Referee-Meta-Learning for Fast Adaptation of Locational Fairness

Weiye Chen[1], Yiqun Xie[1*], Xiaowei Jia[2], Erhu He[2], Han Bao[3], Bang An[3], Xun Zhou[3]

[1]University of Maryland
[2]University of Pittsburgh
[3]University of Iowa
{weiyec, xie}@umd.edu, {xiaowei, erh108}@pitt.edu, {han-bao, bang-an, xun-zhou}@uiowa.edu

Abstract

When dealing with data from distinct locations, machine learning algorithms tend to demonstrate an implicit preference of some locations over the others, which constitutes biases that sabotage the spatial fairness of the algorithm. This unfairness can easily introduce biases in subsequent decision-making given broad adoptions of learning-based solutions in practice. However, locational biases in AI are largely understudied. To mitigate biases over locations, we propose a locational meta-referee (Meta-Ref) to oversee the few-shot meta-training and meta-testing of a deep neural network. Meta-Ref dynamically adjusts the learning rates for training samples of given locations to advocate a fair performance across locations, through an explicit consideration of locational biases and the characteristics of input data. We present a three-phase training framework to learn both a meta-learning-based predictor and an integrated Meta-Ref that governs the fairness of the model. Once trained with a distribution of spatial tasks, Meta-Ref is applied to samples from new spatial tasks (i.e., regions outside the training area) to promote fairness during the fine-tune step. We carried out experiments with two case studies on crop monitoring and transportation safety, which show Meta-Ref can improve locational fairness while keeping the overall prediction quality at a similar level.

Introduction

Locational bias has been widely studied in many social sectors and linked with social disparities of various types (Fan et al. 2020; Kontokosta and Hong 2021), such as the impact of climate change and related disasters (e.g., floods, food shortage), resource distribution (e.g., subsidies in agriculture), and infrastructure quality and safety. With the increasing of adoptions of machine learning methods in broad domains (e.g., climate resilience, food security, public resource management), fairness issues associated with these prediction models have become a major subject, directly impacting the trust from the public and the sustained use of the systems in the long-term. While fairness issues related to races or genders have been extensively examined in machine learning (Hardt, Price, and Srebro 2016; Dwork et al. 2012; Zafar et al. 2017; Agarwal et al. 2018; Creager et al. 2019; Kusner et al. 2017), very few studies have attempted to consider locational fairness. The lack of consideration for locational fairness in machine learning applications may result in unintended consequences (e.g., biased resource distribution). In this study, we exemplify this issue by examining two important social problems: (1) Agricultural monitoring: The climate change and the growing population have raised alarms and attention on global food security (e.g., G20's GEOGLAM initiative). With the broad deployment of machine learning in satellite-based crop monitoring (e.g., NASA Harvest), it is critical to explicitly consider locational fairness in mapping results, informing key decisions such as subsidy distribution or farm insurance (Bailey and Boryan 2010; NASEM 2018). (2) Transportation safety: Given the complexity of traffic accident risk estimation, machine learning algorithms have been increasingly used to account for heterogeneous information from diverse sources. However, without the awareness of locational fairness, these methods can produce biased risk maps, further adding bias in investments for infrastructure improvements (Kontokosta and Hong 2021; Bednarek, Boyce, and Sileo 2022).

We aim to create a new meta-learning framework that explicitly models locational fairness and enables rapid adaptation of fairness to new locations within different regions (e.g., different cities). Unlike traditional fairness formulation using pre-defined groups such as races and genders, locational fairness faces more challenges when being transferred between training and test data. First, in traditional fairness definitions, the groups considered in fairness evaluation are the same in the training and test datasets. In contrast, locational fairness often deals with entirely different sets of locations in the training region (e.g., city A) and test region (e.g., city B). Second, the change of spatial regions between training and test also introduces distribution shifts in the data. Third, fields such as agricultural monitoring require labor-intensive field survey, resulting in scarce availability of labeled data. Addressing these challenges requires the meta-learning model to learn the initial weights that not only can quickly adapt to the new distribution, but also adapt to an unknown fairness criterion (i.e., fairness defined on a new set of locations), from limited amount of labeled data. Several directions have been explored in related work:

Fair learning: Fairness-aware learning formulations have been extensively studied and most existing works focus on

[*]Corresponding author.

pre-defined groups (Mehrabi et al. 2021). A mainstream direction is to minimize the correlation between learned features with sensitive attributes, such as gender or race. The approaches include sensitive information encryption or removal (Kilbertus et al. 2018; Johndrow and Lum 2019), feature decorrelation (Zhao et al. 2022b), agnostic representation learning (Creager et al. 2019; Morales et al. 2020), representation neutralization (Du et al. 2021), regularization (Yan and Howe 2019), and so on. However, these methods do not consider the scenario faced in this problem, where the groups represented by locations involved in fairness evaluation are different from training to test.

Locational fairness: Recent studies (Xie et al. 2022; He et al. 2022, 2023) examined fairness formulations with respect to locations, and they focus on the case where space partitions are used for fairness evaluation. Similarly, they only consider problems where the spatial region remains the same from training to testing, and cannot address the issue of non-stationary groups.

Domain shifts: Domain adaptation methods mitigate covariance shift and learn invariant domains to reduce the effects of distribution shifts on model bias (Singh et al. 2021). Sample-reweighting and self-training approaches (Bickel, Brückner, and Scheffer 2007; An et al. 2022a; He et al. 2023) also aim to reduce the distribution gap between training and testing sets by assigning higher weights to samples more similar to test samples feature-wise, or include high-confidence pseudo-labels on test samples during training. In addition, heterogeneity-aware learning tackles variability by data partitioning and network branching (Xie et al. 2021, 2023). While these methods address distribution shifts, they also do not consider the changes of groups (locations) between training and test for fairness applications.

Meta-learning: Model-agnostic meta learning (MAML)'s gradient-by-gradient training allows it to learn an initial model that can be quickly fine-tuned to the test data with only a small number of observations (Finn, Abbeel, and Levine 2017; Ren et al. 2018; Xie et al. 2023; Chen et al. 2023). Recent developments have also started exploring the use of MAML in fairness-aware learning (Zhao et al. 2020, 2022a). These methods enable a fair model's prediction to remain independent from the sensitive attributes, and can let it adapt to changing distributions. However, they similarly have not considered the case where training and test sets have completely different groups (locations). Moreover, we focus on a different class of fairness definitions – prediction quality parity instead of protected attribute decorrelation (though both are commonly used standard definitions) (Zhang, Lemoine, and Mitchell 2018; Du et al. 2020) – for our targeted applications.

We propose a referee-meta-learning framework to address the challenges. Our contributions are:

- We propose a locational meta-referee (Meta-Ref) which learns to dynamically adjust learning rates of data samples in a task to make the prediction model fairer for samples at different locations after the gradient updates.
- We propose a three-phase training framework to update parameters of Meta-Ref and its corresponding prediction model using a distribution of spatial tasks.
- We experiment with real-world data for satellite-based crop classification and traffic accident risk estimation.

Our results on crop monitoring and transportation safety show that Meta-Ref can effectively improve fairness over locations in new test regions while keeping aggregated global performances similar to the baselines.

Concepts and Problem Formulation

The goal of this work is to mitigate locational biases in the prediction results, i.e., to reduce the variation of model performances, or **prediction quality disparity** (Du et al. 2020), over locations in a spatial region.

Definition 1 A **location** i is defined as a specific point or region in the geographical space, with s_i representing all data points associated with the location i. A data point can be an one-dimensional vector, a time-series, an image, etc.

Definition 2 **Locational fairness**, $\mathcal{L}_{fair}$, measures the locational biases of a deep neural network $\mathcal{F}$ using the performance scores $\mathcal{M} = \{m_{s_i}\}$ for data points from a set of distinct locations $\mathcal{S} = \{s_i\}$:

$$\mathcal{L}_{fair}(\mathcal{S}) = \sqrt{\frac{1}{|\mathcal{S}|}\sum_{s_i \in \mathcal{S}} (m_{s_i} - \hat{\mathcal{M}})^2} \tag{1}$$

where $\hat{\mathcal{M}}$ is the average performance for all data points.

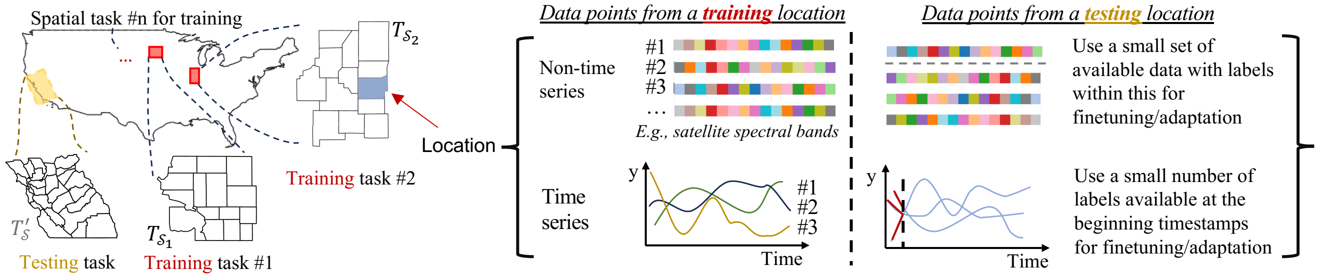
Definition 3 A **spatial task** $T_{\mathcal{S}}$ refers to the set of geolocated data points $\mathcal{S}$ in a study area of interest, where a deep neural network $\mathcal{F}$ with parameters $\mathbf{\Theta}$ is learned to make predictions. $T_{\mathcal{S}}$ also defines the set of locations where locational fairness is evaluated.

Fig. 1(a) shows an example of sampling spatial tasks for training and testing, with counties as distinct locations. Fig. 1(b) illustrates that a location may be associated with non-time series and/or time series data. A standard machine learning algorithm does not consider locational fairness, potentially resulting in an unfair distribution of prediction quality scores across the spatial task. Conversely, we expect a fairness-driven algorithm to enhance parity among locations in terms of prediction quality scores.

Problem Formulation. Given a set of spatial tasks $\{T_{\mathcal{S}_1}, T_{\mathcal{S}_2}, ...\}$ with associated features $\mathbf{X}$ and labels $\mathbf{y}$ from training locations, we aim to train a deep neural network $\mathcal{F}_{\mathbf{\Theta}}(\cdot)$ with awareness of locational fairness (Eq. (2)). The goal is that $\mathcal{F}_{\mathbf{\Theta}}(\cdot)$ can be quickly adapted to a new spatial task $T_{\mathcal{S}'}$ from test locations, where $T_{\mathcal{S}'} \cap \{T_{\mathcal{S}_1}, T_{\mathcal{S}_2}, ...\} = \phi$, using only a small amount of test samples $\mathbf{X}'$ and $\mathbf{y}'$ (Fig. 1(c)). The adaptation should consider both the prediction and fairness objectives.

Method

In this section, we introduce our locational meta-referee (Meta-Ref). Meta-Ref works in conjunction with any neural network-based prediction models and dynamically assigns learning rates to mini-batches to enforce fairness. Trained using a meta-learning framework, Meta-Ref can adapt to various spatial tasks and can be fine-tuned for unseen ones. We will provide details on its structure as well as its training and transfer strategies in the following sections.


Spatial task #n for training
...
$T_{\mathcal{S}_2}$
Location
Training task #2
$T'_{\mathcal{S}}$
$T_{\mathcal{S}_1}$
Testing task
Training task #1
Data points from a training location
Non-time series
#1
#2
#3
...
E.g., satellite spectral bands
Time series
y
#1
#2
#3
Time
Data points from a testing location
Use a small set of available data with labels within this for finetuning/adaptation
y
Time
Use a small number of labels available at the beginning timestamps for finetuning/adaptation

(a) Composition of spatial tasks (b) A location and its data points $s_i \in \mathcal{S}$ (c) Adaptation of fairness for an unseen spatial task

Figure 1: Illustrative examples of spatial tasks for training and testing, their encompassing locations and data points, as well as the data requirements for finetuning/adaption for spatial tasks from test locations.

Model-Agnostic Meta-Learning

Model-Agnostic Meta-Learning (MAML) (Finn, Abbeel, and Levine 2017) is a scheme which trains a generalizable model that can be quickly adapted to new tasks through gradient-by-gradient update strategies. It moderates and learns through gradient updates of over a set of tasks and finds a gradient leading to better generalization. The goal of MAML is to learn an initial model from a distribution of tasks such that the initial model can be quickly fine-tuned to the optimal parameters of individual tasks using only a few samples. Given a distribution of tasks $\{T_i | T_i \sim p(\mathcal{T})\}$, each gradient update in MAML is given by:

$$\mathbf{\Theta}'_i \leftarrow \mathbf{\Theta} - \beta \nabla_{\mathbf{\Theta}} \mathcal{L}(\mathbf{X}^{(i)}, \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y}^{(i)}) \tag{2}$$

$$\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \alpha \sum_{T_i \sim p(\mathcal{T})} \nabla_{\mathbf{\Theta}} \mathcal{L}(\tilde{\mathbf{X}}^{(i)}, \mathcal{F}_{\mathbf{\Theta}'_i}, \tilde{\mathbf{y}}^{(i)}) \tag{3}$$

where $\mathbf{\Theta}'_i$ represents temporary parameters of the deep neural network $\mathcal{F}$ for a task T_i; α and β are the hyper-parameters of the step size of gradient updates; $\mathbf{X}^{(i)}$ and $\mathbf{y}^{(i)}$ are the training mini-batches, and $\tilde{\mathbf{X}}^{(i)}$ and $\tilde{\mathbf{y}}^{(i)}$ are the validation mini-batches for task T_i, respectively.

This combination of task-specific and global gradient update simulates the scenarios encountered in testing, where we are given one initial model and aim to reach good performance of a new task after updating with a mini-batch. Thus, using the sequence of gradient updates (gradients of gradients), MAML learns a set of parameters that are not necessarily optimal for any given task, but can be quickly adapted to one specific task with a small number of points. In this work, we define the tasks using spatial tasks $\{T_{\mathcal{S}}\}$, which contain geo-located data points from different spatial regions.

A Locational Meta-Referee

We propose a locational meta-referee (Meta-Ref), which aims to adapt fairness learned from a distribution of spatial tasks $\{T_{\mathcal{S}_1}, T_{\mathcal{S}_2}, ...\}$ in the training area to a new spatial task $T_{\mathcal{S}'}$ in the test region. Fig. 2 provides an illustration of the framework during training. The main ideas of Meta-Ref are:

- It learns to produce learning rates for different locations within each spatial task as a function of the locations' features and the prediction model's relative performances on them, using only a few observations.
- The prediction model is optimized with variable learning rates dynamically estimated by Meta-Ref, with the goal of collaboratively minimizing the fairness loss over the locations in a given spatial task.
- To quickly adapt to a new spatial task with a different set of locations, Meta-Ref creates a distribution of spatial tasks during training by randomly generating local subsets of data points at different locations, where the training spatial tasks may overlap with each other.

In other words, Meta-Ref $\mathcal{F}^{MR}$ with parameters $\mathbf{W}$ is trained in coordination with the prediction model $\mathcal{F}$ to learn a mechanism to enforce fairness on a given spatial task and make sure the mechanism is easily transferable across a distribution of spatial tasks.

Specifically, Meta-Ref takes three types of inputs: (1) Performance metrics (e.g., RMSE) of data sample generated by the current prediction model $\mathcal{F}$'s parameters $\mathbf{\Theta}$, which help evaluate the current performance on the data points and their potential impact on the locational fairness; (2) The global performance metrics to benchmark the level of locational fairness and convert absolute performance metrics to relative scores; and (3) The encoding generated by $\mathcal{F}$ over data samples. The encoding reflects the characteristics of samples that can better guide the learning rate estimation. For example, a large loss may not always entail a high learning rate as a sample may be a very difficult case, whose loss can hardly be further reduced without causing significant negative impacts on other samples. Denoting the encoding process of the prediction model as $\mathcal{F}^{enc}$, we consider the prediction model as $\mathcal{F}(\cdot) = \mathcal{F}^{dec}(\mathcal{F}^{enc}(\cdot))$, where $\mathcal{F}^{dec}(\cdot)$ is the decoder. Meta-Ref explicitly considers the performance metric $m_i = M(\mathbf{X}_i, \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y}_i)$ yielded by the prediction model on data samples. The performance metrics are further standardized by subtracting the global performance $\hat{\mathcal{M}}$ to obtain relative performances, so that Meta-Ref becomes invariant of the state of the overall performance, making it more transferable. Formally, we represent Meta-Ref as a neural network $\mathcal{F}^{MR}$ with parameters $\mathbf{W}$, which outputs a fairness factor η_i for each location s_i in a spatial task $T_{\mathcal{S}}$ using:

$$\eta_i = \mathcal{F}^{MR}_{\mathbf{W}}(\mathcal{F}^{enc}(\mathbf{X}_i), M(\mathbf{X}_i, \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y}_i)) - \hat{\mathcal{M}}) \tag{4}$$

The fairness factors will be translated into learning rates during meta-training, detailed in the next section.

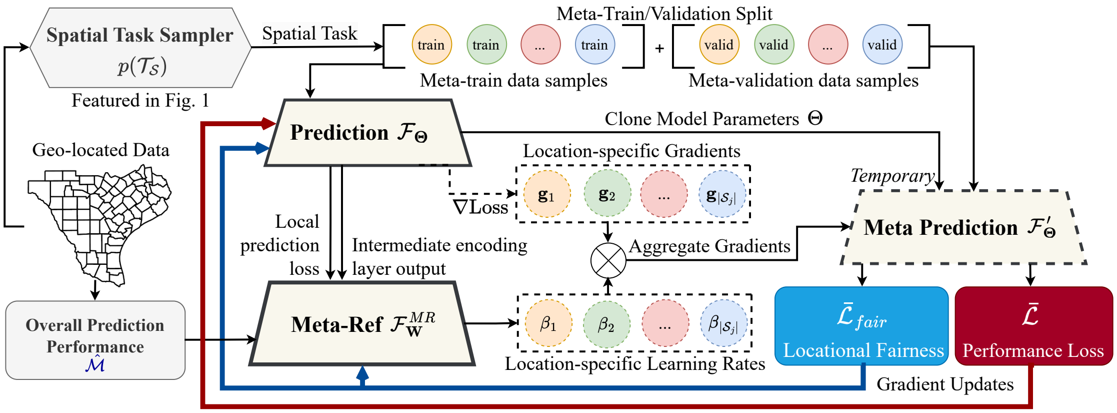

Spatial Task Sampler
$p(\mathcal{T}_\mathcal{S})$
Featured in Fig. 1
Spatial Task
Meta-Train/Validation Split
train
train
...
train
+
valid
valid
...
valid
Meta-train data samples
Meta-validation data samples
Geo-located Data
Prediction $\mathcal{F}_\Theta$
Clone Model Parameters Θ
Location-specific Gradients
∇Loss
$\mathbf{g}_1$
$\mathbf{g}_2$
...
$\mathbf{g}_{|\mathcal{S}_j|}$
Temporary
Meta Prediction $\mathcal{F}'_\Theta$
Local prediction loss
Intermediate encoding layer output
Aggregate Gradients
Overall Prediction Performance
$\hat{\mathcal{M}}$
Meta-Ref $\mathcal{F}^{MR}_\mathbf{W}$
β_1
β_2
...
$\beta_{|\mathcal{S}_j|}$
Location-specific Learning Rates
$\bar{\mathcal{L}}_{fair}$
Locational Fairness
$\bar{\mathcal{L}}$
Performance Loss
Gradient Updates

Figure 2: An illustration of the training framework to enforce locational fairness with Meta-Ref.

Three-Phase Training of Meta-Ref

We apply MAML's gradient-by-gradient update strategies to train both Meta-Ref and the prediction model. Below, we present a three-phase framework for the training process.

Generation of spatial task distribution $\mathcal{T}(\mathcal{S})$. Before the start of the three-phase training, it is important to first generate a diverse distribution of spatial tasks so that we can learn a more transferable initial set of parameters for the prediction model $\mathcal{F}$ in conjunction with Meta-Ref. We combine two strategies to generate spatial tasks for the distribution $\mathcal{T}(\mathcal{S})$: (1) Locations are grouped according to administrative boundaries (e.g., cities, states), zones, or their attributes (e.g., precipitation, altitude) to create spatial tasks. To enlarge the variety of spatial tasks, the elements of two spatial tasks do not need to be disjoint, so there's a possibility that $\exists \mathcal{S}_i, \mathcal{S}_j \in \{\mathcal{S}\} : \mathcal{S}_i \cap \mathcal{S}_j \neq \emptyset$. (2) To create spatial tasks that are more distinctive from regular ones from the training area, we further include spatial tasks whose locations are randomly sampled from the entire training area.

Phase 1: Prediction performance estimation. Prior to each epoch of training, we begin by sampling an array of spatial tasks $\mathbf{T} = \{T_{\mathcal{S}_i} \sim p(\mathcal{T}(\mathcal{S}))\}$. The first step is to generate the necessary inputs for Meta-Ref according to Eq. (4). In this phase, Meta-Ref gains knowledge of how the prediction model performs on spatial tasks and assesses the level of locational biases in them. On a training mini-batch $(\mathbf{X}_i^{(j)}, \mathbf{y}_i^{(j)})$ at location $s_i^{(j)}$ in the spatial task $T_{\mathcal{S}_j}$, we assess the prediction loss as $l_i^{(j)} = \mathcal{L}(\mathbf{X}_i^{(j)}, \mathcal{F}_\Theta, \mathbf{y}_i^{(j)})$ and the corresponding metrics $m_i^{(j)} = M(\mathbf{X}_i^{(j)}, \mathcal{F}_\Theta, \mathbf{y}_i^{(j)})$.

We apply the same procedure on all locations in a spatial task $T_{\mathcal{S}_j}$, getting an losses and metrics, $\mathbf{L}^{(j)}$ and $\mathbf{M}^{(j)}$:

$$\mathbf{L}^{(j)} = \left[l_i^{(j)} \middle| i \in [1, |\mathcal{S}_j|] \right], \mathbf{M}^{(j)} = \left[m_i^{(j)} \middle| i \in [1, |\mathcal{S}_j|] \right] \quad (5)$$

With loss values calculated for all locations, we compute gradients of each location as $\mathbf{g}_i^{(j)} = \nabla_\Theta l_i^{(j)}$. Additionally, we evaluate the overall performance for all training data points regardless of their location origin $(\mathbf{X}, \mathbf{y})$:

$$\hat{\mathcal{M}} = M(\mathbf{X}, \mathcal{F}_\Theta, \mathbf{y}) \quad (6)$$

Phase 2: Fairness-aware learning rate estimation. Using outputs $\mathbf{L}^{(j)}$, $\mathbf{M}^{(j)}$, $\mathbf{g}_i^{(j)}$ and $\hat{\mathcal{M}}$ from Phase 1, Meta-Ref dynamically adjusts the step sizes of gradients associated with data points at different locations in a spatial task:

$$\Theta'^{(j)} \leftarrow \Theta - \sum_{s_i \in \mathcal{S}_j} \beta_i^{(j)} \mathbf{g}_i^{(j)} \quad (7)$$

where $\beta_i^{(j)}$ represents the learning rate of data points in location $s_i^{(j)}$, assigned by Meta-Ref, and $\Theta'^{(j)}$ is the temporary parameters of $\mathcal{F}$ through updates on data points of all locations. Different from MAML (Eq. (2)), step sizes are no longer identical over all data points in a mini-batch; instead, they become dependent on the locations and spatial tasks. Step sizes are assigned via translating Meta-Ref-generated fairness factors $\mathbf{N}^{(j)} = \left[\eta_i^{(j)} \middle| i \in [1, |\mathcal{S}_j|] \right]$ with Eq. (4). Specifically, we perform the translation by standardizing the fairness factors from all locations within each spatial task into a stationary range to improve the stability of training:

$$\tilde{\eta}_i^{(j)} = \frac{\eta_i^{(j)} - \min(\mathbf{N}^{(j)})}{\max(\mathbf{N}^{(j)}) - \min(\mathbf{N}^{(j)})} \quad (8)$$

$$\beta_i^{(j)} = \tilde{\eta}_i^{(j)} \times (\beta^+ - \beta^-) + \beta^- \quad (9)$$

where $\eta_i^{(j)} = \mathcal{F}^{MR}_\mathbf{W}\left(\mathcal{F}^{enc}_\Theta(\mathbf{X}_i^{(j)}), m_i^{(j)} - \hat{\mathcal{M}} \right)$; and β^+ and β^- represent the upper and lower limits of step size of the gradient update, respectively.

To stabilize training at the early stage, we constrain the variance of learning rates across data samples from different locations, var $\left(\left\{ \beta_i^{(j)} \middle| i \in [1, |\mathcal{S}_j|] \right\} \right)$, by adjusting β^+ and β^-. Given a baseline learning rate β^0 and a scaling factor ρ, we set the upper and lower bounds of $\beta_i^{(j)}$ at iteration t as $\beta^+ = \frac{1}{1+\mathrm{e}^{-t/\rho}} \cdot \beta^0$ and $\beta^- = \frac{\mathrm{e}^{-t/\rho}}{1+\mathrm{e}^{-t/\rho}} \cdot \beta^0$. The gap between these bounds expands gradually through a sigmoid-shaped curve as training progresses, so after the early training stage learning rates approach a constant range, allowing more flexibility to improve fairness.

Finally, the location-dependent learning rates are applied via Eq. (7) for gradient updates on the temporary prediction model parameters within the inner meta-training loop.

Phase 3: Dual meta-updates. In this dual meta-update phase, we consider both the prediction performance and the locational fairness to make final gradient updates. Specifically, we use two different losses on the validation data for a spatial task $T_{\mathcal{S}_j}$, prediction loss $\bar{\mathcal{L}}^{(j)}$ and locational fairness loss $\bar{\mathcal{L}}^{(j)}_{fair}$, to meta-update the parameters of the prediction model and Meta-Ref. The prediction loss measures the collective performance of temporarily updated prediction model parameters $\tilde{\eta}_i^{(j)}$, and the locational fairness loss reflects the effectiveness on the coordination of Meta-Ref and the prediction model. We compute two losses using:

$$\bar{\mathcal{L}}^{(j)} = \mathcal{L}(\tilde{\mathbf{X}}^{(j)}, \mathcal{F}_{\mathbf{\Theta}'^{(j)}}, \tilde{\mathbf{y}}^{(j)}) \tag{10}$$

$$\bar{\mathcal{L}}^{(j)}_{fair} = \mathcal{L}_{fair}(\mathcal{S}_j) = \sqrt{\frac{1}{|\mathcal{S}_j|}\sum_{s_i \in \mathcal{S}_j} (\tilde{m}_i^{(j)} - \hat{\mathcal{M}})^2} \tag{11}$$

where $\tilde{m}_i^{(j)} = M(\tilde{\mathbf{X}}_i^{(j)}, \mathcal{F}_{\mathbf{\Theta}'^{(j)}}, \tilde{\mathbf{y}}_i^{(j)})$; $(\tilde{\mathbf{X}}^{(j)}, \tilde{\mathbf{y}}^{(j)})$ are validation data from the whole $\mathcal{S}_j$; and $(\tilde{\mathbf{X}}_i^{(j)}, \tilde{\mathbf{y}}_i^{(j)})$ are validation mini-batch sampled from location s_i.

For the prediction loss $\bar{\mathcal{L}}^{(j)}$, we use its gradients to update only the prediction model. For $\bar{\mathcal{L}}^{(j)}_{fair}$, we use its gradients to update both the prediction model and Meta-Ref. In this way, Meta-Ref focuses on the fairness side, and it coordinates with the prediction model to address both prediction performance and fairness:

$$\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \alpha_1 \nabla_{\mathbf{\Theta}} \bar{\mathcal{L}}^{(j)} \tag{12}$$

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha_2 \nabla_{\mathbf{W}} \bar{\mathcal{L}}^{(j)}_{fair} \tag{13}$$

$$\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \alpha_3 \nabla_{\mathbf{\Theta}} \bar{\mathcal{L}}^{(j)}_{fair} \tag{14}$$

where α_1, α_2, and α_3 are learning rates set for the three meta-update operations, respectively. Through Eqs. (10 - 14), we can see that each of the three-way meta-update involves gradients over gradients (since $\mathbf{\Theta}'^{(j)}$ remains in expanded forms of Eqs. (12 - 14)). This makes the gradient updates consider the contribution of each location in a spatial task, and mimic the actual fine-tuning process during testing. Algorithm 1 summarizes the procedure of the three-phase training of Meta-Ref.

Fine-Tuning on Test Region

Meta-trained parameters of the prediction model are expected to demonstrate good generalizability but not necessarily optimal to any tasks. Though, we do not fine-tune Meta-Ref, considering that it is not directly related to the prediction performance, while also avoiding overfitting.

Given a spatial task in the test area $T_{\mathcal{S}'}$, where $T_{\mathcal{S}'} \cap \{T_{\mathcal{S}_1}, T_{\mathcal{S}_2}, ...\} = \phi$, we fine-tune the test data $\mathbf{X}'$ and $\mathbf{y}'$ in a slightly different fashion than three-phase meta-updates. The fine tuning has two phases: 1) Prediction performance estimation, and 2) Meta-Ref-guided optimization.

Algorithm 1: Three-Phase Training of Meta-Ref

```
Require: $p(\mathcal{T}_{\mathcal{S}})$: distribution of spatial tasks
Parameters: $\alpha_1, \alpha_2, \alpha_3, \beta^0, \rho$
sample batch of spatial tasks $\mathbf{T} = \{T_{\mathcal{S}_j} \sim p(\mathcal{T})\}$
for all $T_{\mathcal{S}_j} \in \mathbf{T}$ do
    [Phase 1]
    for all $s_i^{(j)} \in \mathcal{S}_j$ do
        Evaluate local prediction loss $l_i^{(j)} = \mathcal{L}(\mathbf{X}_i^{(j)}, \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y}_i^{(j)})$
    end for
    Evaluate global prediction loss $\hat{\mathcal{M}} = M(\mathbf{X}, \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y})$
    [Phase 2]
    Assign step size $\beta_i^{(j)}$ with Eqs. (8 - 9) for $s_i^{(j)} \in \mathcal{S}_j$
    Update $\mathbf{\Theta}'^{(j)} \leftarrow \mathbf{\Theta} - \sum_{s_i \in \mathcal{S}_j} \beta_i^{(j)} \mathbf{g}_i^{(j)}$ [Eq. 7]
    [Phase 3]
    Evaluate $\bar{\mathcal{L}}^{(j)}$ with Eq. (10)
    Evaluate $\bar{\mathcal{L}}^{(j)}_{fair}$ with Eq. (11)
    Update $\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \alpha_1 \nabla_{\mathbf{\Theta}} \bar{\mathcal{L}}^{(j)}$
    Update $\mathbf{W} \leftarrow \mathbf{W} - \alpha_2 \nabla_{\mathbf{W}} \bar{\mathcal{L}}^{(j)}_{fair}$
    Update $\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \alpha_3 \nabla_{\mathbf{\Theta}} \bar{\mathcal{L}}^{(j)}_{fair}$
end for
```

Phase 1 of fine-tuning is similar to the Phase 1 of training Meta-Ref on a spatial task, with the differences being that the input data are from different regions, and that the overall performance metrics $\hat{\mathcal{M}}$ remains being calculated from the training data. Formally, we evaluate the local prediction loss on every location $s_i' \in \mathcal{S}'$ with $l_i' = \mathcal{L}(\mathbf{X}_i', \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y}_i')$ and the global overall performance metrics $\hat{\mathcal{M}} = M(\mathbf{X}, \mathcal{F}_{\mathbf{\Theta}}, \mathbf{y})$.

Phase 2 follows with the assignment of learning rates of all locations within this spatial task, following Eqs. (8 - 9) using Meta-Ref, producing β_i' for each location in the test region. Then we optimize $\mathbf{\Theta}$ with $\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \sum_{s_i' \in \mathcal{S}'} \beta_i' \nabla_{\mathbf{\Theta}} l_i'$. This two-phase fine-tuning effectively simulates the behavior of Eq. (7) where we update the prediction model with Meta-Ref-assigned learning rates.

Experiments

Case Study Datasets

Satellite-based crop classification: Crop mapping is important for various downstream tasks including acreage estimation, subsidy distribution and farm insurance. Our study area is a ∼6700 km^2 region in Central Valley, California, which is a major region in the US for walnut plantation. The satellite imagery we use is from Sentinel-2 multispectral data. As the 10 spectral bands we use from Sentinel-2 have different spatial resolution (e.g., 10m, 20m), we sample all bands to 20m (with an image tile size of 4096 × 4096), a common choice in applications. The image tile was captured in August 2018. The labels are from the USDA Crop Data Layer (CDL) (CDL 2017). For walnut plantation mapping, we preprocess the labels into binary walnut and non-walnut classes. Since the fairness paradigm is based on prediction quality parity, fairness calculation needs the performances from the locations as inputs. For classification, this requires a certain level of aggregation (e.g., F1 or accuracy is not meaningful for an individual point). In our experiment, locations are thus represented by 128×128 non-overlapping local patches from the image tile instead of individual pixels. We use a 50% by 50% train-test split for the locations. In a test location, 5% of randomly sampled data points are used for fine-tuning. Each spatial task $\mathcal{S}_i$ is randomly sampled from training or test locations covered by a random 1280×1280 window (a ∼25km×25km region), with the number of locations ranging from 10 to 15 per spatial task.

Traffic accident risk estimation: Location-based biases in transportation safety estimation can further lead to biases for investment distribution for infrastructure improvements. We use the Iowa traffic accident record dataset shared by (An et al. 2022b), which contains 3 years of traffic accident records and 47 related factors. The dataset has a daily temporal resolution and was spatially aggregated into grid cells, and the total grid size is 64×128 (An et al. 2022b). We use each cell as a location for this regression problem. We partition the dataset into 8-week moving windows, where we use factors in the first 7 weeks to predict the average daily count of accidents in the eighth week. Similarly, we use a 50% by 50% train-test split for the locations. In a test location, only the first 5% of moving windows are used for fine-tuning. Each spatial task is randomly sampled from training or testing locations in a randomly-selected 32×32 window, with 10 to 15 locations per spatial task.

Methods for Comparison

We evaluate the following methods in terms of prediction performance and, particularly, locational fairness: (1) **DNN** and **LSTM**: Plain baselines of neural network, including fully-connected deep neural network for crop classification with image snapshots (non-time-series) and Long-short-term memory (LSTM) model for traffic accident prediction with time series inputs. (2) **Reg**: DNN or LSTM with an additional regularization term (i.e., variance) to enforce locational fairness, a common strategy for prediction quality parity (Kamishima, Akaho, and Sakuma 2011; Yan and Howe 2019). (3) **Adv**: An adversarial training method to learn a location-neutral representation of data samples, inspired by (Zhang, Lemoine, and Mitchell 2018; Alasadi, Al Hilli, and Singh 2019). (4) **Domain**: We add domain adaptation to plain baselines (**Dom-DNN**, **Dom-LSTM**) and **Reg** (**Dom-Reg**), where a discriminator is used to learn domain-invariant features. This can help bridge potential domain gaps between tasks in training and testing. (5) **Bi-Lvl**: A recent state-of-the-art for improving locational fairness by adjusting fairness based on relative performances among samples (Xie et al. 2022; He et al. 2023). It is designed for tasks in the same region and does not consider generalization to new tasks. (6) **MAML**: Model-agnostic meta-learning (Finn, Abbeel, and Levine 2017), designed for fast adaption to new tasks. MAML does not consider the fast adaptation of locational fairness. (7) We compare the aforementioned methods against **Meta-Ref**: Our proposed approach. More details can be found in the technical appendix.

Evaluation Metrics

To evaluate the prediction quality of a test spatial task $T_{\mathcal{S}'}$ from crop classification, we use F1-score p_i to account for the prediction quality for data points at location $s_i \in \mathcal{S}'$. For a regression task in traffic accident risk estimation, we adopt the root mean squared error (RMSE) for p_i. We assess the locational fairness (LF) on the test spatial task from the evaluation metrics at each location by taking their standard deviation. Some methods producing poor prediction quality might achieve better locational fairness on some spatial tasks. Therefore, we also include an adjusted locational fairness (ALF) metric to account for differences in prediction quality when evaluating fairness. Instead of using the mean prediction performance from the current method to calculate the standard deviation, we use the best prediction performance for this task among all methods as the reference mean, denoted as p^*. Then the adjusted fairness score for a spatial task is defined as the average deviation of the prediction quality $\{p_i\}$ from this reference p^*:

$$ALF = \sqrt{\frac{1}{|\mathcal{S}'|} \sum_{s_i \in \mathcal{S}'} (p_i - p^*)^2} \tag{15}$$

By setting the mean to the best prediction performance among all methods, methods that trade performance for fairness (i.e., low variance among $\{p_i\}$ but higher distance to p^*) will be penalized and produce worse scores in ALF.

Results

Training: We sample 1000 spatial tasks from training locations of each dataset. Then we sample another 90 spatial tasks from testing locations from each dataset and split them into three folds for testing. All methods are fine-tuned in each of the 90 spatial tasks in test regions for both datasets.

Comparison to baselines: Table 1 and 2 show the average performance of different models on two datasets. MAML and Meta-Ref have similar performance in terms of prediction metrics on both datasets. In crop classification tasks, these two methods outperform other baselines by significant margins in both performance and fairness measures. In traffic accident risk estimation dataset, Domain methods are close to MAML in both performance and fairness but never surpass MAML. Meta-Ref has reliably demonstrated a lead over MAML in both locational fairness (LF)

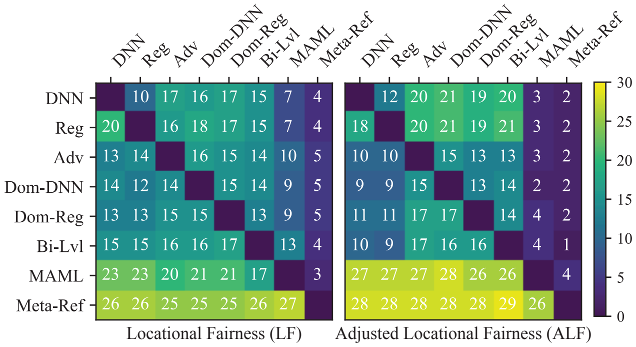


Figure 3: A pairwise comparison matrix for all methods (This is an example from task set 1 for crop classification with 30 tasks in total).

	Task set 1			Task set 2			Task set 3		
	F1	LF	ALF	F1	LF	ALF	F1	LF	ALF
DNN	0.681	0.130	0.138	0.680	0.134	0.141	0.677	0.133	0.139
Reg	0.680	0.129	0.137	0.680	0.134	0.141	0.676	0.133	0.139
Adv	0.660	0.130	0.147	0.640	0.127	0.158	0.639	0.139	0.158
Dom-DNN	0.660	0.133	0.148	0.652	0.148	0.164	0.654	0.143	0.157
Dom-Reg	0.664	0.130	0.144	0.656	0.143	0.159	0.659	0.138	0.151
Bi-Lvl	0.657	0.128	0.147	0.637	0.126	0.156	0.641	0.133	0.157
MAML	0.719	0.119	0.119	0.718	0.124	0.124	0.714	0.126	0.126
Meta-Ref	0.716	**0.107**	**0.107**	0.710	**0.114**	**0.115**	0.706	**0.118**	**0.119**

Table 1: Average metrics for satellite-based crop classification on 90 spatial tasks from test locations.

	Task set 1			Task set 2			Task set 3		
	RMSE	LF	ALF	RMSE	LF	ALF	RMSE	LF	ALF
LSTM	0.139	0.115	0.116	0.137	0.102	0.104	0.150	0.121	0.125
Reg	0.146	0.126	0.128	0.139	0.107	0.110	0.138	0.108	0.109
Adv	0.141	0.129	0.133	0.140	0.111	0.114	0.149	0.110	0.114
Dom-LSTM	0.137	0.110	0.114	0.132	0.089	0.090	0.135	0.088	0.089
Dom-Reg	0.139	0.110	0.115	0.137	0.089	0.092	0.136	0.088	0.090
Bi-Lvl	0.149	0.133	0.135	0.146	0.108	0.112	0.150	0.110	0.114
MAML	0.134	0.108	0.110	0.133	0.084	0.087	0.137	0.087	0.089
Meta-Ref	0.131	**0.102**	**0.103**	0.136	**0.081**	**0.083**	0.138	**0.085**	**0.086**

Table 2: Average metrics for traffic accident risk estimation on 90 spatial tasks from test locations.

and adjusted locational fairness (ALF) on both datasets. The leading positions of Meta-Ref in ALF across both datasets suggest that Meta-Ref does not demonstratively sacrifice its prediction performance for fairness. Meta-Ref tends to generate fairer results even when setting the benchmark to the best-performing model, if not Meta-Ref, for most test tasks. In Fig. 3 and also in the technical appendix, we demonstrate pairwise comparison matrices, where each element indicates the number of spatial tasks where the row method has lower fairness metrics (LF or ALF) than column method. It shows that Meta-Ref maintain better fairness on most spatial tasks compared to baselines. In addition, as shown in Fig. 4, Meta-Ref has demonstrated fairer predictions than MAML on most tasks, confirming its effectiveness.

Sensitivity analysis: The training of Meta-Ref relies on three outer gradient updates (Eqs. (12, 13, 14)) to coordinate the performance loss and fairness loss with their impacts on prediction and meta-referee. To demonstrate the effectiveness of three outer gradient updates in Meta-Ref, we further conduct an ablation study with the following models: (1) **MR-P2P**: Meta-Ref without applying performance gradient to the prediction model (P2P, Eq. (12)); (2) **MR-F2M**: Meta-Ref without applying fairness gradient to the meta-referee (F2M, Eq. (13)); and (3) **MR-F2P**: Meta-Ref without applying fairness gradient to the prediction model (F2P, Eq. (14)). The experiment results suggest that all gradient updates are essential in the training of Meta-Ref. Without P2P, the model fails to generalize with spatial tasks. Without F2M, meta-referee is no longer updated during training, producing arbitrary locational fairness factors that disrupt the training of prediction model. Trimming F2P from Meta-Ref has the smallest impact among three ablation models, since while MR-F2P still underperforms full MAML and Meta-Ref, it has better prediction and location fairness metrics than other baselines. The technical appendix provides more details about this analysis as well as the robustness of hyperparameters and results.

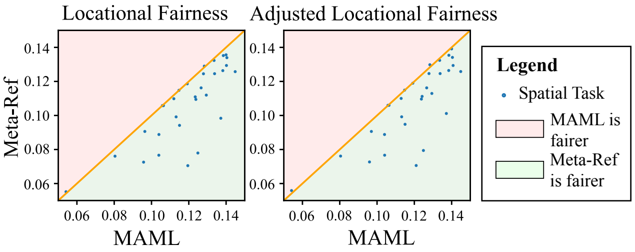


Figure 4: Comparison between MAML and Meta-Ref on fairness metrics among different spatial tasks (an example from task set 1 for crop classification with 30 tasks in total).

Conclusions

This study presented Meta-Ref, a locational fairness meta-referee to address the implicit biases of a neural network on geo-located data. Trained using a meta-learning framework, Meta-Ref attributes the scale of locational biases with characteristics of data and training performance dynamics, and assigns learning rates to data points from different locations. Meta-Ref can be applied to a fine-tuned prediction model on location sets that have never been seen during meta-training. Case studies on crop monitoring and transportation safety showed that Meta-Ref can effectively improve locational fairness. Our future work will explore domain customizations to facilitate implementation in real practices.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2105133, 2126474 and 2147195; NASA under Grant No. 80NSSC22K1164 and 80NSSC21K0314; USGS grants G21AC10207, G21AC10564, and G22AC00266; Google's AI for Social Good Impact Scholars program; the DRI award and the Zaratan supercomputing cluster at the University of Maryland; Pitt Momentum Funds award and CRC at the University of Pittsburgh; and the ISSSF grant from the University of Iowa.

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 60–69. PMLR.

Alasadi, J.; Al Hilli, A.; and Singh, V. K. 2019. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 19–25.

An, B.; Che, Z.; Ding, M.; and Huang, F. 2022a. Transferring Fairness under Distribution Shifts via Fair Consistency Regularization. *arXiv preprint arXiv:2206.12796*.

An, B.; Vahedian, A.; Zhou, X.; Street, W. N.; and Li, Y. 2022b. HintNet: Hierarchical Knowledge Transfer Networks for Traffic Accident Forecasting on Heterogeneous Spatio-Temporal Data. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, 334–342. SIAM.

Bailey, J. T.; and Boryan, C. G. 2010. Remote sensing applications in agriculture at the USDA National Agricultural Statistics Service. Technical report, Research and Development Division, USDA, NASS, Fairfax, VA.

Bednarek, A.; Boyce, A.; and Sileo, A. 2022. Infrastructure investments should be evidence informed and equity focused.

Bickel, S.; Brückner, M.; and Scheffer, T. 2007. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, 81–88.

CDL. 2017. Cropland Data Layer - USDA NASS. https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php. Accessed: 03/20/2022.

Chen, S.; Xie, Y.; Li, X.; Liang, X.; and Jia, X. 2023. Physics-Guided Meta-Learning Method in Baseflow Prediction over Large Regions. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 217–225. SIAM.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

Du, M.; Mukherjee, S.; Wang, G.; Tang, R.; Awadallah, A.; and Hu, X. 2021. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34: 12091–12103.

Du, M.; Yang, F.; Zou, N.; and Hu, X. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4): 25–34.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Fan, C.; Esparza, M.; Dargin, J.; Wu, F.; Oztekin, B.; and Mostafavi, A. 2020. Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters. *Computers, Environment and Urban Systems*, 83: 101514.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

He, E.; Xie, Y.; Jia, X.; Chen, W.; Bao, H.; Zhou, X.; Jiang, Z.; Ghosh, R.; and Ravirathinam, P. 2022. Sailing in the location-based fairness-bias sphere. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–10.

He, E.; Xie, Y.; Liu, L.; Chen, W.; Jin, Z.; and Jia, X. 2023. Physics Guided Neural Networks for Time-Aware Fairness: An Application in Crop Yield Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14223–14231.

Johndrow, J. E.; and Lum, K. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1): 189–220.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650. IEEE.

Kilbertus, N.; Gascón, A.; Kusner, M.; Veale, M.; Gummadi, K.; and Weller, A. 2018. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, 2630–2639. PMLR.

Kontokosta, C. E.; and Hong, B. 2021. Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society*, 64: 102503.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Morales, A.; Fierrez, J.; Vera-Rodriguez, R.; and Tolosana, R. 2020. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2158–2164.

NASEM. 2018. *Improving crop estimates by integrating multiple data sources*. National Academies Press.

Ren, H.; et al. 2018. Learning with Weak Supervision from Physics and Data-Driven Constraints. *AI Magazine*.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3–13.

Xie, Y.; Chen, W.; He, E.; Jia, X.; Bao, H.; Zhou, X.; Ghosh, R.; and Ravirathinam, P. 2023. Harnessing heterogeneity in space with statistically guided meta-learning. *Knowledge and information systems*, 65(6): 2699–2729.

Xie, Y.; He, E.; Jia, X.; Bao, H.; Zhou, X.; Ghosh, R.; and Ravirathinam, P. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*, 767–776. IEEE.

Xie, Y.; He, E.; Jia, X.; Chen, W.; Skakun, S.; Bao, H.; Jiang, Z.; Ghosh, R.; and Ravirathinam, P. 2022. Fairness by “Where”: A Statistically-Robust and Model-Agnostic Bilevel Learning Framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12208–12216.

Yan, A.; and Howe, B. 2019. Fairst: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 552–555.

Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. *Advances in neural information processing systems*, 30.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhao, C.; Chen, F.; Wang, Z.; and Khan, L. 2020. A primal-dual subgradient approach for fair meta learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, 821–830. IEEE.

Zhao, C.; Mi, F.; Wu, X.; Jiang, K.; Khan, L.; and Chen, F. 2022a. Adaptive Fairness-Aware Online Meta-Learning for Changing Environments. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Zhao, T.; Dai, E.; Shu, K.; and Wang, S. 2022b. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 1433–1442.