

Aligning a household-level service array via geospatial and counterfactual modeling: study protocol for a jurisdiction-wide child maltreatment prevention effort

Jamaal Green, Brian Glass, Jordan Purdy, Dyann Daley

Submitted to: JMIR Research Protocols
on: March 28, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Aligning a household-level service array via geospatial and counterfactual modeling: study protocol for a jurisdiction-wide child maltreatment prevention effort

Jamaal Green^{1*} PhD; Brian Glass^{2*} PhD; Jordan Purdy^{2*} PhD; Dyann Daley^{3*} MD

¹ University of Pennsylvania Philadelphia US

² Counterbalance AI Tualitin US

³ Predict-Align-Prevent Grapevine US

*these authors contributed equally

Corresponding Author:

Jamaal Green PhD

University of Pennsylvania
127 Meyerson Hall
210 S. 34th Street
Philadelphia
US

Abstract

Background: Child maltreatment is associated with multiple negative outcomes at the individual and societal level. Children sufferers of maltreatment are at greater risk of a host of negative outcomes (e.g., psychological disorders, substance use, violent delinquency, suicidality, educational outcomes).

Objective: In order to prevent and to ameliorate child maltreatment a combination of geospatial smoothing via a risk terrain modeling framework and counterfactual modeling are proffered here to identify risky areas and to determine optimal (re-)allocation of services to maximally improve maltreatment outcomes.

Methods: A three stage process is proposed which can iteratively be applied within a collaborating jurisdiction to enable responsive and sustained achievement of identified child welfare outcomes. This process makes use of two analytic approaches: geospatial smoothing via a risk terrain framework and counterfactual modeling. Risk terrain modeling (RTM) is a spatial analytic approach that uses spatial machine learning methods to estimate the risk of maltreatment based on prior cases of maltreatment and risk factors of the built environment. Using prior validated cases of maltreatment, violent crime data and built environment data we estimate a series of machine learning models to geospatially smooth the historically identified places at increased risk of child maltreatment. Areas identified as higher risk receive extensive services associated with preventing or limiting child maltreatment such as pre/postnatal care, subsidized daycare and parental counseling. We make use of counterfactual explanation modeling to optimally align service allocation to maximally improve maltreatment outcomes for future service allocations within a collaborating jurisdiction. This technique leverages a statistical model associating household-level information with maltreatment outcomes in order to explore combinations of services which would be predicted to achieve optimal and practical recommendations for future service allocation efforts.

Results: The household level counterfactual recommendations described in the previous section will be aggregated into actionable recommendations. The primary actionable recommendation will directly inform future iterations of Stage 1 by providing a model-informed approach for determining eligible regions.

There are several methods and strategies for aggregating the resulting household-level counterfactual service arrays. The first step is to select, for each household observation, which of the 2^n service array permutations to select. The second step is to aggregate the household-level counterfactual service arrays into spatial subregions. These subregions could be the fishnet grids specified in Stage 1, or any other chosen subregion definition such as ZIP Code, census tract, or school district. The simplest method is to, for each service type, sum the number of households in the subregion that were determined to benefit from the service. Constraints can be introduced to this logic, such as service availability and cost. Algorithmic fairness is also a potential consideration during aggregation, with possibilities for both measuring and balancing metrics such as "recourse fairness".

Conclusions: This protocol sets forth a novel approach for the allocation of supportive services for families at risk of child maltreatment through geospatial smoothing via a RTM framework and the maximization of service impact through CEM. Child maltreatment is an unfortunate, and ubiquitous, issue in the United States. This proposal builds on jurisdiction-wide public health strategies in order to allocate services in a data-informed fashion and further align future iterations of the allocation strategy using outcomes-based counterfactual modeling at the household level. The flexibility of the proposed methodology enables its application regardless of the collaborating jurisdiction's preferences and constraints.

(JMIR Preprints 28/03/2025:71724)

DOI: <https://doi.org/10.2196/preprints.71724>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR publication, my accepted manuscript PDF will be visible to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Aligning a household-level service array via geospatial and counterfactual modeling: study protocol for a jurisdiction-wide child maltreatment prevention effort

Abstract

Background: Child maltreatment is associated with multiple negative outcomes at the individual and societal level. Children sufferers of maltreatment are at greater risk of a host of negative outcomes (e.g., psychological disorders, substance use, violent delinquency, suicidality, educational outcomes).

Objective: In order to prevent and to ameliorate child maltreatment a combination of geospatial smoothing via a risk terrain modeling framework and counterfactual modeling are proffered here to identify risky areas and to determine optimal (re-)allocation of services to maximally improve maltreatment outcomes.

Methods: A three stage process is proposed which can iteratively be applied within a collaborating jurisdiction to enable responsive and sustained achievement of identified child welfare outcomes. This process makes use of two analytic approaches: geospatial smoothing via a risk terrain framework and counterfactual modeling. Risk terrain modeling (RTM) is a spatial analytic approach that uses spatial machine learning methods to estimate the risk of maltreatment based on prior cases of maltreatment and risk factors of the built environment. Using prior validated cases of maltreatment, violent crime data and built environment data we estimate a series of machine learning models to geospatially smooth the historically identified places at increased risk of child maltreatment. Areas identified as higher risk receive extensive services associated with preventing or limiting child maltreatment such as pre/postnatal care, subsidized daycare and parental counseling. We make use of counterfactual explanation modeling to optimally align service allocation to maximally improve maltreatment outcomes for future service allocations within a collaborating jurisdiction. This technique leverages a statistical model associating household-level information with maltreatment outcomes in order to explore

combinations of services which would be predicted to achieve optimal and practical recommendations for future service allocation efforts.

Results: The household level counterfactual recommendations described in the previous section will be aggregated into actionable recommendations. The primary actionable recommendation will directly inform future iterations of Stage 1 by providing a model-informed approach for determining eligible regions.

There are several methods and strategies for aggregating the resulting household-level counterfactual service arrays. The first step is to select, for each household observation, which of the 2^n service array permutations to select. The second step is to aggregate the household-level counterfactual service arrays into spatial subregions. These subregions could be the fishnet grids specified in Stage 1, or any other chosen subregion definition such as ZIP Code, census tract, or school district. The simplest method is to, for each service type, sum the number of households in the subregion that were determined to benefit from the service. Constraints can be introduced to this logic, such as service availability and cost. Algorithmic fairness is also a potential consideration during aggregation, with possibilities for both measuring and balancing metrics such as "recourse fairness".

1 Introduction

In 2022, the United States Children's Bureau reported 558,899 cases of child maltreatment and 1,990 fatalities [1]. Although the rates of child sexual and physical abuse have declined in the U.S. since 1990 [2], the incidence of child maltreatment fatalities has risen since 2008. Child maltreatment is underreported [3], and actual fatalities are estimated to be two to three times higher than reported due to inconsistencies in definitions and reporting standards across states [4]. Child maltreatment is linked to the development of medical illness and psychiatric disorders, as well as poor education, employment, economic, interpersonal, and community outcomes [5–13].

The estimated cost per victim of nonfatal child maltreatment is \$830,928 (2015 USD). The average lifetime cost associated with each child maltreatment fatality is estimated at \$16.6 million (2015 USD). In 2015, the economic burden of child maltreatment in the United States was estimated at \$428 billion for substantiated cases, and \$2 trillion for annually investigated incidents [14]. This economic burden is significant and becomes even more substantial when considered in the broader context of impact on children, family, and communities.

Given the immense personal and societal costs of child maltreatment, it is important to design programs and policies to not only alleviate but prevent abuse in the first place. Young children (ages 0-3) are particularly exposed to the risks of maltreatment [15] and many children who die due to maltreatment are not known to child protection agencies [16, 17]. Given these sobering facts, identifying places and families at increased risk of maltreatment may allow authorities to proactively intervene before maltreatment may occur. Policies such as pre and post-natal counseling, subsidized childcare, and parental counseling may all have protective effects with respect to potential maltreatment if agencies can successfully identify families at risk [18].

The aim of this protocol is to address three enduring issues: targeting of primary prevention resources, expansion of prevention programs to span risk factors, and identification of the bundle of programs that effectively prevents child maltreatment for a given region or household. To accomplish this, the proposed study protocol seeks to identify areas with young children at increased risk of experiencing child maltreatment, and to target these areas with extensive support programs with the goal of reducing and preventing child abuse.

The study is composed of three stages which can iteratively be applied within a collaborating jurisdiction to enable responsive and sustained achievement of the aforementioned goal. Stage 1 of

the study is the development of a smoothed risk surface of child abuse and maltreatment using spatial machine learning approaches within a “risk terrain modeling” framework [19]. Stage 2 of the study will be the resulting service allocation process for areas of increased estimated risk identified in Stage 1. Finally, Stage 3 uses a counterfactual explanation model to determine optimally aligned service allocation to maximally improve maltreatment outcomes for future service allocation.

We propose the use of spatial risk modeling to improve targeting of primary prevention resources, coordination of the delivery of voluntary prevention programs that span individual and contextual risks using targeted universalism, and identification of the most cost-effective bundle of primary prevention programs using counterfactual modeling. It is important to note that this work is not concerned with simply validating whether individual support programs or their various combinations are effective, in a general sense, across a collaborating jurisdiction. This work seeks to expand beyond the utility of classic program evaluation models by identifying household-specific combinations of available support programs that optimally improve household-specific outcomes.

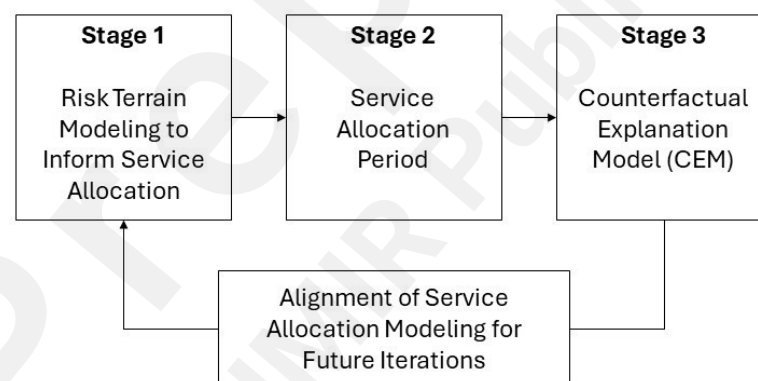


Figure 1: The three stages of the proposed study design.

2 Background

2.1 General Study Design

The study protocol is divided into three stages: Stage 1) Geospatial

service recommendations, Stage 2) Execution of service array in the collaborating jurisdiction, and Stage 3) Counterfactual modeling of service array impact (see Figure 1). The results from Stage 3 can then be used in an iterative fashion to inform the geospatial service recommendations moving forward. In Stage 1, geospatial information about historical child maltreatment events are leveraged in a risk terrain modeling framework to smooth the outcome surface and ultimately yield an actionable fishnet map of recommended locations to receive expanded child welfare related services. In Stage 2, the services are offered and allocated within the collaborating jurisdiction

guided by the geospatial service recommendation map from Stage 1. After a period of service delivery of sufficient duration to allow for a long-arc child maltreatment outcome window (e.g., 5 years), Stage 3 involves the development of a Counterfactual Explanation Model (CEM) to develop actionable jurisdiction-wide recommendations based on the impact of the service array on child maltreatment in the jurisdiction. Rather than provide a simplified pre-post evaluation of service effectiveness, CEM offers a tailored approach to resolve optimal service arrays at the household level, thus further aligning the geospatial service recommendations offered in Stage

1. Several details of the protocol, including the specific services offered, duration of Stage 2, and the final format of the actionable recommendations will be determined during the design scoping phase, which will involve feedback and specifications provided by the collaborating jurisdiction. The current study protocol seeks to propose a design framework robust to the various configurations and challenges that may arise during the final determination of the study details, which will be highly dependent on the governing body in the collaborating jurisdiction.

2.2 Risk Terrain Modeling: Theoretical Background and Use Cases

Risk Terrain Modeling (RTM) is a spatial analytic approach developed by Caplan et al for the purpose of estimating spatial risk with respect to environmental determinants of crime [20]. While RTM comes out of the environmental criminology literature, the basic approach is an application of existing spatial analytic approaches. The ultimate goal is the identification of risky places as opposed to individuals. The environmental criminological literature has long argued that certain environmental features can increase, or decrease, the risk of certain crimes occurring. Brantingham and Brantingham [21] theorized there are two major types of features in the built environment that influence crime: crime attractors and crime generators. Crime generators are areas that attract a large number of people and offer increased opportunities for crime. Stadium or entertainment districts are two examples of crime-generating spaces. Crime attractors, conversely, are places that offer opportunities for crime and are attractive to potential offenders because of the opportunity to engage in certain types of crime. Open-air drug markets would be an example of such a place. RTM, then, is an application of the basic logic set forth by Brantingham and Brantingham (1995) through the use of spatial predictive modeling.

The classic use cases for RTM are, of course, for estimating neighborhood risk profiles with respect to crime of varied kinds. Violent crimes, such as shootings or robberies, are often studied [22]. Recently, RTM has also been used to identify places at higher

risk of child abuse and maltreatment [19] [23].

2.3 Service Allocation Period

The service allocation period will rely on the geospatial analysis from Stage 1, and involves the actual service delivery of various prevention programs to households in the jurisdiction. The jurisdiction-wide service delivery expansion effort is influenced by prior work in public health strategies for the deployment of large scale service delivery efforts. See Section 4.2 for further details.

2.4 Counterfactual Explanation Models: Theoretical Background and Use Case examples

CEMs are statistical models of real world phenomena which can produce actionable recommendations in the interest of maximizing or minimizing the probability that some alternative

outcome would have occurred. CEMs have generated considerable recent interest due to 1) their ability to resolve interpretations from "black box" models [24, 25], 2) their structural alignment with both scientific and everyday reasoning about the world [26], and 3) their ability to offer deployable and testable courses of action for policy makers and practitioners. CEM methodologies can produce counterfactual courses of action for a single observational input (e.g., how a loan applicant can improve their standing, how a medical course of action can improve the survival of an ICU patient, or how a given service combination can improve the child welfare outlook of a single household).

The classic use case example for CEMs is the financial loan scenario: consider a statistical model which predicts whether a lender will repay a loan within a given timeframe. By perturbing the input factors in the statistical model, the lending institution can discover a set of "solutions" which would be predicted to convert an unsuccessful loan applicant into a successful loan applicant, thus offering recommendations to rejected applicants. Another recent example from the field of healthcare is the proposal for CEM in an intensive care unit [27]. In the proposal, a CEM built on historical data offers treatment recommendations to maximize patient survivability. If a patient is predicted to have low chances of survival upon ICU intake, the CEM offers a treatment array sequence to maximally impact the chance of survival. Thus, a CEM leveraged a statistical model linking treatment combinations to health outcomes in order to provide actionable individual-level treatment array recommendations.

2.4.1 Characteristics of Counterfactual Explanation Models

A variety of theoretical work has considered the evaluation of a CEM's quality [28]. This prior work generally considers three domains for the evaluation of a given CEM: Validity, Proximity, Diversity, and Actionability [29].

Validity can refer both to the performance (e.g., AUC, accuracy, smoothness of the data manifold) of a CEM's underlying statistical model, as well as the CEM's ability to provide truly counterfactual predictions [29]. Individual predictions are truly counterfactual when the resulting outcome is in a different class than the original outcome (i.e., perturbing the features of a given input vector results in a different predicted binary outcome). Thus, statistical models whose predicted outcomes are invariant to alterations in the input vector values are not valid for use as a CEM.

Proximity refers to the minimally required alteration of an input vector CEM to generate effective counterfactuals (i.e., altered input vectors which result in a different prediction class). CEM with superior proximity generate counterfactual input vectors using minimal perturbations to the original input vector [30-32].

Diversity refers to the size or cardinality of the set of counterfactual options resolved by a CEM [33]. For example, for a given individual input vector, a CEM might recommend a large and varied set of counterfactual feature configurations which all successfully convert the pre- diction to the more ideal class. While not necessarily a detriment, large sets of "other possible worlds", known as the Rashomon Effect [34], offers a conundrum to CEM deployment, as many differing and often competing courses of action must be further compared and evaluated.

Actionability refers to the level of controllability of the input vector features perturbed by the CEM [35]. For example, an ICU treatment model which counterfactually recommends that a patient become taller is not actionable. This proposal seeks to maximize actionability by only perturbing input vector features which relate to the application of specific family services.

3 Methodology

The study consists of three stages which can iteratively be applied within a collaborating jurisdiction. Stage 1 of the study is the development of a smoothed risk surface of child abuse and maltreatment. Stage 2 of the study will be the resulting service allocation process for areas identified in Stage 1. Finally, Stage 3 uses a counterfactual explanation model to determine optimally aligned service allocation for future iterations of the service array.

3.1 Stage 1 - “Risk Terrain Modeling”

The objective of Stage 1 is neither to predict when or where the outcome will occur, nor to identify regions of latent risk, but rather to apply risk terrain modeling to initially identify the regions where the benefits of service allocation might reasonably be expected. To this end, we propose smoothing the historical outcome surface across the jurisdiction via development of a geospatial machine learning algorithm, where regions corresponding to higher values from the trained smoother will inform the initial service allocation encompassed by Stage 2.

Within the jurisdiction of interest, the observational units of this geospatial smoother are regular polygons the size of which will depend upon the collaborating jurisdiction. Henceforth, these observational units are referred to as “fishnet” grid cells. The outcome being smoothed will pertain to child maltreatment, as defined by the collaborating jurisdiction, with the specific form (e.g., rate, count, etc.) likely being determined after initial exploratory data analysis. The feature set will consist of variables based on the built environment, crime, and census data both within the cell and within neighboring cells, as well as the value of the outcome in neighboring cells (i.e., auto-features), where a variety of isotropic neighborhood structures will be considered (e.g., first-order; second-order; etc.).

Two approaches to creating the smoothed surface will be considered, with goodness-of-fit measures being employed to identify the better of the two. The first, kernel density estimation (KDE) ([20]), is expected to be inferior and will serve as the baseline approach. The second, a tree-boosted learner (e.g., XGBoost [36]), is more complicated, but is expected to be superior. For the boosted learner, k-fold cross validation - with the spatial neighborhood structure dictating the partitioning of the folds - will be used.

3.1.1 Algorithmic Fairness of Geospatial Smoother

Within the paradigm of person-level algorithms and related decision-support tools, there exists a robust literature on algorithmic fairness. Among other things, this literature posits numerous definitions of algorithmic fairness (see, for example, [37])

and [38] for an identified, and potentially multidimensional, protected attribute, along with a number of mitigation procedures for “correcting” the output of such person-level algorithms with respect to the protected attribute (e.g., [39], [40] and [41]). Unfortunately, within the paradigm of geospatial algorithms and related decision-support tools, there is a relative dearth of literature ([42]) pertaining to algorithmic fairness.

In light of this reality, one approach sometimes used is an assessment of the geospatial algorithm’s generalizability across coarse categorizations of poverty (high- versus low-poverty) and/or race (majority white versus majority non-white) based on census data (e.g., [43]). Such assessments are akin to a type of algorithmic fairness audit, but provide no clear course of action in the face of poor generalizability.

Here we propose a relatively simple approach encompassing the algorithmic fairness of the geospatial smoother and the corresponding identification of fishnet cells for service allocation. Our goals with this procedure are to enable a simple, coarse auditing of the geospatial smoother’s outputs across a census-derived protected attribute and to provide a means for the

subsequent and corresponding equitable allocation of services across the levels of the protected attribute.

The approach to algorithmic fairness consists of two steps. In step 1, unsupervised machine learning (e.g., clustering) is used to identify “levels” (e.g., clusters) of a protected attribute. Such unsupervised learning would use relevant - as identified by the jurisdiction of interest

- features from the census data (e.g., racial percentages; poverty percentages; etc.). If the proposed unsupervised learning “fails” (i.e., is uninformative for identifying groups/levels of a protected attribute), then we will default in step 1 to a poverty- and/or race-based protected attribute consisting of only a few levels. In step 2, the limited number of fishnet cells identified for allocation of services within Stage 2 is based on protected attribute level stratification of the geospatial smoother’s output. Such identification of fishnet cells for Stage 2 is consistent with an equitable allocation of services, where the form of equity is ultimately determined by the collaborating jurisdiction. For example, a jurisdiction could elect to allocate each level of the protected attribute an equal number of fishnet cells receiving services, with the specific fishnet cells within each protected attribute level being identified according to higher values of the geospatial smoother.

3.2 Stage 2 - Service Allocation

The service array provided to the households within the fishnet cells identified in Stage 1 will be finalized with input from the collaborating jurisdiction, although the following list presents a core set of service programs and/or service types. The service array will be selected to impact the various known major risk factors for child maltreatment or adverse childhood experiences.

1. Nurse-family partnership [44]
2. Cure Violence [45-47]
3. Crime Prevention Through Environmental Design (CPTED) [48-50]
4. Universal basic income [51, 52]
5. Early Head Start [53, 54]
6. Pregnancy prevention (e.g., Upstream USA) [55, 56]
7. Mobile medical clinics [57, 58]
8. Crisis intervention teams [59]
9. Free college tuition [60, 61]
10. Stewards of Children: Darkness to Light Training [62]

Each service has been demonstrated to address one or more of these major risk factors. This comprehensive multi-service jurisdiction-wide approach uses a public health strategy (see [63, 64]) to expand preventive services to households which may not already be known to the collaborating jurisdiction’s Child Welfare system. The entire set of services chosen by the collaborating jurisdiction will be available to every qualified household of each

fishnet cell identified in Stage 1. Crucially, data records will track household-level service delivery and participation, allowing for the quantification of service array information for use in Stage 3.

3.3 Stage 3 - Service Alignment Recommendation via Counter- factual Explanation Model

3.3.1 Maltreatment Classifier Model Construction and Selection

The first step of Stage 3 is to develop a statistical model to associate available and quantifiable features with the selected child welfare maltreatment outcome of interest. Model specification

and selection will consider the four characteristics of CEMs identified above. While data availability details will not be finalized until Stages 1 and 2 are executed in the collaborating jurisdiction, prior work suggests it is reasonable to anticipate that machine learning models can be trained to associate quantifiable household factors with child maltreatment outcomes with high performance, high algorithmic fairness, and high computational efficiency [41, 65, 66]. One novel contribution of this study protocol will be to extend the use of such models to inform the alignment of jurisdiction-wide service deployment strategies by maximizing the potential for positive outcomes at the household level.

Observational Unit and Data Universe: The granularity of the statistical model will be the household level. All households in the collaborating jurisdiction with at least one child member will comprise the data universe. The data set used to train and validate the statistical model will consist of the union of three mutually inclusive subsets of the data universe: Set 1) Households eligible to receive services as identified in Stage 1 of the protocol, Set 2) Households that actually received services in Stage 2 of the protocol (some of which may not have been in Set 1), and Set 3) Households found in the Statewide Automated Child Welfare Information System (SACWIS) [67]. Thus, the union of these three sets encompasses all households who either were eligible to receive services and/or are known to have the necessary SACWIS records to define an outcome. Households in the jurisdiction universe who fall out of this union set will be available as a counterfactual inference set. That is, despite not having a defined services record, their features will be available to construct an input vector for the statistical model, allowing for a predicted outcome and in turn an optimized counterfactual service array.

Sources of Model Features: A feature engineering stage will attach available quantifiable information to each household observation. These features will be comprised of information drawn from four general sources: Source 1) Household level data available from consumer and market data, Source 2) local built environment data attached to nearby households, Source 3) Governmental reporting and services data, and Source 4) the household's actual service array experience during Stage 2. During counterfactual optimization, Sources 1 and 2 remain static while Source 4 (the service array) is perturbed in order to minimize (or maximize) the predicted probability of the chosen child welfare outcome of interest. In this way, maximum actionability is preserved, since no household level features beyond the controllable service array are perturbed as part of the counterfactual modeling. Importantly, administrative data from the SACWIS system will not be included as model features. This allows for the counterfactual model to be deployable for households which have not been

involved with the collaborating jurisdiction's child welfare system. Governmental reporting and services data (Source 4) can act as important predictors as well as inform a potential post-hoc outcomes analysis (see Section 4.2). The list below presents potential data elements for Source 4.

1. Healthcare / Hospital

- | | |
|---|---|
| (a) Preterm Birth | (h) Prevention Quality Indicators (Hospital Visits) |
| (b) Very Low Birthweight | (i) Pediatric Quality Indicators (Area Level) |
| (c) Failure to Thrive | (j) Psychiatric Hospitalizations |
| (d) Neonatal Abstinence Syndrome | (k) Injuries from Violence |
| (e) Pediatric Physical / Developmental Disabilities | (l) Pediatric Lead Poisoning |
| (f) Teen Birth | (m) Maternal Morbidity / Mortality |
| (g) Preventable Hospital Admissions | (n) Substance Abuse |
| | (o) Substance Related Overdoses |

- (p) Accidental Deaths
- (q) Sexual Violence / Rape
- (r) Prenatal Care Visits
- 2. Dept. of Health / Medical Examiner
 - (a) Infant / Child Deaths
 - (b) Premature Deaths
 - (c) Excess Mortality
 - (d) Marriage / Divorce
- 3. Crime / Emergency Services
 - (a) Arrests
 - (b) Incarcerations
 - (c) Domestic Disturbance / Violence
 - (d) Assault
 - (e) Gunshot / Shooting / Stabbings
 - (f) Intoxication
 - (g) Drug Abuse / Manufacturing
 - (h) Animal Abuse / Control
 - (i) Sexual Assault / Rape
- (j) Arson
- (k) DUI
- (l) Homicide
- (m) Harassment
- (n) Juvenile Kidnapping
- (o) Reckless Driving
- (p) Suicide
- (q) Welfare Checks
- 4. Education / Employment
 - (a) Graduation Rates
 - (b) Truancy Rates
 - (c) Higher Education / Trade School Matriculation
 - (d) Kindergarten Readiness
 - (e) 3rd Grade Reading Level
 - (f) Unemployment
 - (g) Poverty Indices

Model Outcome: The outcome will be defined via a joint venture with the collaborating jurisdiction's governing body and will relate to child maltreatment at the household level. Prior work has considered various maltreatment related outcomes that vary by intensity and prevalence, such as reports/referrals of child maltreatment (i.e., community members reporting alleged child abuse or neglect to a child welfare agency), whether a Child Protective Services (CPS) investigation occurred for an alleged child victim in the household, whether a CPS investigation determined substantiated maltreatment in the home, and whether a child in the home was removed and placed into some form of substitute care as a ward of the state. Outcome availability will differ between households depending on the three data universe subsets listed above,

although it will be possible to both predict and define an outcome for each observation. For example, a household who was eligible for services and received services, but was not found in the SACWIS system (i.e., a member of Sets 1 and 2, but not Set 3), will be defined as "maltreatment absent", that is, there was no detected household maltreatment. However, it will still be possible to calculate a predicted probability of "maltreatment present", thus allowing for defining a counterfactual service array that could have further decreased any potential for maltreatment present in the household.

Model Selection Criteria: Best practice will be followed in order to construct and select between statistical model specifications which best satisfy the requirements for an acceptable associative model for use in the CEM stage [68–70]. This best practice includes 1) exploring a range of diverse ML classifier types (e.g., XGboost, support vector machines,

neural networks), 2) comparing classifier performance on a validation set (e.g., area under the curve (AUC), accuracy, specificity, sensitivity), 3) comparing the performance-fairness trade-off for each model, with the aim of achieving large gains in algorithmic fairness at minimal cost to performance, 4) computational demands (as the CEM stage utilizes an exhaustive search over thousands of perturbed input vectors for each observation), and 5) the ability of the statistical model to underlie a CEM which produces counterfactual results that are 5a) Valid, 5b) Proximal, and 5c) Diverse (see Section 3.4.1 above).

Potential Algorithmic Fairness Correction: The resulting statistical model will produce predicted probabilities that will be used ordinarily within an observation's counterfactual array (see below). However, it is possible that the final deployed CEM will require between-observation comparison at the aggregation stage in order to meet the use case requirements selected in collaboration with the jurisdiction. In the case that predicted probabilities will be directly compared between observations, and therefore between households at different levels of a protected attribute, algorithmic fairness corrections will be considered. The construction of such correctional procedures will follow the guidance of the well-established literature references in Section 4.1.1 above (e.g., [41]).

3.3.2 Compute Counterfactual Service Arrays

The statistical model identified above will become the underlying classification mechanism of the CEM. This CEM will produce service alignment recommendations by computing counterfactual service arrays for each historical observation at the household level. This will be accomplished by optimizing the service array vector for each household observation with respect to certain criteria or constraints, such as minimizing the predicted probability of the maltreatment outcome while also minimizing the number of additional services required. The following computational methodology outlines how this is accomplished: 1) An input vector is constructed for each household observation, which includes both the household characteristic features and the binary service array features (i.e., a 0 signifies the household received Service A, while a 1 signifies the household did not receive Service A), 2) all permutations of the binary service array vector are exhaustively considered to create a set of 2^n candidate input vectors (where n is the number of available services), one of which represents the actually received service array, 3) a predicted probability of the maltreatment outcome is computed for each candidate input vector, and 4) each of the 2^n candidate input vectors are quantified in terms of CEM characteristics for use in the aggregation stage. Importantly, the brute force search employed in this protocol to exhaustively consider all service array permutations precludes various challenges with CEM deployment associated with the possibility of missing globally optimal regions of the search space [71].

In step 4 above, each candidate input vector (of the 2^n defined for each household observation) is described by certain characteristics. These characteristics are A) the change in predicted probability of the maltreatment outcome between original input vector and the candidate input vector (e.g., +4%, -10%), B) the proximity of the candidate service array vector to the original service array vector in Hamming distance (i.e., the number of services which differ between the original and the candidate counterfactual under consideration [72]), and C) the proximity of the candidate input vector to a known alternative input vector from the newly predicted class, representing a proxy for generalizability confidence (by Gower's distance [73], with 1 representing equality and 0 representing maximal distance in the set). In this way, the full set of counterfactual service array candidates for a given household observation can be sorted, filtered, and evaluated to serve specific purposes in the Recommendation Aggregation stage.

For example, consider a household observation whose

maltreatment outcome was positive (with retroactive predicted probability of 85%) and whose actual service array vector was 0, 0, 1, 0, 1, 0 (i.e., in Stage 2, the household received services C and E). And, for clarity, consider three members of its resulting 2^n candidate input vectors. Table 1 describes these candidate input vectors and the resulting CEM characteristics A-D for each candidate. Identifying a single recommendation from the resulting counterfactual candidates requires developing certain standards or rules. These criteria will be developed in conjunction with the collaborating jurisdiction in order to maximally adhere to their pragmatic considerations such as resource

Input Vector Description	Service Array Vector	Predicted Probability of Maltreatment	Predicted Probability of Maltreatment	Change in Services Required	Proximity to Known Member of No-Maltr.
Actual	{0, 0, 1, 0, 1, 0}	85%	-	-	0.31
CEM Candidate 1	{0, 0, 1, 1, 1, 0}	80%	-5%	Net 1 (+1)	0.99
CEM Candidate 2	{1, 1, 0, 1, 1, 0}	50%	-35%	Net 1 (+1, -2)	0.67
CEM Candidate 3	{1, 1, 1, 0, 1, 0}	45%	-40%	Net 3	0.42

Table 1: Example comparison of three counterfactual candidates computed from an actual observation using the Counterfactual Explanation Model (CEM).

constraints. For this example, consider how these candidates might be quantifiably compared to one another. Candidate 1 reduces the predicted probability by 5%, which is less than the other two candidates, however it requires only 1 additional service and is also very similar to a known observation in which maltreatment was in fact prevented. Candidate 2 reduces the predicted probability by 35%, requires a net change of 1 service (1 subtracted, 2 additional) and has moderate similarity to a known observation in the alternative class. Candidate 3 reduces the predicted probability by 40%, more than the other candidates, but requires 3 additional services, and is minimally similar to a known observation in the alternative class. If the determined considerations called for identifying recommendations that would require minimal additional services and maximize proximity to known observations of the maltreatment prevention class, then Candidate 1 would be selected. If the determined considerations called for identifying recommendations that would maximize the reduction of predicted maltreatment, regardless of the number of additional services and regardless of the confidence provided by proximity to known observations, then Candidate 3 would be selected. A balanced approach might result in the selection of Candidate 2. In this way, the comprehensive quantification of each candidate result works to anticipate the myriad potential weighting scenarios from which one will be chosen as the guiding principle for selecting household level counterfactual recommendations.

4. Results

The household level counterfactual recommendations described in the previous section will be aggregated into actionable recommendations. The primary actionable recommendation will directly inform future iterations of Stage 1 by providing a model-informed approach for determining eligible regions. Other secondary recommendations are also possible, if the final project scope calls for other ways to aggregate the results to fit various project goals. The overall conceptual method for generating these recommendations is to aggregate household-level counterfactual service array candidates into an actionable result [74].

There are several methods and strategies for aggregating the resulting household-level counterfactual service arrays. The first step is to select, for each household observation, which of the 2ⁿ service array permutations to select. The previous section details how this selection logic can be specified in order to reflect the jurisdiction's overall project goals and constraints. The second step is to aggregate or "roll up" the household-level counterfactual service arrays

into spatial subregions. These subregions could be the fishnet grids specified in Stage 1, or any other chosen subregion definition such as ZIP Code, census tract, or school district. Bespoke logic can determine the most pragmatic approach for grouping service recommendations into subregions. The simplest method is to, for each service type, sum the number of households in the subregion that were determined to benefit from the service. Constraints can be introduced to this logic, such as service availability and cost. Algorithmic fairness is also a potential consideration during aggregation, with possibilities for both measuring and balancing metrics such as "recourse fairness" [75]. Finally, these results are organized into an actionable recommendation strategy which can be provided to the jurisdiction to improve their future service delivery.

5 Discussion

This protocol sets forth a novel approach for the allocation of supportive services for families at risk of child maltreatment through geospatial smoothing via a RTM framework and the maximization of service impact through CEM. Child maltreatment is an unfortunate, and ubiquitous, issue in the United States. This proposal builds on jurisdiction-wide public health strategies in order to allocate services in a data-informed fashion and further align future iterations of the allocation strategy using outcomes-based counterfactual modeling at the household level. The flexibility of the proposed methodology enables its application regardless of the collaborating jurisdiction's preferences and constraints.

Finally, it is worth noting that while the objective of the presented protocol is to reduce child maltreatment within the collaborating jurisdiction, post-hoc descriptive analyses related to changes in jurisdiction-wide outcomes known to be associated with child maltreatment (e.g., ACES; violent crime, etc.; see Section 4.3.1 for an expanded list) are possible and may be of interest to both the collaborating jurisdiction and corresponding service providers. For example, the Cure Violence program will likely be interested in whether aggregate measures of violent crime in the collaborating jurisdiction have declined pre- versus post-implementation of a given iteration of the implemented service array. Importantly, however, whether such associated outcomes are causally improved as a result of the implemented service array neither can be nor is intended to be answered by the presented protocol. Such associated outcomes may or may not be causally related to child maltreatment, and therefore reducing child maltreatment may not result in, for example, a reduction in violent crime across the collaborating jurisdiction. Regardless, such post-hoc analyses are possible and can be informative for both the collaborating jurisdiction and corresponding service providers.

References

- [1] *Child Maltreatment* 2022. <https://www.acf.hhs.gov/cb/data-research/child-maltreatment>. [Online; accessed 10-June-2024]. 2024.
- [2] Mallory Lucier-Greer et al. "Trends in the Annual Incidence Rates of Child Sexual Abuse and Child Maltreatment over the Past 25 Years in the United States". In: *Child Abuse Review* 33.2 (2024), e2867.
- [3] Meera Viswanathan et al. "Primary Care Interventions to Prevent Child Maltreatment: An Evidence Review for the US Preventive Services Task Force [Internet]". In: (2024).

- [4] Marie Cohen. "A Jumble of Standards: How State and Federal Authorities Have Underestimated Child Maltreatment Fatalities". In: (2024).
- [5] Kaitlyn Petruccelli, Joshua Davis, and Tara Berman. "Adverse childhood experiences and associated health outcomes: A systematic review and meta-analysis". In: *Child abuse & neglect* 97 (2019), p. 104127.
- [6] Vasiliki Tzouvara et al. "Adverse childhood experiences, mental health, and social functioning: A scoping review of the literature". In: *Child Abuse & Neglect* 139 (2023), p. 106092.
- [7] Paul Carlson. "Impact of adverse childhood experiences on academic achievement of school-aged learners". In: (2019).
- [8] Cora Peterson et al. "Economic burden of health conditions associated with adverse childhood experiences among US adults". In: *JAMA network open* 6.12 (2023), e2346323-e2346323.
- [9] Samantha M Brown et al. "The co-occurrence of adverse childhood experiences among children investigated for child maltreatment: A latent class analysis". In: *Child abuse & neglect* 87 (2019), pp. 18-27.
- [10] Christian M Connell et al. "Effects of community-based wraparound services on child and caregiver outcomes following child protective service involvement". In: *Child maltreatment* 29.1 (2024), pp. 190-201.
- [11] Jocelyn Brown et al. "A longitudinal analysis of risk factors for child maltreatment: Findings of a 17-year prospective study of officially recorded and self-reported child abuse and neglect". In: *Child abuse & neglect* 22.11 (1998), pp. 1065-1078.
- [12] *Risk and protective factors*. <https://www.cdc.gov/child-abuse-neglect/risk-factors/index.html>. [Online; accessed 10-June-2024]. 2024.
- [13] Kathryn Maguire-Jack, Tori Negash, and Kenneth J Steinman. "Child maltreatment prevention strategies and needs". In: *Journal of Child and Family Studies* 27.11 (2018), pp. 3572-3584.
- [14] Cora Peterson, Curtis Florence, and Joanne Elevens. "The economic burden of child maltreatment in the United States, 2015". In: *Child abuse & neglect* 86 (2018), pp. 178-183.
- [15] Ru Kwan et al. "Profile of hospitalised maltreated children aged 0 to 3 years and their families". In: *Singapore Medical Journal* (2024), pp. 10-4103.
- [16] Joanna Garstang et al. "Common factors in serious case reviews of child maltreatment where there is a medical cause of death: qualitative thematic analysis". In: *BMJ open* 11.8 (2021), e048689.
- [17] Government Accountability Office. "Child maltreatment: Strengthening national data on child fatalities could aid in prevention". In: *Report# GAO-11-99 prepared for the Chairman, Committee on Ways and Means, House of Representatives* (2011).
- [18] Brenda Jones Harden et al. "The child maltreatment prevention landscape: Where are we now, and where should we go?" In: *The ANNALS of the American Academy of Political and Social Science* 692.1 (2020), pp. 97-118.

- [19] Dyann Daley et al. "Risk terrain modeling predicts child maltreatment". In: *Child abuse & neglect* 62 (2016), pp. 29-38.
- [20] Joel M Caplan. "Mapping the spatial influence of crime correlates: A comparison of operationalization schemes and implications for crime analysis and criminal justice practice". In: *Cityscape* (2011), pp. 57-83.
- [21] Patricia Brantingham and Paul Brantingham. "Criminality of place: Crime generators and crime attractors". In: *European journal on criminal policy and research* 3 (1995), pp. 5-26.
- [22] Grant Drawve, Shaun A Thomas, and Jeffery T Walker. "Bringing the physical environment back into neighborhood research: The utility of RTM for developing an aggregate neighborhood risk of crime measure". In: *Journal of Criminal Justice* 44 (2016), pp. 21-29.
- [23] Jamaal W Green. "The built environment and predicting child maltreatment: An application of random forests to risk terrain modeling". In: *The Professional Geographer* 74.1 (2022), pp. 67-78.
- [24] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.
- [25] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1 (2020), p. 18.
- [26] Ruth MJ Byrne. "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning." In: *IJCAI*. 2019, pp. 6276-6282.
- [27] Zhendong Wang, Isak Samsten, and Panagiotis Papapetrou. "Counterfactual explanations for survival prediction of cardiovascular ICU patients". In: *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings*. Springer. 2021, pp. 338-348.
- [28] André Artelt et al. "Evaluating robustness of counterfactual explanations". In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021, pp. 01-09.
- [29] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 607-617.
- [30] Eoin Delaney, Derek Greene, and Mark T Keane. "Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions". In: *arXiv preprint arXiv:2107.09734* (2021).
- [31] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. "On counterfactual explanations under predictive multiplicity". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 809-818.
- [32] Thibault Laugel et al. "Issues with post-hoc counterfactual explanations: a discussion". In: *arXiv preprint arXiv:1906.04774* (2019).

- [33] Thibault Laugel et al. "Achieving diversity in counterfactual explanations: a re- view and discussion". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 1859–1869.
- [34] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. "Interpretable machine learning-a brief history, state-of-the-art and challenges". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2020, pp. 417–431.
- [35] Peter Schulam and Suchi Saria. "Reliable decision support using counterfactual models". In: *Advances in neural information processing systems* 30 (2017).
- [36] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
- [37] Shira Mitchell et al. "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions". In: *arXiv preprint arXiv:1811.07867v3* (2020).
- [38] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *FairWare' 18: IEEE/ACM International Workshop on Software Fairness, Gothenburg, Sweden*. 2018.
- [39] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Super- vised Learning". In: *arXiv preprint arXiv:1610.02413v1* (2016).
- [40] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. "Does miti- gating ML's disparate impact require disparate treatment?" In: *arXiv preprint arXiv:1711.07076v3* (2019).
- [41] Jordan Purdy and Brian Glass. "The pursuit of algorithmic fairness: On "correct- ing" algorithmic unfairness in a child welfare reunification success classifier". In: *Children and Youth Services Review* 145 (2023), p. 106777.
- [42] Cheryl Flynn et al. *Towards Algorithmic Fairness in Space-Time: Filling in Black Holes*. 2022. arXiv: 2211.04568 [stat.AP].
- [43] Ken Steif. *Public Policy Analytics: Code and Context for Data Science in Govern- ment*. <http://https://urbanspatial.github.io/PublicPolicyAnalytics/>. CRC Press, 2021.
- [44] David Olds et al. "Improving the nurse-family partnership in community practice". In: *Pediatrics* 132.Supplement_2 (2013), S110–S117.
- [45] Jeffrey A Butts et al. "Cure violence: a public health model to reduce gun violence". In: *Annual review of public health* 36 (2015), pp. 39–53.
- [46] Charles L Ransford, Karen Volker, and Gary Slutkin. "The Cure Violence Model for Violence Prevention". In: *Locally Led Peacebuilding: Global Case Studies* (2019), p. 171.
- [47] Sheyla A Delgado et al. "The effects of cure violence in the South Bronx and East New York, Brooklyn". In: (2017).
- [48] Paul Cozens. "Crime prevention through environmental design". In: *Environmental criminology and crime analysis*. Willan, 2013, pp. 175–199.

- [49] Laney A Rupp et al. "Community-engaged neighborhood revitalization and em- powerment: Busy streets theory in action". In: *American Journal of Community Psychology* 65.1-2 (2020), pp. 90-106.
- [50] Banuka De Silva et al. "Crime prevention through environmental design (CPTED): A brief review". In: *De Silva, KBN, Dharmasiri, KS, Buddhadasa, MPAA, Ranaweera, KGNU (2021). Crime prevention through environmental design (CPTED): A brief review. Academia Letters, Article 2337 (2021).*
- [51] Rebecca Hasdell. "What we know about universal basic income". In: *A cross- synthesis. Stanford, Basic Income Lab (2020).*
- [52] Lisa A Gennetian et al. "Behavioral insights into cash transfers to families with children". In: *Behavioral Science & Policy* 7.1 (2021), pp. 71-92.
- [53] Beth L Green et al. "Pathways to prevention: Early Head Start outcomes in the first three years lead to long-term reductions in child maltreatment". In: *Children and Youth Services Review* 118 (2020), p. 105403.
- [54] Inga Saitadze and Darejan Daji Dvalishvili. "Protective role of informal social support and early childhood programs in reducing likelihood of subsequent reports of child maltreatment". In: *Child Abuse & Neglect* 149 (2024), p. 106702.
- [55] Kate Welti and Jennifer Manlove. "Unintended pregnancy in Delaware: estimating change after the first two years of an intervention to increase contraceptive access". In: *Bethesda, MD: Child Trends (2018).*
- [56] Colorado Department of Public Health and Environment. *Taking the unintended out of pregnancy: Colorado's success with long-acting reversible contraception*. 2017.
- [57] Nelson C Malone et al. "Mobile health clinics in the United States". In: *Interna- tional journal for equity in health* 19 (2020), pp. 1-9.
- [58] Dana Weiner et al. "Mobile resource referral technology for preventive child welfare services: Implementation and feasibility". In: *Children and Youth Services Review* 107 (2019), p. 104499.
- [59] Laura Usher et al. "Crisis intervention team (CIT) programs: A best practice guide for transforming community responses to mental health crises". In: *CIT International (2019).*
- [60] Kathleen Bolter and Isabel McMullen. "The Kalamazoo Promise "Sweet 16," Sum- mary Study Results: 16 Key Findings from 16 Years Studying The Kalamazoo Promise". In: (2022).
- [61] Timothy J Bartik, Brad Hershbein, and Marta Lachowska. "The effects of the Kalamazoo Promise Scholarship on college enrollment and completion". In: *Jour- nal of Human Resources* 56.1 (2021), pp. 269-310.
- [62] Alyssa A Rheingold et al. "Child sexual abuse prevention training for childcare professionals: An independent multi-site randomized controlled trial of stewards of children". In: *Prevention Science* 16 (2015), pp. 374-385.
- [63] Matthew Sanders, Daryl Higgins, and Ronald Prinz. "A population approach to the prevention of child maltreatment: Rationale and implications for research, policy and practice". In: *Family Matters* 100 (2018), pp. 62-70.

- [64] Daryl J Higgins et al. "Core components of public health approaches to preventing child abuse and neglect". In: *Handbook of child maltreatment*. Springer, 2022, pp. 445-458.
- [65] Rhema Vaithianathan et al. "Using a machine learning tool to support high-stakes decisions in child protection". In: *AI Magazine* 42.1 (2021), pp. 53-60.
- [66] Devansh Saxena et al. "A human-centered review of algorithms used within the US child welfare system". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1-15.
- [67] *Historical Information: SACWIS/TACWIS Information*. <https://www.acf.hhs.gov/cb/training-technical-assistance/state-tribal-info-systems/historical-info>. [Online; accessed 10-June-2024]. 2024.
- [68] Jun-Jie Zhu, Meiqi Yang, and Zhiyong Jason Ren. "Machine learning in environmental research: common pitfalls and best practices". In: *Environmental Science & Technology* 57.46 (2023), pp. 17671-17689.
- [69] Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. "Machine learning for social science: An agnostic approach". In: *Annual Review of Political Science* 24 (2021), pp. 395-419.
- [70] Matthew Hindman. "Building better models: Prediction, replication, and machine learning in the social sciences". In: *The Annals of the American Academy of Political and Social Science* 659.1 (2015), pp. 48-62.
- [71] Riccardo Guidotti. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery* (2022), pp. 1-55.
- [72] Carlos Aguilera-Ventura et al. "Counterfactual reasoning via grounded distance". In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 19. 1. 2023, pp. 2-11.
- [73] Sandrine Pavoine et al. "On the challenge of treating various types of variables: application for improving the measurement of functional diversity". In: *Oikos* 118.3 (2009), pp. 391-402.
- [74] Mark T Keane et al. "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques". In: *arXiv preprint arXiv:2103.01035* (2021).
- [75] Dylan Slack et al. "Counterfactual explanations can be manipulated". In: *Advances in neural information processing systems* 34 (2021), pp. 62-75.