strategy (Point Protect)	.02	1	0	0.101	IRWIN Database

Continued on next page

Table E1 – continued from previous page

Variable Name	Mean	Maximum	Minimum	Std.Dev	Source
Initial Fire Strategy (Full Suppression)	.90	1	0	0.49	IRWIN Database
Initial Fire Strategy (Monitor)	.037	1	0	0.13	IRWIN Database
Initial Fire Strategy (Confine)	.0319	1	0	0.12	IRWIN Database
Ignited in Inventoried Roadless Area (IRA) WFDSS	0.001	1	0	.034	WFDSS
Ignited in National Recreation Area (NatRec) WFDSS	.001	1	0	.034	WFDSS
Ignited in Class I Airshed	.029	1	0	.163	WFDSS
Ignited in Wilderness Area	.0606	1	0	.2386	WFDSS
Distance to Class I Airshed (km)	117.7	595.9	0	96.0	WFDSS
Distance to National Recreation Area (distNatRec) (km)	328.6	1,079.1	0	227.9	WFDSS
Distance to Inventoried Roadless Area (distIRA) (km)	89.2	544.6	0	100.0	WFDSS
Distance to Wilderness Area (km)	65.2	401.9	65.2	76.0	WFDSS

Continued on next page

Table E1 – continued from previous page

Variable Name	Mean	Maximum	Minimum	Std.Dev	Source
Distance to Census Desginated Place (km)	11.9	79.1	0	11.4	NHGIS

♣: Raster data summary statistics are estimated through random sampling
where values are not included in raster metadata.
♡: Collection of values with ranges defined by row

Table E2. Causal Equations

Outcome Name	Function	Source
Private Property (PP)	f_1(*Am, Characteristics*)	1, 2, 3, 4, 5
Amenities (Am)	f_2(Public Values, Infrastructure, Topography, *wth*, Fuels)	1, 2, 3, 4, 5, 23
Weather (*wth*)	f_3(wind, precipitation, soil/fuel moisture)	27, Assumed
fuels	f_4(soil/fuel moisture, logging, *regional ecosystem*, property locations)	12, 13, 14, 19, 24, 27, 6, 7, 8
Values at Risk(*VAR*)	f_5(Infrastructure, Public Resources, PP)	16, 9, 10, 12, 13, 21, 22, 23, 26, 27
BurnIntensity(*BI*)	f_6(fuels, topography, wind, precipitation, *Historic Wildfire*)	12, 13, 14, 15, 18, 22 , 25, 27
Strategic Concerns (SC)	f_7(fuels, soil/fuel moisture, wind, precipitation, *VAR*, *safety*, GACC, *BI*)	6, 7, 8, 9, 10, 11, 12, 13, 16, 17,18, 21
Stratified Cost Index (SCI)	f_8(*weather*, acres, *VAR*, topography, GACC, fuels)	9, 6
Initial Response Acres (ra)	f_9(fuels, topography, *wth*, *fpr*)	13, 14, 15
Initial Response Perimeter (fpr)	f_{10}(*ra*, topography, fuels)	13, 14, 15
Final Acres (fa)	f_{11}(*ra*, *fpr*, *ffp*, *raf*)	13, 14, 15, 9
Final Fire Perimeter (ffp)	f_{12}(*ra*, *fpr*, *raf*, *fa*)	13, 14, 15, 9
Historic Fire Suppression (hfs)	f_{13}(*Historic Wildfire*, *Historic Logging*, GACC)	11, 24, 25, 22
Logging(l)	f_{13}(fuels, *Historic Logging*, GACC)	11, 21
Resources Assigned to Fire (raf)	f_{14}(*SC*, *Contract Price*, *Budget*)	7, 16, 23
Cost Per Acre (cpa)	f_{14}(raf, *Contract Price*)	7, 9, 10, 16, 22
Budget	f_{15}(*SCI*, Acres, *SC*)	8, 9
Contract Price	f_{17}(GACC, Year)	6, 26
Safety	f_{18}(Topography, *BI*, *wth*)	23

Continued on next page

Table E2 – continued from previous page

<table>
<tr><th>Outcome Name</th><th>Function</th><th>Source</th></tr>
<tr><td colspan="3"><i>Italic Variable</i>: Unobserved
Non-Italic Variable: Observed
green: Treatment
blue: Outcome
1: Poudyal et al. (2010), 2: C. Mason and Quigley (1996), 3: Kain and Quigley (1970), 4: Rodriguez and Sirmans (1994)
5: Heyman et al. (2019), 6: Hand et al. (2016), 7: Hand et al. (2014a)
8: Hand et al. (2014b)
9: Gebert et al. (2007), 10: Gebert and Black (2012), 11: Busenberg (2004)
12: Alexander and Cruz (2013), 13: Finney (2002), 14: Rothermel (1972), 15: Rothermel (1983), 16: Scott et al. (2013)
17: Thompson, Calkin, Finney, Gebert, and Hand (2013)
18: Thompson et al. (2015)
19: Marlon et al. (2012), 20: Martin and Hillen (2016)
21: Jin et al. (2015), 22: Buma, Weiss, Hayes, and Lucash (2020)
23: Bayham and Yoder (2020), 24: Wibbenmeyer and Robertson (2021), 25: Marchal et al. (2017), 26: Gorte and Economics (2013)
27: Boychuk, Braun, Kulperger, Krougly, and Stanford (2008)</td></tr>
</table>

Table E3. Results: All Fire Sizes

Data Set	All GACC Regions		Western Region		Eastern Region	
Variable	**D/DML**	**Point Ignition Data, OLS**	**D/DML**	**Point Ignition Data, OLS**	**D/DML**	**Point Ignition Data, OLS**
log(total property value)	0.0114	0.1603***	0.0101	0.1974***	0.0337	
(*standard errors*)	(0.036)	(0.034)	(0.039)	(0.04)	(0.044)	(0.069)
(*lower bound, upper bound*)	(-0.041, 0.056)	(0.103, 0.220)	(-0.067, 0.075)	(0.120, 0.275)	(-0.11, 0.12)	(-0.166, 0.099)
N	1750	1750	931	931	770	770
R^2 (IS)	.93	.398	.93	.47	.93	.21
R^2 (OOS)	.859	0.003	.859	0.010	.859	-0.10
R^2 (OOS, SCI Baseline)	.56	-.25	.22	-.2	.34	-.4
Controls Used						
Aspect/Slope/Elevation	✓	✓	✓	✓	✓	✓
Forest Service Spread Model Controls	✓	✓	✓	✓	✓	✓
Private Property	✓	✓	✓	✓	✓	✓
Fuel Model Fixed Effects	✓	✓	✓	✓	✓	✓
LANDFIRE data	✓	✓	✓	✓	✓	✓
Resources Currently Deployed Elsewhere	✓	✓	✓	✓	✓	✓
GACC Fixed Effects	✓	✓	✓	✓	✓	✓
NARR/Gridmet weather variables	✓	✓	✓	✓	✓	✓
Total Impacted Homes	×	✓	×	✓	×	✓
Population Location	✓	×	✓	×	✓	×
Full Suppression Strategy Designation	✓	✓	✓	✓	✓	✓
Other Strategy Fixed Effects	×	✓	×	✓	×	✓
Month of Year Fixed Effects	✓	✓	✓	✓	✓	✓
Distances to/Ignition within...						
National Recreation Areas	×	✓	×	✓	×	✓
Class I Airshed	×	✓	×	✓	×	✓
Communication Towers	✓	✓	✓	✓	✓	✓
Inventoried Roadless Areas (IRA)	×	✓	×	✓	×	✓
National Recreation Areas	×	✓	×	✓	×	✓
Critical Habitat Region	×	✓	×	✓	×	✓
National Park Service Buildings	×	✓	×	✓	×	✓
Critical Habitat Region	×	✓	×	✓	×	✓
Census Designated Place	×	✓	×	✓	×	✓

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table E4. Results: Fires over 100 Acres

Data Set	All GACC Regions		Western Region	
Variable	**D/DML**	**Point Ignition Data, OLS**	**D/DML**	**Point Ignition Data, OLS**
log(total property value)	0.0136	0.1110***	0.0145	0.1636***
(*standard errors*)	(0.04)	(0.033)	(0.042)	(0.04)
(*lower bound, upper bound*)	(-0.056, 0.084)	(0.048, 0.175)	(-0.067, 0.088)	(0.086, 0.240)
N	1345	1345	736	736
R^2 (IS)	.90	.47	.90	.48
R^2 (OOS)	0.91	0.01	0.91	-.01
R^2 (OOS, SCI Baseline)	.46	-.34	.10	-.3
Controls Used				
Aspect/Slope/Elevation	✓	✓	✓	✓
Forest Service Spread Model Controls	✓	✓	✓	✓
Private Property	✓	✓	✓	✓
Fuel Model Fixed Effects	✓	✓	✓	✓
LANDFIRE data	✓	✓	✓	✓
Resources Currently Deployed Elsewhere	✓	✓	✓	✓
GACC Fixed Effects	✓	✓	✓	✓
NARR/Gridmet weather variables	✓	✓	✓	✓
Total Impacted Homes	×	✓	×	✓
Population Location	✓	×	✓	×
Full Suppression Strategy Designation	✓	✓	✓	✓
Other Strategy Fixed Effects	×	✓	×	✓
Month of Year Fixed Effects	✓	✓	✓	✓
Distances to/Ignition within...				
National Recreation Areas	×	✓	×	✓
Class I Airshed	×	✓	×	✓
Communication Towers	✓	✓	✓	✓
Inventoried Roadless Areas (IRA)	×	✓	×	✓
National Recreation Areas	×	✓	×	✓
Critical Habitat Region	×	✓	×	✓
National Park Service Buildings	×	✓	×	✓
Critical Habitat Region	×	✓	×	✓
Census Designated Place	×	✓	×	✓

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table E5. Results: Nonlinear Ignition-Point Models

Data Set	All GACC Regions	
Variable	**XGBoost**	**Random Forest**
log(total property value)	0.0640	0.110
(*standard errors*)	-	-
(*lower bound*, *upper bound*)	-	-
N	2779	2779
R^2 (IS)	-	-
R^2 (OOS)	0.456	0.412
Controls Used		
Aspect/Slope/Elevation	✓	✓
Forest Service Spread Model Controls	✓	✓
Private Property	✓	✓
Fuel Model Fixed Effects	✓	✓
LANDFIRE data	✓	✓
Resources Currently Deployed Elsewhere	✓	✓
GACC Fixed Effects	✓	✓
NARR/Gridmet weather variables	✓	✓
Total Impacted Homes	✓	✓
Population Location	✓	✓
Full Suppression Strategy Designation	✓	✓
Other Strategy Fixed Effects	✓	✓
Month of Year Fixed Effects	✓	✓
Distances to/Ignition within...		
National Recreation Areas	✓	✓
Class I Airshed	✓	✓
Communication Towers	✓	✓
Inventoried Roadless Areas (IRA)	✓	✓
National Recreation Areas	✓	✓
Critical Habitat Region	✓	✓
National Park Service Buildings	✓	✓
Critical Habitat Region	✓	✓
Census Designated Place	✓	✓

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

All estimates are produced from a 5-fold, twice repeated and cross-validated version of the model, where hyperparameters are chosen by grid search. Gaussian Processes and Elastic Net Models were also evaluated, but both struggled to perform out of sample ($R^2_{OOS} < .15$)

Estimates of the standard errors failed to converge for these models, and thus are not reported

APPENDIX E
APPENDIX CHAPTER 1

E.1 Appendix: Math

E.1.1 Wedge *a* covariance and correlation. Recall that $e = x - \hat{x}$, i.e.e is the first-stage prediction's residual.

$$
\begin{aligned}
\text{Corr}(\hat{x}, e) &= \frac{\text{Cov}(\hat{x}, e)}{\sigma_{\hat{x}}\sigma_e} \\
&= \frac{\text{Cov}(\hat{x}, x)}{\sigma_{\hat{x}}\sigma_e} - \frac{\text{Var}(\hat{x})}{\sigma_{\hat{x}}\sigma_e} \\
&= \frac{\text{Cov}(\hat{x}, x)}{\sigma_{\hat{x}}\sigma_e}\frac{\sigma_x}{\sigma_x} - \frac{\sigma_{\hat{x}}}{\sigma_e} \\
&= \frac{\text{Corr}(\hat{x}, x)\,\sigma_x}{\sigma_e} - \frac{\sigma_{\hat{x}}}{\sigma_e} \\
&= \sigma_e^{-1}\Big(\text{Corr}(\hat{x}, x)\,\sigma_x - \sigma_{\hat{x}}\Big) \qquad \text{(F1)}
\end{aligned}
$$

E.1.2 Appendix: MLP/Neural cross-validation procedure. Unlike many of the other methods explored in this paper, MLP (Multi-Layer Perceptrons) are difficult to cross-validate in a consistent way. This is for three main reasons.

The first reason is due to one of neural methods' advantages for prediction problems—that they are highly adaptable to many different problem spaces, varying both in more traditional hyper-parameters such as learning-rate and neural network width, but also in much-more nuanced choices such as optimization method or input structure. "Neural Networks," despite the term's usage in many settings, is actually less of a single model and more a label placed on an entire class of iteratively-optimized models. The work's aim has been to use "off-the-shelf" machine learning methods to understand what empirical concerns exist when placing these models naively in an otherwise-recognizable econometric instrumental variables setting, but for a neural network, the off-the-shelf model is highly dependent on the problem at hand. Unfortunately, this advantage of MLPs and other Neural Networks makes a full grid-search of the hyper-parameter space intractable. This requires us to restrict the grid-space somewhat to create a tractable solution, while allowing the model a good shot at choosing the "correct" specification. This restriction potentially handicaps the neural model's flexibility, and may produce higher average out-of-sample loss than the full set of model specifications could potentially produce.

Second, neural networks are computationally expensive to train—each combination of hyperparameters must be trained separately over many iterations, and for most optimization procedures it is useful to utilize different re-orderings of the data-set to reach a satisfactory loss-minimizing point. Even for less-complex data such as ours with relatively few observations, cross-validating even 100 hyperparameter combinations over 1000 synthetic datasets leads to prohibitively long training periods given our computational resources.

Last, neural networks are sensitive to their initialization point, which is not the case for any of our other methods included in this analysis. Unlike most other methods, neural-class models can find values for one of a number of potential loss-minimizing local optima for its parameters, and while those local optima can perform similarly well, they do not necessarily produce identical predictions and can fail on different subsets of data in different ways.

These differences make analyzing how cross-validating a neural network dictates hyper-parameters given a reasonable loss function more interesting because the choices a five-fold cross-validation approach might make are indicative of how the most flexible model chooses a specification given different search spaces. For our cross-validation, we use a five-fold cross-validation procedure to match our other methods, and use mean-squared error as our loss function. We fix a few hyperparameters in place—we use no regularization on the weights, and use an "Adam" optimizer with it's out-of-box/off-the-shelf learning rate of .001. We trained all models over 40 epochs, and used a batch-size of 10 observations. One unusual step we take is to introduce a leaky rectified linear activation function (ReLU) activation function to connect hidden layers. This was done to prevent the model from suffering from "dying weights" which is when parameters accidentally force a large number of activations to inappropriately "ignore" activations due to $a(x) = 0$ for all or many values of x.[1]

We began by creating three separate hyper-parameter search spaces distinguished by maximum-allowable width and depth. The 'shallow/wide' neural networks are allowed to choose from hidden-layer representations that are 16, 32, 64, 256, and 512 nodes in hidden layer width respectively. This model is then restricted to contain at most a single hidden layer, but is also allowed to choose from a model that maps inputs to outputs directly, using a linear activation function. This functional form is sensible given x is linear in z for our DGPs. Excluding the 0-hidden layer case would prevent the cross-validation procedure from finding the easiest approximation for a linear functional form. The cross-validation procedure allows differences in regulation by choosing between a dropout rate of .1 or .2.

The 'narrow/deep' neural network is instead allowed to choose from a representation with two, three, four, or five hidden layers each with a number of

---

[1] ReLU was tried initially, however using ReLU on our data seemingly led to a complete shutdown of the predictive power when we attempted it on our weaker-instrument setting. Even using dropout, the MLP's performance was poor in predicting out of sample. We cannot be certain that the dying weights problem was the cause of this issue, but adding a slight negative slope for activations less than 0 seems to have mitigated the behavior.

nodes (width) equal to 16, 32, or 64. This model, too, is allowed to choose from the simple linear mapping of inputs to output, for the same reasons as above.

The last search space, referred to as the "unrestricted" neural network is allowed to choose from any combination of the hyperparameters offered to the narrow or shallow networks—from zero through five hidden layers and using the full complement of widths.

To get a sense of how these search-spaces choose models on average, these three search spaces were used to cross-validate over our first 25 datasets, from which we generated a full list of 125 folds. These cross-dataset folds were then used to find, for each search space, the average out of sample MSE across all 125 folds.

From the set of available models available to every search space and for each iteration, we chose two models at random weighted by their average out of sample MSE. These probabilities were chosen using the weighted upper-tail normal CDF, normalized such that all weights for a given search space sum to one. Formally, where i is a given set of hyperparameters, j is a search-space and μ_j is the mean out of sample MSE for a given search space:

$$p_{i,j}^{chosen} = \int_{z_{i,j}^{mse}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} dz \, , \tag{F2}$$

$$z_{i,j}^{mse} := \frac{mse_{i,j} - \mu_j}{se(mse_j)} \, . \tag{F3}$$

For each iteration and using the probabilities above, two models are chosen at random for each search space, and then cross-validated again using a 5-fold procedure. The "winning" model is chosen by lowest out-of-sample MSE, and is used to predict $\hat{x}$ for the first stage.

For a visual explanation and overview of the results from the selection method, see Figure A6.

E.1.3 Low-bias methods. Because machine learning algorithms are designed to minimize loss (maximizing fit), the fact that $\mathrm{Cov}(\hat{x}, e) \neq 0$ is partially by design. To see this fact, consider any prediction method that minimizes mean-squared error (MSE)—conditional on the training data $\{x, \mathbf{z}\}$:

$$\mathrm{MSE}(\hat{x}, x|x, \mathbf{z}) = \underbrace{\Big(x - \mathrm{E}[\hat{x} \mid x, \mathbf{z}]\Big)^2}_{\big(\text{Bias of } x \text{ for } \hat{x}\big)^2} + \underbrace{\mathrm{E}\Big[\, (\hat{x} - \mathrm{E}[\hat{x} \mid x, \mathbf{z}])^2 \,\Big|\, x, \mathbf{z} \,\Big]}_{\text{Cond. Var}(\hat{x})} + \underbrace{\mathrm{E}\Big[\, \varepsilon^2 \,\Big|\, x, \mathbf{z} \,\Big]}_{\text{Cond. Var}(\varepsilon)} \tag{F4}$$

where ε is the irreducible error—the unknowable disturbance from the DGP of x, i.e.$x = f(\mathbf{z}) + \varepsilon$.[2]

Equation (F4) highlights that in an MSE-minimization problem, $\hat{x}$ is the only component of MSE that a learning algorithm can change (x, $\mathbf{z}$, and ε are all data dependent). This fact leads to the widely discussed variance-bias tradeoff—an arbitrary estimator will generally face a negotiate between low-variance predictions and low-bias predictions. Many traditional econometric estimators result from prioritizing zero bias and then selecting the minimum-variance estimator from this class of unbiased estimators. As (F4) points out, these estimators could reduce their out-of-sample MSE by *accepting* some bias and reducing variance. This tradeoff is at the heart of the prediction improvement many out-of-the-box algorithms offer relative to plain OLS.

E.2 Appendix: Neural approaches to measuring causal response

Neural networks and their offspring offer just such a route to explore data that falls outside of the traditional bounds, whether that is to use transformers for text data, or convolutional neural networks (CNNs) for imaging. However, many of the same problems apply to these tools as were outlined in the paper. In order to make full use of them in a two stage approach, the same stringent restrictions are required to generate meaningful and unbiased coefficients in a two-stage framework. Indeed, when running a cross-validated feed-forward network to produce a meaningful first stage estimation with our high-complexity data most simulations with any hidden layers simply reproduced an approximation of β close to that of naive OLS.[3] There is a burgeoning field of research in machine learning that strives to understand IV problems under less-parametric (though generally still somewhat parametric) causal structures and these methods are seeing success in both simulated and real-world data. The downside to using these powerful methods is that they require a new framework in which to understand them, and make interpretation of treatment effects more challenging.

The first of the recent batch of machine learning instrumental variables papers is referred to as "Deep IV" Hartford et al. (2017). The authors throw away the linear functional form for $x = f(z, u)$ in the "first stage," but assume linearly additive confounding variables and learn the causal structure with a two-part one-pass neural network model. The authors do this by recasting the econometric approach to instrumental variables into two interlinked problem spaces - estimating the conditional distribution $g(x|z)$ and then using the approximation of x given z to predict y. This creates difficulties because such methods produce good counterfactual predictions, but have a harder time matching the clean interpretable causal effect of X on y when compared to traditional econometric

---

[2] These expectations are conditional on the given dataset; ε and $\hat{x}$ are conditionally independent by definition. The expectation term is conditional on data observed, so for simplicity, the term $E_D(\hat{x}(z; D))$ will simply be referred to as $\hat{x}$.

[3] See Appendix Section E.1.2 for full details of the MLP methods.

approaches. Further, because of the flexibility in functional form, models of this category tend to have more trouble outside of $supp(z_{train})$ or $supp(g(x|z_{train}))$ that are observed in a training sample - and it's difficult to apply a post-analysis structure to such a model to gather understanding on counterfactuals where z_{test} is considerably different than z_{train}. Further, many other methods have been created and can be used to estimate instrument-identified causal effects using a similar semi-parametric two-stage function that can identify complex functional forms in either first or second stages Bennett et al. (2020) and Xu et al. (2020) and improve on edge-of-support marginal effects.[4] Both of these papers and Deep IV are able to produce causal inferences using images as instruments—something that a regression would not be able to meaningfully do without some form of pre-model dimensionality reduction.

Neural approaches to causal inference are also not limited to use semi-parametric structural forms for heterogenous treatment effects. Kilbertus et al. (2020) created a neural network to identify the total set of conditional causal effects given a fully non-parametric instrumental variable analysis. (With the very reasonable assumption placed on the function of unobserved noise that it does not feature infinite discontinuities, for example.)

In spite of the massive technical improvements these models have made, trying to extract a beta-equivalent from the existing models is difficult, though interpretable machine learning methods do exist. Unfortunately, extracting meaningful information using prediction-explanation methods such as Ribeiro, Singh, and Guestrin (2016) about how a result is produced, or to infer what kind of economic information can be gathered from the weights within a model, neural networks are hard to interpret Wang, Wang, and Zhao (2019). This makes comparing such models to traditional 2SLS or econometric approaches for ATE approximation difficult—and produces complications in choosing benchmarks as to how to evaluate them.

E.3 Appendix: The complications of machine learning and the monotonicity assumption

One often overlooked assumption in instrumental variables irrelevant under the assumption of constant treatment effects, but, without guaranteed constant treatment effects, is referred to as 'monotonicity' of heterogenous treatment effects. If a researcher is willing to simply find the 'Local Average Treatment Effects' or LATE J. Angrist and Pischke (2009), 2SLS can under certain circumstances recover that estimate. Simply put, this means that while treatment effects can vary across our population, response of endogenous variable x_i to instrument z_i must move in the same direction for all individuals i. For canonical

---

[4] The methodology contained in Xu et al. (2020) is particularly useful, because it does not predict X directly, but rather applies Neural Networks to the task of learning polynomial forms to pass through first and second stages in a 2SLS (with an L2 penalty) and may mostly avoid components a and b as described earlier.

cases of binary instruments and treatments, this boils down to an assumption of 'no defiers'.

Our results as written do not conclude one way or another about the implications of the monotonicity assumption with regards to the machine learning in the first stage, as our primary datasets feature homogenous treatment effects are constant and equal to β. However, relaxing from constant treatment effects to continuous, monotonicity-preserving and heterogenous treatment effects, under some circumstances using nonlinear methods can lead to inefficiency in estimates of the LATE.

To illustrate a very simple example, imagine a case estimating the coefficient β_1 where $x_i = 1 + \sum_{v=1}^{7} z_i * \gamma_i + \varepsilon_i + u_i$, where $\gamma_i \sim \gamma(.5, 4)$, thus $E(\gamma_i) = 2$, $min(\gamma_i) = 0$, $Y = 1 + X_i\beta_1 + u_i + \epsilon_i$ and all error terms $\epsilon_i, \varepsilon_i, u_i$ $N(0, 1)$. This extends the strengthened exclusion restriction from $E(z|\varepsilon) = 0$ to $E(z|\varepsilon) = E(z|\eta X) = 0$ where $\eta = \gamma_i - E(\gamma_i)$.[5] In this case, the instruments are not weak, but have weak effects on varying members of the sample. If a machine learning algorithm conditions its predictions on the instruments directly, and the coefficients γ_i are sufficiently varied, $x_{\hat{ssml}}$ will only be a better estimate in expectation.[6]

In this case; under the stronger exclusion restriction described above and using a MLSLS strategy as described in J. Chen et al. (2020) with a non-linear meta-model, will produce an unbiased result, but also may result in inflated standard errors relative to a 2SLS or SSIV approach. This is because, under this kind of monotonicity, linear IV becomes a weighted average of marginal treatment effects.[7] If the econometrician believes this type of variation exists, the assumptions required for 2SLS' validity are strong, and MLSLS' are stronger. ML makes no guarantees on recovering this weighted sum in the same manner, so there exists important future work to examine how exactly this may impact structural estimates of β.

---

[5] See J. Heckman, Urzua, and Vytlacil (2006) for the original treatment of η in the IV case.

[6] See A7 for a simulation of this. As noted by J. Chen et al. (2020) in appendix D, the asymptotic results for MLIV are not directly comparable here.

[7] which the reader should turn to J. J. Heckman and Vytlacil (2005) and J. Heckman et al. (2006) to see a formal treatment of.

APPENDIX F

APPENDIX CHAPTER 2

F.1 Model Training and Hyperparameter Selection

The conditional ViT models were trained using Pytorch, on a single Nvidia RTX3090 GPU. Initial batchsize was selected following the recommended protocol: 'as big as can fit in memory', which resulted in batch sizes of 5, and 2 for the initial model and final model respectively. Accumulating gradients across multiple steps, actual batch size b was defined as $b \in \{8, 16, 32, 64, 128\}$, and was considered one of the many hyperparameters to select from in the model.

The ADAMW[1] optimizer was chosen and fixed to update the weights, to allow for 'true' L2 normalization as described in the work that proposed the technique. Then, hyperparameter sets were selected to be depth of transformer, transformer head size, convolutional kernel size, study area, multi-layer perceptron latent dimension, as well as two versions of the class token.

F.1.1 Hyper Parameter Selection. A single out of sample test was performed using grid search over the combination of hyperparameters above, and found a depth of 5 sequential encoder layers to be marginally better than the largest transformer (6 layers). By far the most important hyperparameter choices were learning rate and batch size. Batch sizes of 16 and 32 performed very well, while batch sizes of 8 or less tended to be unstable during training, and batch sizes of 64 or 128 overfit the training regimin, and led to poor out-of-sample performance. Learning rates, selected from $\{1e^{-5}, 2e^{-5}, 3e^{-5}, 5e^{-5} and 1e^{-4}\}$, tended to prefer smaller learning rates, even with the cosine annealing learning rate scheduler.

The convolutional kernel also proved critical, and an unusually large kernel was selected, 13×13 appeared to be the most performant kernel. It's unclear whether this is because there are important variations in texture over larger areas, or if the model would be well served to add a second convolution layer to the encoding step.

F.1.2 On instability. Many times, due to mechanical failure, power loss, or unnoticed memory leak, the training was interrupted for both the selection models and also the final evaluation at otherwise random intervals. The author did not set up a stateful dataset function, meaning once the model is restarted from a save point it would be using the same data in the same groups in the same order, which can give misleading estimates of performance. To get around this, when the model failed, the initial seed was multiplied by the last successful epoch, re-randomizing the ordering of the data.

---

[1] as proposed in Loshchilov and Hutter (2017)

F.2 Integrated Gradient estimation of Cost-Important Factors in Selected Fires

Integrated gradients are a common approach to get pixel/channel-level importance of features in natural images. A good approximation of the non-linear 'importance' can be calculated by

$$\frac{\partial f}{\partial x} * x$$

However, in cases where the gradient is saturated for variable x, it can mislead readers into believing features are not important when they have diminishing responses. Given a 'baseline' image represented by a $H \times W \times C$ matrix, whose activation is believed to be 0 across all classes, integrated gradients calculate the integral from activations at that 0 point to the observed image.

Unlike typical applications of the technique, activations of '0' are actually quite challenging to find. This is because, after normalization, an activation of 0 represents the exact average-costed fire, conditional on ignition.

Several images were tried to find this 'true' 0 image for this study, but the one that worked is the 'average baseline', which found the pixel-level average of all images in the set and combined them. i.e.for $X \in \mathcal{X}, X_avg = \frac{\sum_{i=0}^{n} X}{n}$.

These integrated gradients *appeared* impressive, and can be interpreted as the model doing substantially well at identifying features that would lead to risk, as well as the likelihood that they would be at risk, given the ignition, it's equally likely given the size of the image that this represents apophenia bias.

That being said, the results are intriguing, and thus they were included in the figures here to examine. Used in these figures are two wildfires from the 2020 season whose boundaries and events are validated an well understood, first, the Holiday Farm Wildfire in Lane County and second, the Silverado Wildfire near Irvine California. These incidents were well documented, and controversy/challenges were covered well. The model appears to align with reports and information contained in incident command logs (aside from considerations arising from smoke, which the model is unlikely to be able to find) all of which are publicly available on inciweb.

Results from these estimations can be seen in figures C14 and C15.

F.3 Out of Distribution Performance

The dataset was reconstructed over the early months of 2022 (Conducted exactly on data drawn at 04-22-2022) to include all wildfires occurring from December 30th, 2019 to March 30th of 2022 to collect a new set of fires whose properties would closely resemble the actual forecasting problem at hand.

Using fires from 2022, which had a number of abberant, out of season wildfires, a new training/testing split was created to determine if the fire cost prediction model would; without modification, generalize to new types of damaging wildfires, as well as fires burning under seasonal-weather-regional combinations unseen in the training data .[2]

Once the old model was refined and tested, a new model was developed to determine how much this performance could be improved by modifying hyperparameters, both those associated with the model directly but also those attached to the data generation process. Notably, increasing individual ’study area’ images from 15km × 15km to a larger field of vision, 20km × 20km while maintaining similar resolution in the latent space - ie, 1001 × 1001 becomes 1334 × 1334 after the convolution step.

Despite improving the performance of the model in both the property value and suppression cost per acre cases, estimates of the effect of property value on suppression costs are nearly indistinguishible from those generated from the baseline model, and come at the cost of being generated after seeing initial results. Thus, the original model estimates are provided to minimize the author’s bias from creeping into the engineering choices.

The results for the fire cost portion of the model are fairly consistent with the original model, with one notable exception - out of sample performance improved in Californian regions. This makes some sense, as Californian wildfires tend to be, on average, larger (in acreage terms), and typically burn for longer and thus potentially threaten communities further away from the point of ignition. For both of these reasons, and thus having access to pixels further from the ignition point is likely to be more valuable

F.4 The Non-linear model Hypothesis (testing other regression models)

One possibility is that the immense computational power required to train the ViT model in this work is not worth the effort, and that comparable if not higher quality results can be produced by point-level data, simply used nonlinearly.

Using the expanded 2022 dataset from the prior section, several tree based models were crossvalidated and tested for out of sample performance to check the hypothesis.

Gaussian processes, often cited as a go to solver for nonlinear regression problems, performed only marginally better than OLS estimates or regularized linear models (which themselves improve on OLS estimates), but the tree-based methods performed better than either. It is unclear why the gaussian process regression performed poorly, despite typically being the best nonlinear regression solver available for datasets of reasonable size. It is possible that a different kernel than the traditional RBF kernel would improve estimates given the outcome rejects on tests for being normally distributed, though it is unclear how much room there is for improvement. More likely, the number of covariates makes

---

[2] A criticism one might level is that given month of year and weather patterns, along with sufficiently strong spatial fixed effects could uniquely identify wildfires and thus map these combinations to their exact per-acre costs.

learning the gaussian process sufficiently well to perform well on the observed outcome while simultaneously conditioning out variables required to block causal pathways provide too much flexibility to a gaussian process while also being able to perform well out of sample.

These models were able to improve markedly on linear models, implying there is indeed some advantage to nonlinearity in this problem, but none of the crossvalidated models exceeded the performance achieved by the ViT model used

To see how these point-level models differed from the raster-based ViT in terms of estimating the effect of property value on suppression costs per acre, a doubly robust meta-model was constructed, and the coefficient θ was estimated. Unfortunately, predictions from these models failed to converge for standard error reporting, but point estimates are reported regardless.

REFERENCES CITED

Abatzoglou, J. T., Balch, J. K., Bradley, B. A., & Kolden, C. A. (2018). Human-related ignitions concurrent with high winds promote large wildfires across the USA. *International Journal of Wildland Fire*, *27*(6), 377. Retrieved from `https://doi.org/10.1071/wf17149` doi: 10.1071/wf17149

Abt, K. L., Prestemon, J. P., & Gebert, K. M. (2009, 06). Wildfire Suppression Cost Forecasts for the US Forest Service. *Journal of Forestry*, *107*(4), 173-178. Retrieved from `https://doi.org/10.1093/jof/107.4.173` doi: 10.1093/jof/107.4.173

Albini, F. A. (1979). *Spot fire distance from burning trees-a predictive model*. Intermountain Forest; Range Experiment Station, Forest Service, U.S. Dept. of Agriculture.

Alexander, M. E., & Cruz, M. G. (2013, June). Limitations on the accuracy of model predictions of wildland fire behaviour: A state-of-the-knowledge overview. *The Forestry Chronicle*, *89*(03), 372–383. Retrieved from `https://doi.org/10.5558/tfc2013-067` doi: 10.5558/tfc2013-067

Anderson, T. W., & Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63. doi: 10.1214/aoms/1177730090

Angrist, J., & Frandsen, B. (2020). *Machine Labor*. (NBER Working Paper No. 26584)

Angrist, J., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, New Jersey: Princeton University Press.

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, *14*(1), 57-67.

Angrist, J. D., & Krueger, A. B. (1995, April). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, *13*(2), 225–235. doi: 10.1080/07350015.1995.10524597

Angrist, J. D., & Krueger, A. B. (2001, November). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. Retrieved from `https://doi.org/10.1257/jep.15.4.69` doi: 10.1257/jep.15.4.69

Bayham, J., & Yoder, J. K. (2020, January). Resource allocation under fire. *Land Economics*, *96*(1), 92–110. Retrieved from https://doi.org/10.3368/le.96.1.92 doi: 10.3368/le.96.1.92

Baylis, P., & Boomhower, J. (2019, December). *Moral hazard, wildfires, and the economic incidence of natural disasters* (Tech. Rep.). Retrieved from https://doi.org/10.3386/w26550 doi: 10.3386/w26550

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429. Retrieved from https://doi.org/10.3982/ecta9626 doi: 10.3982/ecta9626

Belloni, A., Chernozhukov, V., & Hansen, C. (2011). *Lasso methods for gaussian instrumental variables models.* Retrieved from https://doi.org/10.2139/ssrn.1908409 (MIT Department of Economics Working Paper No. 11-14) doi: 10.2139/ssrn.1908409

Belloni, A., Chernozhukov, V., & Hansen, C. (2013, November). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650. Retrieved from https://doi.org/10.1093/restud/rdt044 doi: 10.1093/restud/rdt044

Ben-Or, D., Kolomenkin, M., & Shabat, G. (2020). Generalized quantile loss for deep neural networks. *CoRR*, *abs/2012.14348*. Retrieved from https://arxiv.org/abs/2012.14348

Bennett, A., Kallus, N., & Schnabel, T. (2020). *Deep generalized method of moments for instrumental variable analysis.*

Bevis, L. E., & Villa, K. (2020, August). Intergenerational transmission of maternal health: Evidence from cebu, the philippines. *Journal of Human Resources*, 0819–10372R2. Retrieved from https://doi.org/10.3368/jhr.58.1.0819-10372r2 doi: 10.3368/jhr.58.1.0819-10372r2

Biewen, M., & Kugler, P. (2020, August). *Two-stage least squares random forests with an application to angrist and evans (1998).* (IZA Discussion Paper No. 13613)

Boychuk, D., Braun, W. J., Kulperger, R. J., Krougly, Z. L., & Stanford, D. A. (2008, March). A stochastic forest fire growth model. *Environmental and Ecological Statistics*, *16*(2), 133–151. Retrieved from https://doi.org/10.1007/s10651-007-0079-z doi: 10.1007/s10651-007-0079-z

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, *26*(3), 801–824. Retrieved 2022-06-23, from http://www.jstor.org/stable/120055

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. Retrieved from https://doi.org/10.1023/a:1010933404324 doi: 10.1023/a:1010933404324

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). *Geometric deep learning: Grids, groups, graphs, geodesics, and gauges.*

Buma, B., Weiss, S., Hayes, K., & Lucash, M. (2020, February). Wildland fire reburning trends across the US west suggest only short-term negative feedback and differing climatic effects. *Environmental Research Letters*, *15*(3), 034026. Retrieved from https://doi.org/10.1088/1748-9326/ab6c70 doi: 10.1088/1748-9326/ab6c70

Busenberg, G. (2004, March). Wildfire management in the united states: The evolution of a policy failure. *Review of Policy Research*, *21*(2), 145–156. Retrieved from https://doi.org/10.1111/j.1541-1338.2004.00066.x doi: 10.1111/j.1541-1338.2004.00066.x

Butry, D. T., Gumpertz, M., & Genton, M. G. (2008). The production of large and small wildfires. In (pp. 79–106). Springer Netherlands. Retrieved from https://doi.org/10.1007/978-1-4020-4370-3_5 doi: 10.1007/978-1-4020-4370-3_5

Calkin, D. E., Venn, T., Wibbenmeyer, M., & Thompson, M. P. (2013). Estimating US federal wildland fire managers' preferences toward competing strategic suppression objectives. *International Journal of Wildland Fire*, *22*(2), 212. Retrieved from https://doi.org/10.1071/wf11075 doi: 10.1071/wf11075

Chen, D. L., & Yeh, S. (2020). *Government expropriation increases economic growth and racial inequality: Evidence from eminent domain.* (TSE Working Paper No. 16-693) doi: 10.2139/ssrn.2977074

Chen, J., Chen, D. L., & Lewis, G. (2020). *Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models.*

Chen, W., Chen, X., Hsieh, C.-T., & Song, Z. (2019). A forensic examination of china's national accounts. *Brookings Papers on Economic Activity*, *2019*(1), 77–141. Retrieved from https://doi.org/10.1353/eca.2019.0001 doi: 10.1353/eca.2019.0001

Chen, X., & White, H. (1999, March). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, *45*(2), 682–691. Retrieved from `https://doi.org/10.1109/18.749011` doi: 10.1109/18.749011

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018, January). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. Retrieved from `https://doi.org/10.1111/ectj.12097` doi: 10.1111/ectj.12097

Chernozhukov, V., Newey, W., Singh, R., & Syrgkanis, V. (2020). *Adversarial estimation of riesz representers.*

Chernozhukov, V., Newey, W. K., & Singh, R. (2021). *A simple and general debiased machine learning theorem with finite sample guarantees.*

Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., & Chernozhukov, V. (2016, September). *Double machine learning for treatment and causal parameters* (Tech. Rep.). Retrieved from `https://doi.org/10.1920/wp.cem.2016.4916` doi: 10.1920/wp.cem.2016.4916

Colangelo, K., & Lee, Y.-Y. (2021). *Double debiased machine learning nonparametric inference with continuous treatments.*

Cunningham, S. (2021). *Causal inference*. New Haven, CT: Yale University Press.

Derenoncourt, E. (2019, December). *Can you move to opportunity? evidence from the great migration.* (Unpublished)

Dillon, G. (2020, May). Results and application of the national wildfire risk assessment. In *Proceedings of the fire continuum-preparing for the future of wildland fire* (pp. 252–257). Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Donovan, G. H., Noordijk, P., & Radeloff, V. (2004). Estimating the impact of proximity of houses on wildfire suppression costs in oregon and washington. In *In: 2004. proceedings of 2nd symposium on fire economics, planning and policy: A global view.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale.*

Farley, B., & Clark, W. (1954, September). Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 76–84. doi: 10.1109/tit.1954.1057468

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, *89*(1), 181–213. Retrieved from `https://doi.org/10.3982/ecta16901` doi: 10.3982/ecta16901

Finney, M. A. (2002, August). Fire growth using minimum travel time methods. *Canadian Journal of Forest Research*, *32*(8), 1420–1424. Retrieved from `https://doi.org/10.1139/x02-068` doi: 10.1139/x02-068

Finney, M. A. (2006). An overview of flammap fire modeling capabilities. In *In: Andrews, patricia l.; butler, bret w., comps. 2006. fuels management-how to measure success: Conference proceedings. 28-30 march 2006; portland, or. proceedings rmrs-p-41. fort collins, co: Us department of agriculture, forest service, rocky mountain research station. p. 213-220* (Vol. 41).

Florec, V., Thompson, M. P., & y Silva, F. R. (2019). Cost of suppression. In (pp. 1–11). Springer International Publishing. Retrieved from `https://doi.org/10.1007/978-3-319-51727-8_96-1` doi: 10.1007/978-3-319-51727-8_96-1

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Friedman, J. H. (2002, February). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. doi: 10.1016/s0167-9473(01)00065-2

Fuller, W. A. (1977, May). Some properties of a modification of the limited information estimator. *Econometrica*, *45*(4), 939–953. Retrieved from `https://doi.org/10.2307/1912683` doi: 10.2307/1912683

Gebert, K. M., & Black, A. E. (2012, March). Effect of suppression strategies on federal wildland fire expenditures. *Journal of Forestry*, *110*(2), 65–73. Retrieved from `https://doi.org/10.5849/jof.10-068` doi: 10.5849/jof.10-068

Gebert, K. M., Calkin, D. E., & Yoder, J. (2007, July). Estimating suppression expenditures for individual large wildland fires. *Western Journal of Applied Forestry*, *22*(3), 188–196. Retrieved from `https://doi.org/10.1093/wjaf/22.3.188` doi: 10.1093/wjaf/22.3.188

Gorte, R., & Economics, H. (2013). *The rising cost of wildfire protection*. Headwaters Economics Bozeman, MT.

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric of quality change. In *The price statistics of the federal goverment* (p. 173-196). National Bureau of Economic Research, Inc. Retrieved from https://EconPapers.repec.org/RePEc:nbr:nberch:6492

Gude, P. H., Jones, K., Rasker, R., & Greenwood, M. C. (2013). Evidence for the effect of homes on wildfire suppression costs. *International Journal of Wildland Fire*, *22*(4), 537. Retrieved from https://doi.org/10.1071/wf11095 doi: 10.1071/wf11095

Hand, M. S., Gebert, K. M., Liang, J., Calkin, D. E., Thompson, M. P., & Zhou, M. (2014a). *Economics of wildfire management: the development and application of suppression expenditure models*. Springer Science & Business Media.

Hand, M. S., Gebert, K. M., Liang, J., Calkin, D. E., Thompson, M. P., & Zhou, M. (2014b). Regional and temporal trends in wildfire suppression expenditures. In *Economics of wildfire management* (pp. 19–35). Springer.

Hand, M. S., Thompson, M. P., & Calkin, D. E. (2016, January). Examining heterogeneity and wildfire management expenditures using spatially and temporally descriptive data. *Journal of Forest Economics*, *22*, 80–102. Retrieved from https://doi.org/10.1016/j.jfe.2016.01.001 doi: 10.1016/j.jfe.2016.01.001

Hansen, C., Hausman, J., & Newey, W. (2008, October). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, *26*(4), 398–422. Retrieved from https://doi.org/10.1198/073500108000000024 doi: 10.1198/073500108000000024

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning, pmlr* (Vol. 70, pp. 1414–1423). Retrieved from http://proceedings.mlr.press/v70/hartford17a.html

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., & Shi, H. (2021). *Escaping the big data paradigm with compact transformers.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. Retrieved from https://doi.org/10.1007/978-0-387-84858-7 doi: 10.1007/978-0-387-84858-7

Heckman, J., Urzua, S., & Vytlacil, E. (2006, October). *Understanding instrumental variables in models with essential heterogeneity* (Tech. Rep.). Retrieved from https://doi.org/10.3386/w12574 doi: 10.3386/w12574

Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*(3), 669–738. Retrieved 2022-06-25, from http://www.jstor.org/stable/3598865

Heyman, A. V., Law, S., & Berghauser Pont, M. (2019). How is location measured in housing valuation? a systematic review of accessibility specifications in hedonic price models. *Urban Science*, *3*(1). Retrieved from https://www.mdpi.com/2413-8851/3/1/3 doi: 10.3390/urbansci3010003

Ho, T. K. (1995). Random decision forests. In *Proceedings of the third international conference on document analysis and recognition (volume 1)* (pp. 278–282). IEEE Computer Society.

Huang, X. (2020). *100 m population grid in the conus by disaggregating census data with open-source microsoft building footprints.* Harvard Dataverse. Retrieved from https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/DLGP7Y doi: 10.7910/DVN/DLGP7Y

Huang, X., Wang, C., Li, Z., & Ning, H. (2020, July). A 100 m population grid in the CONUS by disaggregating census data with open-source microsoft building footprints. *Big Earth Data*, *5*(1), 112–133. Retrieved from https://doi.org/10.1080/20964471.2020.1776200 doi: 10.1080/20964471.2020.1776200

Ioffe, S., & Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer New York. Retrieved from https://doi.org/10.1007/978-1-4614-7138-7 doi: 10.1007/978-1-4614-7138-7

Jin, Y., Goulden, M. L., Faivre, N., Veraverbeke, S., Sun, F., Hall, A., . . . Randerson, J. T. (2015, September). Identification of two distinct fire regimes in southern california: implications for economic impact and future change. *Environmental Research Letters*, *10*(9), 094005. Retrieved from https://doi.org/10.1088/1748-9326/10/9/094005 doi: 10.1088/1748-9326/10/9/094005

Kain, J. F., & Quigley, J. M. (1970, June). Measuring the value of housing quality. *Journal of the American Statistical Association*, *65*(330), 532–548. Retrieved from https://doi.org/10.1080/01621459.1970.10481102 doi: 10.1080/01621459.1970.10481102

Katuwal, H., Calkin, D. E., & Hand, M. S. (2016, January). Production and efficiency of large wildland fire suppression effort: A stochastic frontier analysis. *Journal of Environmental Management*, *166*, 227–236. Retrieved from https://doi.org/10.1016/j.jenvman.2015.10.030 doi: 10.1016/j.jenvman.2015.10.030

Kilbertus, N., Kusner, M. J., & Silva, R. (2020). *A class of algorithms for general instrumental variable models.*

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization.*

Koenker, R., & Hallock, K. F. (2001, November). Quantile regression. *Journal of Economic Perspectives*, *15*(4), 143–156. Retrieved from https://doi.org/10.1257/jep.15.4.143 doi: 10.1257/jep.15.4.143

LANDFIRE. (2019). *Landfire: Associated products.* https://landfire.gov/version_download.php. (Accessed: 2021-09-30)

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2020). *Valid t-ratio inference for iv.*

Liang, J., Calkin, D. E., Gebert, K. M., Venn, T. J., & Silverstein, R. P. (2008). Factors influencing large wildland fire suppression expenditures. *International Journal of Wildland Fire*, *17*(5), 650. Retrieved from https://doi.org/10.1071/wf07010 doi: 10.1071/wf07010

Liao, L., Chen, Y.-L., Yang, Z., Dai, B., Wang, Z., & Kolar, M. (2020). *Provably efficient neural estimation of structural equation model: An adversarial approach.*

Liu, R., Shang, Z., & Cheng, G. (2020). *On deep instrumental variables estimate.*

Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in adam. *CoRR*, *abs/1711.05101*. Retrieved from http://arxiv.org/abs/1711.05101

Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization.*

Marchal, J., Cumming, S. G., & McIntire, E. J. B. (2017, June). Land cover, more than monthly fire weather, drives fire-size distribution in southern québec forests: Implications for fire risk management. *PLOS ONE*, *12*(6), e0179294. Retrieved from `https://doi.org/10.1371/journal.pone.0179294` doi: 10.1371/journal.pone.0179294

Marlon, J. R., Bartlein, P. J., Gavin, D. G., Long, C. J., Anderson, R. S., Briles, C. E., ... Walsh, M. K. (2012, February). Long-term perspective on wildfires in the western USA. *Proceedings of the National Academy of Sciences*, *109*(9), E535–E543. Retrieved from `https://doi.org/10.1073/pnas.1112839109` doi: 10.1073/pnas.1112839109

Martin, J., & Hillen, T. (2016, June). The spotting distribution of wildfires. *Applied Sciences*, *6*(6), 177. Retrieved from `https://doi.org/10.3390/app6060177` doi: 10.3390/app6060177

Mason, C., & Quigley, J. M. (1996, July). Non-parametric hedonic housing prices. *Housing Studies*, *11*(3), 373–385. Retrieved from `https://doi.org/10.1080/02673039608720863` doi: 10.1080/02673039608720863

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. *Advances in neural information processing systems*, *12*, 512–518.

McCoy, S. J., & Walsh, R. P. (2018, September). Wildfire risk, salience & housing demand. *Journal of Environmental Economics and Management*, *91*, 203–228. Retrieved from `https://doi.org/10.1016/j.jeem.2018.07.005` doi: 10.1016/j.jeem.2018.07.005

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., ... Shi, W. (2006, March). North american regional reanalysis. *Bulletin of the American Meteorological Society*, *87*(3), 343–360. Retrieved from `https://doi.org/10.1175/bams-87-3-343` doi: 10.1175/bams-87-3-343

Mueller-Smith, M. (2015). *The criminal and labor market impacts of incarceration.* (Unpublished)

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. Retrieved from https://doi.org/10.1257/jep.31.2.87 doi: 10.1257/jep.31.2.87

Ng, S., & Bai, J. (2009, January). Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, *1*(1). doi: 10.2202/1941-1928.1014

Pearl, J. (2010, August). 3. the foundations of causal inference. *Sociological Methodology*, *40*(1), 75–149. Retrieved from https://doi.org/10.1111/j.1467-9531.2010.01228.x doi: 10.1111/j.1467-9531.2010.01228.x

Pearson, K. (1901, November). LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. Retrieved from https://doi.org/10.1080/14786440109462720 doi: 10.1080/14786440109462720

Plantinga, A., Walsh, R., & Wibbenmeyer, M. (2021, August). Priorities and effectiveness in wildfire management: Evidence from fire spread in the western united states. *WP*. Retrieved from http://mwibbenmeyer.com/research/

Poudyal, N. C., Hodges, D. G., Fenderson, J., & Tarkington, W. (2010, May). Realizing the economic value of a forested landscape in a viewshed. *Southern Journal of Applied Forestry*, *34*(2), 72–78. Retrieved from https://doi.org/10.1093/sjaf/34.2.72 doi: 10.1093/sjaf/34.2.72

Preisler, H. K., Westerling, A. L., Gebert, K. M., Munoz-Arriola, F., & Holmes, T. P. (2011). Spatially explicit forecasts of large wildland fire probability and suppression costs for california. *International Journal of Wildland Fire*, *20*(4), 508. Retrieved from https://doi.org/10.1071/wf09087 doi: 10.1071/wf09087

Prior-Magee, J. S., Johnson, L. J., Croft, M. J., Case, M. L., Belyea, C. M., & Voge, M. L. (2020). *Protected areas database of the united states (pad-us) 2.1 (provisional release.* U.S. Geological Survey. Retrieved from https://www.sciencebase.gov/catalog/item/5f186a2082cef313ed843257 doi: 10.5066/P92QM3NT

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). “why should i trust you?”. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* ACM. Retrieved from https://doi.org/10.1145/2939672.2939778 doi: 10.1145/2939672.2939778

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rodriguez, M., & Sirmans, C. (1994, 01). Quantifying the value of a view in single-family housing markets. *Appraisal Journal*, *62*, 600-603.

Rothermel, R. C. (1972). *A mathematical model for predicting fire spread in wildland fuels*. Intermountain Forest amp; Range Experiment Station, Forest Service, U.S. Dept. of Agriculture.

Rothermel, R. C. (1983). *How to predict the spread and intensity of forest and range fires* (Tech. Rep.). Retrieved from https://doi.org/10.2737/int-gtr-143 doi: 10.2737/int-gtr-143

Santosa, F., & Symes, W. W. (1986, October). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307–1330. Retrieved from https://doi.org/10.1137/0907087 doi: 10.1137/0907087

Scott, J. H., Thompson, M. P., & Calkin, D. E. (2013). *A wildfire risk assessment framework for land and resource management* (Tech. Rep.). Retrieved from https://doi.org/10.2737/rmrs-gtr-315 doi: 10.2737/rmrs-gtr-315

Short, K. C., Finney, M. A., Vogler, K. C., Scott, J. H., Gilbertson-Day, J. W., & Grenfell, I. C. (n.d.). *Spatial datasets of probabilistic wildfire risk components for the united states (270m) (2nd edition)*. Forest Service Research Data Archive. Retrieved from https://doi.org/10.2737/rds-2016-0034-2 doi: 10.2737/rds-2016-0034-2

Singh, R., Sahani, M., & Gretton, A. (2019). *Kernel instrumental variable regression.*

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014, January). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Storm, H., Baylis, K., & Heckelei, T. (2019, August). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, *47*(3), 849–892. Retrieved from https://doi.org/10.1093/erae/jbz033 doi: 10.1093/erae/jbz033

Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). *Roformer: Enhanced transformer with rotary position embedding.*

Suppression costs. (2020). https://www.nifc.gov/fire-information/statistics/suppression-costs. NIFC.