UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FAIRNESS IN MACHINE LEARNING AND OPTIMIZATION: APPLICATIONS IN COMPUTATIONAL CRIMINOLOGY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements

for the Degree of

DOCTOR OF PHILOSOPHY

BY

KAREN ROBERTS-LICKLIDER

Norman, Oklahoma

2025

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/341692

FAIRNESS IN MACHINE LEARNING AND OPTIMIZATION:
APPLICATIONS IN COMPUTATIONAL CRIMINOLOGY

A DISSERTATION APPROVED FOR THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Theodore Trafalis, Chair

Dr. Talayeh Razzaghi

Dr. Charles Nicholson

Dr. Andrés Gonzalez

Dr. Trina Hope

© Copyright by KAREN ROBERTS-LICKLIDER 2025
All Rights Reserved.

Abstract

This research leverages machine learning, optimization, and systems modeling to address fairness in treatment and criminal justice.

This dissertation offers an overview of fairness in machine learning, discussing how bias can be introduced through data and model selection, and how fairness metrics such as demographic parity, disparate impact, equalized odds, and statistical parity difference can be employed to identify and mitigate inequality.

In addition, the dissertation generalizes these concepts to optimization, pre-processing methods such as reweighting methods, and social welfare metrics such as the Gini coefficient, the Hoover index, which can be embedded into optimization functions.

The facility location problem is solved to increase access to drug and alcohol rehabilitation for Oklahoma facilities. Through the lens of distance measurements such as Haversine, Euclidean, Manhattan, and Chebyshev, and constraint conditions such as fairness, the model seeks to find placement of facilities that can minimize travel cost and achieve better fairness for service delivery. Contrary to the common assumption that equity is expensive, adding fairness constraints reduced total travel cost while improving access. Simply, fairness pulls facilities slightly away from already well served corridors toward under served counties, which cuts many long trips and improves overall coverage. We also find that this equity shift pairs naturally with a grid like distance, reinforcing both lower cost and fairer access.

Next, compartmental system models are fitted to explore the consequences of treatment-oriented and punishment-oriented strategies to incarceration and recovery. Results suggest that treatment prioritization may lower incarceration rates and costs. Extending this more, an epidemic model where income governs

the addiction levels (SIRV2 model) is considered. The model demonstrates that poverty influences outcomes, influences relapse patterns and constrains treatment options, and resource distribution based on such considerations of fairness enhances treatment equity.

Lastly, predictive machine learning (ML) models are constructed with the Substance Abuse and Mental Health Services Administration (SAMHSA) rehabilitation's data set in predicting treatment completion. Various methods are evaluated — support vector machines (SVM), decision trees, random forest, and neural networks — as well as fairness interventions like over-sampling, over-sampling using intersectionality, and over-sampling to the worst-case ratio. The results indicate that decision trees and random forests strike the best trade-off between accuracy and fairness.

In summary, this work connects algorithmic fairness, optimization, and system modeling with policy that can be made in practice. It offers policy recommendations to improve rehabilitation, decrease incarceration, and foster more equitable systems of addiction recovery.

Acknowledgments

First, I would like to sincerely thank my advisor, Dr. Theodore Trafalis. His knowledge, patience, and consistent support have played a major role in shaping my Ph.D. experience. Under his guidance, I've learned to think more analytically, ask meaningful questions, and raise the standard of my work. I'm genuinely grateful for the mentorship and encouragement he has provided throughout this journey.

I'm also very grateful to my dissertation committee members—Dr. Andrés Gonzalez, Dr. Talayeh Razzaghi, Dr. Charles Nicholson, and Dr. Trina Hope. Their thoughtful feedback, wide-ranging expertise, and valuable perspectives have helped improve the quality and depth of this research. I appreciate the time and effort each of them dedicated to supporting this work.

Finishing a Ph.D. wouldn't have been possible without the support of a few very important people in my life. My wife, Shelly Roberts-Licklider, has been by my side through all of it—encouraging me, being patient, and believing in me even during the times when I didn't believe in myself. Her support has meant everything.

I'm incredibly thankful for my family. My parents who have supported me through all my ups and downs and never given up on me, my brother who has given me a reason to better myself to be a good example for him and to never give up, and both of my grandmothers have been a huge part of my journey. Although my grandmothers sadly passed away before they could see this moment, their love and support helped me reach this milestone.

I'm especially thankful for my dear friends Linda Neikirk, Lori Murphy, and Cindy Rose. They were there for me during the hardest times—when I felt like giving up, they reminded me to keep going. Their encouragement and friendship helped me stay focused and finish what I started. I also want to thank my friend John Bridges for being my mentor throughout my academic career and my friend Jason Papayik, who has been an invaluable sounding board, offering insight and helping me think through many of the analytical ideas that shaped this academic journey.

Contents

Acknowledgments vi

1 Introduction 1

2 Literature Review 6

3 Fairness in ML 9

3.1 Introduction . . . 9
3.2 Research Questions . . . 10
3.3 Fairness Measures . . . 11
3.4 Fairness Concepts . . . 11
3.4.1 Individual Fairness . . . 11
3.4.2 Group Fairness . . . 12
3.5 Discrimination Types . . . 12
3.5.1 Disparate Impact . . . 12
3.5.2 Disparate Treatment . . . 13
3.5.3 Disparate Mistreatment . . . 13
3.6 Levels of Discrimination . . . 14
3.6.1 Structural . . . 14
3.6.2 Organizational . . . 14
3.6.3 Interpersonal . . . 15
3.7 Bias Types . . . 15
3.7.1 Confirmation Bias . . . 15
3.7.2 Population Bias . . . 16

3.7.3 Sampling/Selection Bias . . . 16
3.7.4 Interaction Bias . . . 16
3.7.5 Decline Bias . . . 16
3.7.6 Hindsight Bias . . . 17
3.7.7 In-Group Bias . . . 17
3.7.8 Emergent Bias . . . 17
3.7.9 Label Bias . . . 17
3.7.10 Outcome Proxy Bias . . . 18
3.7.11 Statistical and Psychological Phenomenon in Bias . . . 18
3.7.11.1 Dunning-Kruger Effect . . . 18
3.7.11.2 Simpson's Paradox . . . 18
3.8 Fairness Measures . . . 18
3.8.1 Disparate Impact . . . 19
3.8.2 Demographic Parity . . . 19
3.8.3 Statistical Parity Difference . . . 20
3.8.4 Equal Opportunity . . . 20
3.8.5 Equalized Odds . . . 21
3.9 Intersectional Fairness . . . 22
3.10 Worst-case Ratios . . . 25
3.10.1 Demographic Parity Ratio . . . 26
3.10.2 Disparate Impact Ratio . . . 26
3.10.3 Conditional Statistical Parity Ratio . . . 26
3.10.4 Multiclass Equal Opportunity Ratio . . . 27
3.10.5 Multiclass Equalized Odds Ratio . . . 27
3.10.6 Fairness Through Awareness . . . 29
3.10.7 Fairness Through Unawareness . . . 29
3.11 Conclusion . . . 30

4 Fairness in Optimization **31**
4.1 Introduction . . . 31
4.2 Research Questions . . . 31
4.3 Key Terms for Bias Mitigation Algorithms . . . 32

4.4 Pre-Processing Algorithms . . . 33
4.4.1 Suppression . . . 33
4.4.1.1 Covariance Analysis . . . 33
4.4.2 Reweighting Process . . . 34
4.4.3 Resampling . . . 35
4.4.4 Massaging . . . 36
4.4.5 Optimized Pre-Processing . . . 36
4.4.6 Learning Fair Representation . . . 36
4.4.7 Disparate Impact Remover . . . 37
4.5 In-Processing Algorithms . . . 37
4.5.1 Loss Function Optimization in Classification Problems 37
4.5.1.1 Logistic Regression . . . 38
4.5.1.2 Linear SVM . . . 38
4.5.1.3 Non-Linear SVM . . . 39
4.6 Post-Processing Algorithms . . . 40
4.6.1 Fair Classification . . . 40
4.6.2 Fair Regression . . . 42
4.6.3 Fair Clustering . . . 42
4.6.4 Fair Embedding . . . 43
4.6.4.1 Debiasing Algorithm . . . 43
4.6.5 Equalized Odds . . . 46
4.7 Equity in Optimization . . . 46
4.7.1 Gini Coefficient . . . 47
4.7.2 Hoover Index . . . 48
4.7.3 McLoone Index . . . 48
4.7.4 Max-Min Fairness . . . 48
4.7.5 Price of Fairness . . . 49
4.8 AI and Ethics . . . 49
4.8.1 Use in Law Enforcement . . . 49
4.8.2 Data Protection and User Privacy . . . 49
4.9 Conclusion . . . 50

5 Optimizing Treatment Facility Locations in Oklahoma Using Haversine, Euclidean, Manhattan, and Chebyshev Distance Optimization **51**

5.1 Introduction . . . 51
5.2 Problem Statement . . . 53
5.3 Research Questions . . . 53
5.4 Our Contributions . . . 54
5.5 Methodology and Assumptions . . . 55
5.6 Background . . . 57
5.7 Analysis . . . 59
5.7.1 The Network . . . 59
5.7.2 The Euclidean Model . . . 59
5.7.3 The Haversine Parameter . . . 62
5.7.4 The Manhattan Parameter . . . 63
5.7.5 The Chebyshev Parameter . . . 63
5.8 Fairness Measures . . . 64
5.9 Results . . . 67
5.10 Conclusions . . . 72

6 A Treatment Focused Compartmental Model Approach to Drug Offense Mitigation in Oklahoma **74**

6.1 Introduction . . . 74
6.2 Research Questions . . . 75
6.3 Assumptions . . . 76
6.4 Background . . . 76
6.4.1 Current State . . . 76
6.5 Methodology . . . 81
6.5.1 Future State . . . 84
6.6 Analysis . . . 85
6.7 Results . . . 88
6.8 Sensitivity Analysis via Python Translation . . . 93
6.9 Validation . . . 95

6.10 Conclusions 96

7 Poverty-Structured $SIRV_2$ Model with Fairness Modeling Methamphetamine Addiction in Oklahoma 98

7.1 Introduction 98
7.2 Research Questions 99
7.3 Methods 99
7.3.1 The Model 99
7.3.2 Parameter Definitions 102
7.3.3 Assumptions 103
7.3.4 The Data 104
7.3.5 Results 109
7.3.6 The Statistics 116
7.4 Conclusions 131

8 Machine Learning Techniques with Fairness for Prediction of Completion of Drug and Alcohol Rehabilitation 132

8.1 Introduction 132
8.2 Research Questions 133
8.3 Methodology 134
8.4 Data Clean Up Process 134
8.5 Encoding and Balancing Data 135
8.6 Data Sensitivities/Distributions 137
8.6.1 Pareto Charts 137
8.6.2 Bucketized Categories 139
8.7 Fairness Measures 139
8.7.1 Disparate Impact 139
8.7.2 Disparate Impact – Multiclass 143
8.7.3 Statistical Parity Difference 144
8.7.4 Statistical Parity Difference – Multiclass 145
8.7.5 Equal Opportunity 147
8.7.6 Equalized Odds 149
8.7.7 Equalized Odds - Multiclass 150

8.8 The Model . . . 153
8.8.1 Support Vector Machines . . . 153
8.8.2 Decision Trees . . . 155
8.8.3 Random Forests . . . 157
8.8.4 Neural Networks . . . 160
8.9 Interpretation of Results . . . 162
8.10 Reweighting . . . 166
8.10.1 Interactions . . . 166
8.10.1.1 Dual Interaction . . . 166
8.10.1.2 Three Way Interaction . . . 167
8.10.2 Intersectionalization . . . 167
8.11 New Fairness Calculations . . . 167
8.11.1 Worst Case Demographic Parity . . . 167
8.11.2 Worst Case Demographic Parity – Multiclass . . . 169
8.11.3 Worst Case Disparate Impact . . . 169
8.11.4 Conditional Statistical Parity Ratio . . . 171
8.11.5 Equal Opportunity Ratio - Multiclass . . . 172
8.11.6 Equal Odds Ratio . . . 174
8.12 Conclusions . . . 180

9 Conclusion **181**

10 Future Work **183**

Bibliography **185**

Chapter 1

Introduction

This dissertation shows that by integrating algorithmic fairness, optimization, ML and system modeling provides a realistic optimal site location, and a county-level system model that supports a first treatment approach that can reduce incarceration, lower public costs, and improve access in Oklahoma rehabilitation centers. The conclusions hold across reasonable distance assumptions and parameter ranges. We acknowledge older data in places, but we use them to demonstrate a clear, transparent method that policymakers can apply as newer data become available.

According to (CSG Justice Center, 2025), in 2023, there are more drug and alcohol related deaths per 100,000 people in Oklahoma than homicides or suicides, with 30 drug overdoses and 20 alcohol-induced deaths. More than 600 of these deaths between 2015 and 2023 were due to methamphetamine. In terms of poverty, 23% of Oklahoma renters paid a rent that was more than their income. In 2020, approximately 1,200 people were incarcerated in prison in Oklahoma per 100,000 people, about 85% of the prison population being drug offenses. According to data for 2012, of the people released from prison in 2012, 71% were arrested within 5 years of release, 54% were convicted of a new offense and 46% returned to prison. 22% of the people who left prison in Oklahoma in 2014 were re-incarcerated in 3 years. The disparities in the arrest rate have remained the same, where the arrest rate for black individuals is double the rate for white people. The prison incarceration rate remains nearly 5 times higher for the black population compared to the White Population. This shows one example of why we need to look at fairness. This dissertation relies in places on drug court records from 2002–2005. That limits how closely our estimates match current needs, especially for counties that do not have drug court programs and could shift demand compared to incarceration based measures. We chose these data because they are accessible and structured enough to demonstrate the workflow, and the models are built so we can plug in newer datasets later and refresh all results.

This dissertation will look at how drug addiction data can be examined from a computational criminology perspective using methods of optimization of treatment centers to ensure that optimal and fair coverage is applied throughout the state of Oklahoma. We look at how machine learning models can be applied to look at fairness in predicting whether an addict will complete a rehabilitation program, as well as how many times it may take them. The classic SIR model is expanded to explore compartmental models to show how drug addiction spreads as a disease. The first model looks at punishment vs treatment approaches while the second model looks at the addiction model with a poverty component and how the Gini index of a county can be applied as a fairness component. All of this ties together to show how the spread of addiction and access to rehabilitation in Oklahoma can be significantly improved if we apply these techniques.

For our contributions, We focus on four simple goals. First, place more rehab centers across Oklahoma in a more equitable spread so that people don't have to travel far to get help. Second, look at how many times people go to rehab before they finish and see if we can fairly predict who is likely to complete treatment. Third, test whether a "treatment-first" approach can reduce repeat offending compared to sending people directly into the justice system. Finally, study how poverty affects addiction over time and identify which counties are hit hardest, using a straightforward fairness score (the Gini index) to keep access more equal. The layout of how the dissertation reads is as follows:

In Chapter 2 we begin our literature review of what authors contributed to each section of the extensive literature reviews that are continued in Chapters 3 and 4.

In Chapter 3, we provided an extensive review of fairness measures and bias mitigation strategies in machine learning, focusing on foundational approaches to ensure fairness in algorithmic decision-making. That survey covered a range of techniques including pre-processing, in-processing, and post-processing methods, examining how these methods can address bias in machine learning models. The key fairness metrics explored in Part 1—such as demographic parity, disparate impact, and equal opportunity—laid the groundwork for understanding fairness in algorithmic systems. Part 1 also highlighted the challenges of measuring and implementing fairness, especially in settings where sensitive attributes such as race, gender, and age play a crucial role.

Building on the foundations of Part 1, Chapter 4, expands the scope by exploring more advanced and emerging areas of fairness in machine learning and optimization. This chapter extends the discussion to include Fair Classification, Fair Clustering, and Fair Regression, examining their role in reducing bias and ensuring fairness in diverse algorithmic systems. In addition, we explore the optimization of fairness measures and their practical impact on sensitive demographic data, with a particular focus on attributes such as

gender, race, and age.

The primary objectives of this second survey are to provide a comprehensive overview of fairness and bias mitigation algorithms in both machine learning and optimization, and to demonstrate how fairness measures, bias reduction techniques, and data privacy work together to promote ethical AI development. This chapter also emphasizes the practical implications of fairness in sensitive areas such as law enforcement and healthcare, where the need for unbiased and equitable decision making is paramount.

In this survey, we explore various fairness measures used in algorithmic decision making and optimization, focusing on reducing bias in protected classes. Sensitive demographic data, where bias often centers on attributes such as sex, age, religion, marital status, veteran status, race, pregnancy, disability, and sexual orientation, form the basis of our analysis. Our goal is to show how fairness measures, loss function optimization, and bias reduction techniques interact within machine learning processes to create more ethical and equitable AI systems.

In addition to providing an overview of key fairness measures, including demographic parity, disparate impact, equalized odds, and statistical parity difference, we explore how these measures can be optimized in multiclass scenarios. The survey also covers various types of bias that affect machine learning models, such as confirmation bias, sampling/selection bias, interaction bias, and in-group bias, to understand how these contribute to unfair outcomes. By integrating theoretical concepts with real-world applications, this survey contributes to the ongoing discussion on fairness in AI and optimization for ethical decision making.

Ultimately, this chapter aims to offer a detailed understanding of fairness measures and their implementation in machine learning models, with a focus on reducing bias in protected classes. Through this exploration, our goal is to demonstrate how fairness, bias reduction, and data privacy can collectively lead to more equitable and ethical AI systems. As machine learning algorithms increasingly impact sectors like finance and law enforcement, it is crucial to develop systems that provide fair and unbiased results across diverse demographic groups. Ensuring fairness requires evaluating and mitigating bias at multiple stages of the machine learning process: data pre-processing, algorithmic decision making, and output interpretation, while extending fairness metrics to ensure equitable outcomes across different intersectional groups. In Chapter 5, *Optimizing Treatment Facility Locations in Oklahoma Using Haversine, Euclidean, Manhattan, and Chebyshev Distance Optimization* , which we will refer to in this dissertation as the *Location Problem* chapter, Computational Criminology (CC) applies computational models and simulations to address criminological challenges (Berk, 2008). This study uses a Facility Location Problem (FLP) framework to optimize the placement of rehabilitation treatment facilities in Oklahoma (Cantlebary and Li, 2020).

A multiobjective optimization model is developed to minimize costs and travel distances while maximizing facility coverage across regions, adhering to constraints from the Maximum Covering Location Problem (MCLP) and the P-Center Problem (Rader, 2010). The P-Center model focuses on minimizing the maximum distance between demand sites and their nearest facilities, ensuring equitable access between regions.

Oklahoma's overcapacity prisons and lack of accessible treatment facilities contribute to high recidivism rates, with 68% of drug offenders rearrested within three years (Belenko et al., 2013b). By focusing on equitable access to treatment, this research aims to reduce re-offending and addiction-related crimes. Using fairness measures such as the Gini coefficient and Hoover index (Chen and Hooker, 2023), combined with distance metrics (e.g. Haversine (Akkihal, 2006), Manhattan (Batta et al., 1990), Euclidean (Maria et al., 2020) and Chebyshev (Mora et al., 2016)), facilities are optimally distributed. The Chebyshev metric, which considers the maximum travel distance in any direction, is included to address various geographical demands.

The Gurobi optimizer is used to prioritize objectives, solving for cost, distance, and coverage (Gurobi Optimization, LLC, 2023). Starting locations are defined using the center of each county, with squared Haversine distances used for minimization. This approach integrates fairness measures and P-Center optimization to strategically distribute facilities, demonstrating how improved treatment access can address addiction and reduce incarceration rates.

By combining advanced optimization techniques with fairness measures, this research highlights a practical approach to addressing systemic challenges in addiction treatment and incarceration. These findings contribute to a more equitable and efficient allocation of resources, ultimately promoting better social outcomes.

Substance abuse is a significant contributor to mental illness, with prevalence rates varying between age groups (National Survey on Drug Use and Health, 2022). In order to address these issues, tools such as COMPAS (Correctional Offender Management Profile for Alternative Sanctions) have been employed in criminal justice to predict recidivism, but have been criticized for racial bias (Dressel and Farid, 2018). This highlights the importance of incorporating fairness into assessments, especially when considering alternatives to incarceration, such as rehabilitation programs.

We also touch on the fact that methamphetamine abuse poses a significant challenge in the United States, with Oklahoma ranked ninth in the prevalence of meth labs and reported nearly 333 deaths from meth overdoses in 2016 (Cosgrove, 2017; Advanced Recovery Systems, 2018). Meth use is strongly correlated with poverty, with many users coming from high-poverty backgrounds. In Chapter 6, we look at how a punishment focused model compares to a treatment focused model when looking at the number of people

incarcerated as well as the amount of tax dollars spent over a 10 year period. This uses a compartmental model to see how addicts move through free, probation, prison, and treatment using simulation techniques. To explore these dynamics in chapter 7, we utilize methods like the White and Comiskey SIR model (Battista, 2009) by incorporating county-level poverty data as a second independent variable. The model categorizes individuals into susceptible, actively using and recovered groups, stratified by poverty levels.

Using this enhanced model, we aim to map poverty levels by county in Oklahoma and compare them with the locations of Alcoholics Anonymous (AA) and Narcotics Anonymous (NA) meetings. Hotspot analyzes will identify correlations between poverty levels, prevalence of addiction, and accessibility to recovery meetings, offering information on targeted interventions to address meth addiction in vulnerable populations.

In chapter 8, the study uses treatment episode data from the Substance Abuse and Mental Health Services Administration (Substance Abuse and Mental Health Services Administration, 2021) to predict rehabilitation outcomes, including program completion and treatment history. Machine learning models such as Support Vector Machines (SVM), Random Forests, Neural Networks, and Decision Trees are utilized to develop predictive frameworks. Techniques like SMOTEN for class balancing (Chawla et al., 2002) and fairness measures, including Demographic Parity, Equalized Odds, and Disparate Impact (Irfan et al., 2023), are incorporated to ensure equitable predictions. These models are evaluated using fairness metrics for both binary and multiclass scenarios, considering complex interactions and biases (Ghosh et al., 2021).

By integrating advanced algorithms with fairness-focused approaches, this work seeks to address Oklahoma's high incarceration rates by exploring rehabilitation as an alternative to prison. Improved predictive accuracy and fairness in decision making can help reduce recidivism and promote equitable outcomes for people facing substance abuse challenges.

Chapter 9 will conclude the dissertations work and Chapter 10 will have a comprehensive look at the future work from all chapters.

In terms of gaps in the literature, most studies focus on only one piece of the problem. They either predict who will finish treatment without checking if the results are fair, pick facility locations using a simple straight line distance that does not match how people actually drive in Oklahoma, or run system models that ignore equity and differences between counties. Very few connect treatment-versus-punishment choices to both incarceration and public cost while accounting for poverty, rural access, and uncertain parameters, especially with Oklahoma specific data.

Chapter 2

Literature Review

The literature on machine learning and fairness measures is vast, focusing on bias mitigation, fairness constraints, and their applications in different fields. Several key papers explore discrimination and bias in AI (Zafar et al., 2017), (Mukherjee, 2023b), (Barocas et al., 2023b). Additional research emphasizes fairness in ML systems and explores practical fairness measures such as the Gini Index, Hoover Index, and intersectional fairness (Ghosh et al., 2022), (Irfan et al., 2023). These studies are foundational for understanding the challenges and opportunities in applying fairness in computational criminology. The following tables show the concepts mentioned in each source.

Table 2.1: Discrimination and Bias in AI

Method	Disparate Impact	Disparate Treatment	Disparate Mistreatment	Levels of Discrimination	Confirmation Bias	Population Bias	Sampling/Selection Bias	Interaction Bias	Decline Bias	Hindsight Bias	In-Group Bias	Emergent Bias	Label Bias	Outcome Proxy Bias
(Zafar et al., 2017)	✓	✓	✓											
(Mukherjee, 2023b)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
(Barocas et al., 2023b)	✓	✓												
(Barocas et al., 2023c)				✓										
(Leslie et al., 2015)				✓										
(Mahoney et al., 2020a)							✓			✓	✓		✓	✓
(Peters, 2022)					✓									
(Rose, 2010)						✓								
(Berg, 2005)							✓							
(Baeza-Yates, 2018)								✓						
(Verma et al., 2021)							✓						✓	
(Roese and Vohs, 2012)										✓				
(Taylor and Doria, 1981)											✓			
(Ekstrand et al., 2021)	✓	✓	✓											

Table 2.2: Fairness Metrics

Metric	Demographic Parity	Disparate Impact	Stat. Parity Diff.	Equal Opportunity	Equalized Odds	Disparate Mistreatment	Demographic Parity Ratio	Disparate Impact Ratio	Cond. Stat. Parity Ratio	MC Equalized Odds Ratio	MC Equal Opportunity Ratio	Fairness Through Awareness	Fairness Through Unawareness	Intersectional Fairness
(Mukherjee, 2023b)	✓		✓	✓	✓	✓						✓	✓	
(Ghosh et al., 2022)		✓					✓	✓	✓	✓	✓			
(Nielsen, 2020)		✓	✓	✓	✓									
(Irfan et al., 2023)		✓	✓	✓	✓					✓	✓			

Table 2.3: Fairness Algorithms in AI

Algorithm	Pre-processing	Reweighting	Removing Bias Data	Resampling	Massaging	Covariance Analysis	In-processing	Logistic Regression Loss	Linear SVM Loss	Non-linear SVM Loss	Post-Processing	Fair Regression	Fair Clustering	Fair Embedding	Fair Classification	Equalized Odds
(Mukherjee, 2023b)	✓						✓				✓					
(Nielsen, 2020)		✓														
(Mukherjee, 2023a)								✓	✓	✓		✓	✓	✓	✓	

Table 2.4: Equity in Optimization

Algorithm	**Equity in Optimization**	**Hoover Index**	**Gini Index**	**McLoone Index**	**Max-Min Fairness**	**Price of Fairness**
(Ogryczak et al., 2014)	✓					✓
(Chen and Hooker, 2023)	✓	✓	✓	✓	✓	

Chapters 4 and 5 go into an extensive survey of all the work that has been done in fairness in ML and optimization. Some of the literature gaps reviewed would be as follows. Most prior work looks at only one piece at a time: prediction, where to put facilities, or system behavior, not all together. Fairness is usually checked narrowly (one or two metrics, little attention to worst-case groups). Distance is often treated as a straight line, which ignores Oklahoma's grid roads and corridor travel. There is also little Oklahoma-

specific county-level evidence, with limited testing of how results change under different assumptions.

This dissertation addresses these gaps by looking at how we can connect the pieces end to end: fair prediction, distance-aware siting, and a county level system model, so findings add up. We evaluate fairness broadly (including worst case groups), use distance measures that better reflect how people actually travel, and show that adding fairness can even lower overall cost by cutting very long trips. We ground everything in Oklahoma data, state the limits clearly, and test robustness so the policy message (for example, treatment first versus punishment first) is practical and reliable.

Chapter 3

Fairness in ML

3.1 Introduction

Reducing bias in machine learning problems has been of growing interest in recent years. Research done in German Credit data, Adult Census Income data (Valentim et al., 2022), Home Credit data (Ángel Pavón Pérez et al., 2023), ProbPublica COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) data, and facial recognition software Fitzpatrick Skin Type classification system (Mukherjee, 2023a), are some examples of data and machine learning software that has proven to show bias in their predictions. The German Credit data is an offline banking data set while the Home Credit data is referring to online banking. The findings between the two were slightly different. The Home Credit data is biased towards the unbanked individuals which tends to be females. The German Credit data has 7.5% more males with good credit than females, while the Home Credit data favours 3% more females than males (Ángel Pavón Pérez et al., 2023). The Adult Census dataset contains demographic data extracted from the 1994 Census Bureau database with the main task being to predict whether a person earns over $50,000 per year is classified into high or low income. Sex is the protected group in this dataset. Female is the bias variable in this data with only 15.04% having high income (Valentim et al., 2022). In the COMPAS dataset the task is to predict whether an offender is going to re-offend. In this dataset, black offenders had a higher overall score of recidivism (re-offending) with 45% more being assigned a higher score than whites. Female defendants were 19.4% more likely to be assigned a higher score than males even though females had a lower score of criminality (Mukherjee, 2023a). In the skin type classification system two benchmark datasets were used, IJB-A and Audience. Gender shades were then classified by gender and skin type in an intersectional classification. The IJB-A data set was released in 2015 by the National Institute of Standards and Technology (NIST). It contains 79.6%

light skin types, where darker females were least represented with only 4.4% of the population and darker males comprised of 16%. Note that lighter females represented 20.2% of the data, where lighter males were the majority with 59.4% of the dataset (Mukherjee, 2023a). The Audience dataset was released in 2014 and had 86.2% with lighter skin type. Darker males were 6.4%, and darker females 7.4% of the data. The lighter males were 41.6%, and lighter females were the majority with 44.6% of the population (Mukherjee, 2023a). There was also the case of Amazon same day delivery in which they tried removing the protected class from the data set all together but found that there was still enough correlation with terms such as zip code to cause bias. In this case, they tried to do Prime free same day delivery orders on orders over $35. This was delivered to 27 metropolitan areas. It was found that six major cities were excluding predominantly black zip codes (Chicago, Dallas, Washington, Boston, Atlanta, and New York), black citizens were half as likely as white residents to live in neighborhoods with access to Amazon same-day delivery (Mahoney et al., 2020a).

In this survey, we aim to look at various discrimination types that may arise in machine learning processes. These include disparate impact, where discrimination manifests as an outcome even when there is no discriminatory intent, and disparate treatment, which involves formal and informal discrimination (Zafar et al., 2017). Additionally, we will explore disparate mistreatment, a form of discrimination where individuals are subjected to unfair treatment when the misclassifications rates are different for the protected class as compared to the other classes (Zafar et al., 2017).

The primary objectives of the survey is to review various fairness measures used in algorithmic decision making, optimization of their loss function, and reduction of bias in protected classes. A focus on demographic data is important since the bias appears when it comes to sensitive attributes such as gender, age, religion, marital status, veterans status, race, pregnancy, disability, and sexual orientation. The fairness measures that will be reviewed will be demographic parity, disparate impact, equalized odds, equal opportunity, and statistical parity difference. We will also look at the multi-class cases which include the worst-case ratios. Our unique contribution is to explore how fairness measures and bias reduction are related. We want to show how these aspects work together in a way that is easy to understand.

3.2 Research Questions

- **What are the current fairness and bias mitigation algorithms used in machine learning and optimization?**
  - The research provides an overview of various algorithms that aim to ensure fairness in machine learning models and optimization

processes.

- **How can fairness measures be applied to sensitive demographic data such as gender, race, and age?**
  - The research examines how fairness measures can be used to mitigate biases in demographic data, focusing on key protected attributes.
- **What are the methods to ensure fairness in classification, clustering, and regression models?**
  - The research explores fairness-specific algorithms in Fair Classification, Fair Clustering, and Fair Regression models.
- **What are the challenges and trade-offs between fairness and predictive accuracy in machine learning models?**
  - The research discusses the balance between maintaining fairness and achieving high predictive accuracy in ML models.

3.3 Fairness Measures

There are several fairness measures to be considered such as demographic parity, disparate impact, statistical parity difference, equal opportunity, equalized odds, and disparate mistreatment. When looking at the multi-class case we look at worst-case ratios such as disparate impact ratio, demographic parity ratio, conditional statistical parity ratio, multi-class equal opportunity ratio, and multi-class equalized odds ratio. In addition the concept of fairness through awareness and fairness through unawareness will be explored.

3.4 Fairness Concepts

3.4.1 Individual Fairness

Individual Fairness states that individuals have similar experiences and should be treated similarly. It ensures that statistical measures of outcomes are equal for similar individuals (Mahoney et al., 2020a). According to (Ekstrand et al., 2021), individual fairness is typically operationalized with a distance metric Δ_{in} between individuals and another Δ_{dec} between decision distributions. It is fair if similar individuals receive similar decisions:

$$\Delta_{in}(v_i, v_j) \leq \delta \Longrightarrow \Delta_{dec}(\delta(v_i, \delta(v_j)) \leq \epsilon \tag{3.1}$$

In (Friedler et al., 2021), two key concepts are presented. The first, WYSIWYG ("What You See Is What You Get"), suggests that the data we have is a direct and unbiased representation of reality, indicating an absence of systematic bias or discrimination in the data gathering process. The second concept, WAE ("We are All Equal"), asserts that various groups are fundamentally similar regarding the task at hand, and any systematic differences in observations should be attributed to discrimination and, therefore, adjusted.

3.4.2 Group Fairness

Different groups have similar experiences in the system. These groups can be race, gender, ethnicity, veteran status, marital status, education, etc. These are denoted protected and unprotected groups, and the goal is to ensure that the unprotected group is not unfairly mistreated (Ekstrand et al., 2021). Group fairness partitions a population into groups defined by protected attributes and seeks to ensure that statistical measures of outcomes are equal across groups (Mahoney et al., 2020a). WYSIWYG and WAE apply to group fairness as well. When it comes to groups, WAE maintains that all groups have the same abilities, while WYSIWYG assumes that observations reflect the abilities of the groups (Mahoney et al., 2020a).

3.5 Discrimination Types

3.5.1 Disparate Impact

Disparate impact occurs when a decision making system provides outputs that benefit or hurt a group of people sharing a value of sensitive attributes more frequently than other groups of people (Zafar et al., 2017). Groups receive outcomes at different rates (Ekstrand et al., 2021). In the case of a job hire, the plaintiff can still win the case by providing evidence that an alternative practice is potentially more non-discriminatory (Mukherjee, 2023b). The employer then has the opportunity to explain whether the reason for the different selection rates has a business justification. The burden reverts back to the plaintiff to the alternative employment practice that would have achieved the employers aim while being less discriminatory. This is about practices that have a disproportionate effect on a protected class, even if unintentional, which must be both unjustified and avoidable (Barocas et al., 2023b).

3.5.2 Disparate Treatment

Disparate treatment occurs when a decision making system provides different outputs for groups of people with the same or similar values of non-sensitive attributes of features but different values of sensitive attributes (Zafar et al., 2017). Groups are intentionally treated differently (Ekstrand et al., 2021). There are two types of disparate treatment, formal and intentional. Formal discrimination is the direct denial of opportunities based on membership of a protected class and the concept of rational racism which refers to whatever is rational must be rightful. One example of rational racism would be that all old and disabled people should be left to die in a disaster due to the cost to rescue them being greater than the benefit they would bring to society if they were rescued (Mukherjee, 2023b). Intentional discrimination is best understood through burden-shifting under the Title VII employment law in which the victim has the responsibility to establish a case for discrimination against them. This goes back to the job hiring case in which the plaintiff is able to establish the case of discrimination thus shifting the burden to the defendant to show an account of non-discrimination (Mukherjee, 2023b). The plaintiff must establish "prima facie" (based on first impression) case of discrimination by showing they are a member of a protected class, were qualified for a position, were denied that position, and that the position remained open or that it was given to someone that was not in the protected class. If the plaintiff is successful, the employer must produce a legitimate, nondiscriminatory reason for the adverse decision (Barocas et al., 2023b). This type of discrimination matches the average person's conception of what discriminatory behavior looks like.

3.5.3 Disparate Mistreatment

Disparate mistreatment can occur in any automated decision making system whose outputs or decisions are not 100% accurate, i.e. the misclassification rates may be different for groups of people having different values of sensitive attributes (e.g. males and females; blacks and whites) (Zafar et al., 2017). Groups receive erroneous effects at different rates (Ekstrand et al., 2021). In some cases the false positive rate is more important while in other cases it is the false negative rate. An example of this is the case of the COMPAS data in which they assess the risk of recidivism; the false positive rate across the class of different races where it is better to let a guilty person go than to accuse an innocent one. In the case of the German credit data, it is more important to look at the false negative rates so that deserving candidates from the protected class are not discriminated against (Mukherjee, 2023b).

3.6 Levels of Discrimination

There are three levels of discrimination; structural, organizational, and interpersonal which can be looked at as direct or indirect. Direct discrimination occurs when references are made directly to the protected group. Indirect discrimination is when the actions or decision make no such references yet disadvantages one or more groups (Barocas et al., 2023c). An example of these direct and indirect relationships for each of the levels can be seen in the following table.

Table 3.1: Examples of Direct and Indirect Discrimination

Type	Direct	Indirect
Structural	Segregation Laws	English Language Proficiency Requirement
Organizational	Unequal Pay Policies	Seniority Based Advancement
Interpersonal	Gender Based Harassment	Stereotyping

3.6.1 Structural

Structural factors are the way society is organized. Laws that limit opportunities for certain groups is a direct form of structural discrimination (Barocas et al., 2023c). Examples of direct discrimination would be laws against a group, such as segregation laws against African Americans or anti marriage laws against same-sex couples. Indirect discrimination can occur in situations such as English proficiency requirements discriminating against non English speaking citizens or schools in higher income neighborhoods tending to be better funded and attracting more qualified teachers leading to education advantages to children of these higher income parents. An example of machine learning in structural discrimination is when the objective function in predictive models reflects the incentive of the organization deploying them such as in the 2019 study finding racial bias in a system used to identify patients with a high risk of adverse health outcomes with black patients being assigned lower scores compared to white who were equally at risk (Obermeyer et al., 2019). This discovery emerged due to the model being configured to anticipate healthcare costs rather than needs. It revealed that the healthcare system allocates fewer resources to care for black patients compared to their white counterparts, even when they share similar health conditions.

3.6.2 Organizational

Organizational discrimination are at the level of organizations and are structured by the decision-making rules and processes they put in place and

the context in which individual actors operate (Barocas et al., 2023c). Direct examples are unequal pay policies or lack of disability accommodations while indirect are seniority based advancement which discriminate against younger employees. One of the mitigation measures is adding a fairness constraint to the optimization or other types of algorithmic fairness interventions by use of redistribution or reallocation (Barocas et al., 2023c).

3.6.3 Interpersonal

Interpersonal discrimination are the attitudes and beliefs that result in discriminatory behaviour by individuals (Barocas et al., 2023c). Direct examples would be gender based harassment while indirect example would be stereotyping. A 2015 study suggests that academic fields in which achievement is believed to be driven by innate brilliance exhibits a greater gender disparity having fewer women (Leslie et al., 2015). The innate brilliance in women is a type of gender stereotyping in which women would self-select out of disciplines (or perform more poorly than they otherwise would) (Barocas et al., 2023c). One of the ways in which stereotyping has been attempted to be reduced has been through diversity training. Elizabeth Paluck and Donald Green conducted a review of thousands of intervention in 2009 (Paluck et al., 2020). The study involved implementing various strategies, including facilitating interactions with individuals from diverse groups, redefining social identity categories, raising awareness, addressing emotions, aligning with consistent values and self-worth, engaging in cooperative learning, and leveraging peer influence. However, limited evidence exists to substantiate the efficacy of this mitigation approach. Another method is to withhold the decision subject's identity from the decision maker. This is easier with technology than with in person hiring interviews (Chohlas-Wood et al., 2021).

3.7 Bias Types

3.7.1 Confirmation Bias

Confirmation bias is the tendency to only believe in or search for information that supports one's preconceived beliefs in which the individual avoids believing any contrary views or opinions (Mukherjee, 2023b). An example of this would be during elections when one supports politicians without thinking anything bad about the candidate and only accepts positive opinions. It is people's tendency to search for information that supports their beliefs and ignore or distort data contradicting them (Peters, 2022).

3.7.2 Population Bias

Population bias occurs when certain characteristics in the user population differ from the original target population such as demographics or representative groups (Mukherjee, 2023b). An example given is in 2010 facial detection software in cameras having a problem detecting non-Caucasian faces when people took photographs of people of Asian demographics. The Nikon Coolpix S630 digital camera would show an error message asking if someone blinked (Rose, 2010)

3.7.3 Sampling/Selection Bias

Sampling bias occurs when the target population is sampled in a non random way. (Mukherjee, 2023b). This occurs when one population is over-represented of under-represented in a training dataset. An example would be a job recruiting tool that has been trained on predominantly white male job candidates (Mahoney et al., 2020a). An example of this is a non-response bias when referring to the mechanism by which some survey respondents choose not to answer survey questions (thereby selecting themselves out of the sample) (Berg, 2005). Selection bias occurs when taking samples from demographic groups inadvertently shows correlations between the features pertaining to that demographic group and the training labels (Verma et al., 2021).

3.7.4 Interaction Bias

Interaction bias arises by how a user interacts with a system (Mukherjee, 2023b). An illustration of this phenomenon is evident in the context of reading habits, such as the left-to-right orientation in America. In online search results, items positioned towards the top left corner tend to receive disproportionately more attention. Another manifestation is observed in recommendation bias, where users might exclusively click on articles suggested to them, neglecting other articles that are not part of the recommendations (Baeza-Yates, 2018).

3.7.5 Decline Bias

Decline bias occurs when there is a consistent inclination to perceive the present as progressively worsening during comparisons with the past and present which is shown by a tendency attributed to an individual's challenges in adapting to ongoing changes (Mukherjee, 2023b).

3.7.6 Hindsight Bias

Hindsight bias closely aligns with lottery predictions, as individuals frequently tend to create logical explanations for past outcomes. They may develop a belief that they can accurately forecast future events, despite their initial information often leading to inaccurate predictions (Mukherjee, 2023b). Sometimes people feel like they "knew it all along" by believing that an event is more predictable after it becomes known than it was before it became known (Roese and Vohs, 2012). According to Roese and Vohs, hindsight bias stems from cognitive inputs, meta-cognitive inputs, and motivational inputs. Cognitive inputs involve the selective recall of information in alignment with one's current understanding, with individuals engaging in sense-making to attribute meaning to their knowledge. Meta-cognitive inputs refer to the potential misattribute of the ease with which a past outcome is comprehended to its presumed prior likelihood. Motivational inputs occur when individuals possess a need to perceive the world as organized and predictable, aiming to avoid being held responsible for problems.

3.7.7 In-Group Bias

In-Group bias occurs when groups sacrificed self-interest in order to protect group status, even in the face of failure (Taylor and Doria, 1981). An example of this would be when an employer may attempt to recruit people from their own social group without objectively evaluating their competence for this role (Mukherjee, 2023b).

3.7.8 Emergent Bias

Emergent bias materializes when shifts occur in the population, cultural disparities emerge, or new societal insights become known, typically after the system has been operational for a period. This phenomenon is commonly noted in Human-Computer Interaction (HCI) designs (Mukherjee, 2023b).

3.7.9 Label Bias

Label bias occurs when the annotation process introduces bias during the creation of training data such that the people labeling the data might not represent a diverse group of demographics and can bring their implicit personal bias into their labels (Mahoney et al., 2020a). This happens when the training labels which are mostly generated manually are afflicted by human bias (Verma et al., 2021). An example of this could be facial images from

young males and females from one ethnic group. This makes it label biased to not be able to predict older males and females of other ethnic groups.

3.7.10 Outcome Proxy Bias

Outcome proxy bias arises when the machine learning task is inadequately defined. For instance, predicting a person's likelihood to commit a crime using arrests as a proxy can introduce bias. This is problematic as arrest rates tend to be higher in areas with increased police presence, and an arrest itself doesn't necessarily indicate guilt (Mahoney et al., 2020a).

3.7.11 Statistical and Psychological Phenomenon in Bias

3.7.11.1 Dunning-Kruger Effect

The Dunning-Kruger effect is when someone with limited knowledge or skills considers themselves an expert on the subject and overestimates their abilities to perform the task successfully (Mukherjee, 2023b).

3.7.11.2 Simpson's Paradox

Simpson's paradox occurs when trends at the subgroup level of the dataset differ from those that occur at the aggregate level (Mukherjee, 2023b). An example of this is given in (Bickel et al., 1975) where graduate admissions was reported based on gender. It was shown that women were admitted at a lower rate than men for graduate admissions for graduate studies. However it showed that women candidates typically applied to more departments where the admission rates were lower for both men and women.

3.8 Fairness Measures

There are several fairness measures to be considered such as demographic parity, disparate impact, statistical parity difference, equal opportunity, equalized odds, and disparate mistreatment. When looking at the multiclass case we look at worst-case ratios such as disparate impact ratio, demographic parity ratio, conditional statistical parity ratio, multiclass equal opportunity ratio, and multiclass equalized odds ratio. Also the concept of fairness through awareness and fairness through unawareness will be explored. discussed in detail in the following table.

Fairness Measure	Criteria for Fairness
Demographic Parity	Probability outcome same for both groups $p_+(z=0)$ and $p_+(z=1)$
Disparate Impact	Closer to 1
Statistical Parity Difference	Closer to 0
Equal Opportunity	TPR same for both groups
Equalized Odds	TPR and FPR same for both groups and demographics

Table 3.2: Fairness Measures and Criteria for Fairness

3.8.1 Disparate Impact

$$DI = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \tag{3.2}$$

This compares the proportion of individuals receiving a favorable outcome for a privileged and underprivileged group. This is the ratio of the rate of a positive outcome for the disfavored group to the rate of a positive outcome for the favored group (Nielsen, 2020). The closer to 1 it is, the more fair it is. $\hat{Y}$ is the model predictions and A is the protected attribute, with 0 being the underprivileged class and 1 being the privileged class (Irfan et al., 2023). Typically the 80% rule is followed for the threshold to be considered discriminatory (Ghosh et al., 2022). To comply with disparate impact, we must comply with the p%-rule. This rule states that the ratio of the percentage of the individuals from the protected class with $d_\theta(x) \geq 0$ to that of the other class should be no less than $p : 100$ (Mukherjee, 2023a). We assume there is only one binary sensitive attribute $z = 0, 1$ giving us the following formula:

$$\min\left(\frac{P(d_\theta(x) \geq 0 \mid z = 1)}{P(d_\theta(x) \geq 0 \mid z = 0)}, \frac{P(d_\theta(x) \geq 0 \mid z = 0)}{P(d_\theta(x) \geq 0 \mid z = 1)}\right) \geq \frac{p}{100} \tag{3.3}$$

3.8.2 Demographic Parity

For demographic parity, membership in a protected class should have no correlation with the decision; meaning it should be independent of the protected attribute (Hardt et al., 2016). In the binary case of $\hat{Y} \in \{0, 1\}$ and a binary protector $A \in \{0, 1\}$ we have:

$$Pr\left\{\hat{Y} = 1 \mid A = 0\right\} = Pr\left\{\hat{Y} = 1 \mid A = 1\right\} \tag{3.4}$$

Becker et al. explores maximizing the overall coverage while ensuring that the coverage of all groups remains identical; thus enforcing demographic parity (Becker et al., 2023). They state that demographic parity is defined as equality in probability of being selected conditional on group membership. Since we are striving for equality, a demographic parity closer to zero is the goal while a demographic parity closer to one is considered more unfair. When it comes to the confusion matrix, this means that the true and false positive rates need to be the same for both classes to have demographic parity (Mukherjee, 2023b).

3.8.3 Statistical Parity Difference

As with demographic parity, the true and false positive rates must be equal for the two classes to have statistical parity (Mukherjee, 2023b). The closer the result is to 0, the more fair it is. The formula for statistical parity difference can be seen as follows.

$$SPD = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \tag{3.5}$$

Where $\hat{Y}$ is the models predictions, $A = 0$ is the protected attribute for the unprivileged class and $A = 1$ is the protected attribute for the privileged class. (Irfan et al., 2023). An example of statistical parity difference is in the loan data in which we want each applicant to have an equivalent opportunity to obtain a good credit score regardless of their gender, thus having equal probability for male and female applicants to have good predicted credit score:

$$P(d = 1|G = m) - P(d = 1|G = f) \tag{3.6}$$

where G is gender (m is male, f is female) and d is the predicted credit score (1 is good, 0 is bad) (Verma and Rubin, 2018).

3.8.4 Equal Opportunity

When the true positive rate (TPR) is the same for both the privileged and unprivileged groups, it is considered fair (Irfan et al., 2023). Let $A = 1$ and $A = 0$ equal the privileged and unprivileged groups demographics respectively.

$$P\left(\hat{Y} = 1 \mid Y = 1, A = 0\right) = P\left(\hat{Y} = 1 \mid Y = 1, A = 1\right) \tag{3.7}$$

Considering the above equation for $y = 1$, the equation shows the TPR across privileged and unprivileged groups, while $y = 0$ represents the FPR across privileged and unprivileged groups (Irfan et al., 2023).

$$\text{TPR}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3.8}$$

$$\text{EqOpp}_{\text{diff}} = \max_i(\text{TPR}_i) - \min_i(\text{TPR}_i) \tag{3.9}$$

The two classes $z = 0$ and $z = 1$ should have the same true positive rates, i.e. a person belonging to the positive class should correctly predicted to be in the positive class irrespective of the value of z (Mukherjee, 2023b). This holds with the following formula:

$$P\left(\hat{y} = 1 | z = 0, y = 1\right) = P\left(\hat{y} = 1 | z = 1, y = 1\right) \tag{3.10}$$

A classifier satisfies the false negative error rate balance definition if both protected and unprotected groups have equal FNR, the probability of a subject in a positive class to have a negative predictive value. An example would be in the probability of an applicant with an actual good credit score to be incorrectly assigned a bad predicted credit score and should be the same for both male and female applicants (Verma and Rubin, 2018):

$$P\left(d = 0 | Y = 1, G = m\right) = P\left(d = 0 | Y = 1, G = f\right) \tag{3.11}$$

Mathematically a classifier with equal FNR will also have equal TPR

$$P\left(d = 1 | Y = 1, G = m\right) = P\left(d = 1 | Y = 1, G = f\right) \tag{3.12}$$

where G is gender (m is male, f is female) and d is the predicted credit score (1 is good, 0 is bad).

3.8.5 Equalized Odds

Let $A = 1$ and $A = 0$ be the privileged and unprivileged groups respectively. When $Y = 1$, the next equation shows the TPR and when $Y = 0$ it shows the FPR. This shows the ROC where TPR and FPR are equal for the privileged and unprivileged demographics.
(Irfan et al., 2023).

$$P\left(\hat{Y} = 1 | Y = y, A = 0\right) = P\left(\hat{Y} = 1 | Y = y, A = 1\right) \tag{3.13}$$

where $y \in \{0, 1\}$
The positive and negative rates should be the same, thus the following formula should hold (Mukherjee, 2023b):

$$P(\hat{y} = 1|z = 0, y \in 0, 1) = P(\hat{y} = 1|z = 1, y \in 0, 1) \tag{3.14}$$

In the example of the loan application, the probability of an applicant with an actual good credit score to be correctly assigned a good predicted credit score and the probability of an applicant with an actual bad credit score to be incorrectly assigned a good predicted credit score should both be the same for male and female applicants, i.e., these should have similar classifications regardless of their genders. (Verma and Rubin, 2018):

$$P(d = 1|Y = i, G = m) = P(d = 1|Y = i, G = f), i \in 0, 1 \tag{3.15}$$

3.9 Intersectional Fairness

Intersectional fairness is defined as the maximum disparity between subgroups that combine membership from different protected attributes (Chen et al., 2023). There are three common limitations to group fairness algorithms commonly used to optimize machine learning models: single-axis definitions of fairness, dichotomization of privilege, and group size dependence (Lett and La Cava, 2023). The idea of restricting fairness considerations to just one protected attribute is flawed, as it oversimplifies the issue by only addressing a single aspect of potential discrimination. Oversimplification of the severity of discrimination within a protected attribute can vary between classes; thus differences in the extent of discrimination experienced within a particular protected attribute among different groups call for an intersectional approach to fairness. Such an approach should be capable of recognizing and addressing the diverse and complex nature of discrimination that occurs along various protected attributes. Three common remedies to group size dependence are including a population frequency weight in the fairness measure, imposing a threshold that excludes small groups from the fairness measure/algorithm, specifying a Bayesian prior that smooths fairness estimates for small groups (Lett and La Cava, 2023).

The idea of fairness gerrymandering is considered when looking at intersectionality. This is when we have unfair classification due to subgroups not having positive outcomes. An example would be not having black women and white men that passed but having all black men and white women passing. The bias is observed only while looking at the subgroups when we take race and gender as protected attributes (Ghosh et al., 2022).



Figure 3.1: Fairness Gerrymandering

An example of intersectional fairness without demographics is given by (Gohar and Cheng, 2023) using distributionally robust optimization (DRO) to minimize the worse case loss for any subgroup. We minimize the loss over all distributions that are close to the input distribution by considering the worst-case loss over all distributions with χ^2 divergence less that r, where r is the radius of a chi-squared ball ($B(P, r)$) around the input probability distribution P. The DRO function is defined as follows:

$$L_{DRO}(f, r) = \sup_{Q \in B(P,r)} \mathbb{E}_Q \left[l(f; \chi) \right] \tag{3.16}$$

In (Hashimoto et al., 2018), this is done by maximizing the minimum utility for all subgroups by using Rawlsian's Max-Min theory by reducing the worst-case risk where $\mathbb{E}\left[l(f; \chi) \right]$ is the expected loss for a loss function e.g. log loss (Gohar and Cheng, 2023). This fairness objective can be shown as:

$$L_{max}(f) = max_{s \in S} \mathbb{E}\left[l(f; \chi) \right] \tag{3.17}$$

Some of the fairness metrics mentioned in (Chen et al., 2023) are Statistical Parity Difference (SPD), Average Odds Difference (AOD), and Equal Opportunity Difference (EOD). The fairness metric and their intersectional metric can be seen in the following tables:

Statistical Parity Difference and Equal Opportunity Difference have been previously discussed. Average Odds Difference is the average discrepancy in FPR and TPR between the privileged and unprivileged groups (Chen et al., 2023).

Table 3.3: Fairness metrics

Metric	Definition
SPD	$P[\hat{Y}=1\|A=0]-P[\hat{Y}=1\|A=1]$
AOD	$\frac{1}{2}\Big(P[\hat{Y}=1\|A=0,Y=0]-P[\hat{Y}=1\|A=1,Y=0]$ $+P[\hat{Y}=1\|A=0,Y=1]-P[\hat{Y}=1\|A=1,Y=1]\Big)$
EOD	$P[\hat{Y}=1\|A=0,Y=1]-P[\hat{Y}=1\|A=1,Y=1]$

Table 3.4: Intersectional Fairness Metrics

Metric	Definition
SPD	$\max_{s\in S} P[\hat{Y}=1\|A=s]$ $-\min_{s\in S} P[\hat{Y}=1\|A=s]$
AOD	$\frac{1}{2}\Big(\max_{s\in S}(P[\hat{Y}=1\|A=s,Y=0]$ $+P[\hat{Y}=1\|A=s,Y=1])-\min_{s\in S}(P[\hat{Y}=1\|A=s,Y=0]$ $+P[\hat{Y}=1\|A=s,Y=1])\Big)$
EOD	$\max_{s\in S} P[\hat{Y}=1\|A=s,Y=1]$ $-\min_{s\in S} P[\hat{Y}=1\|A=s,Y=1]$

(Foulds and Pan, 2018) explains Differential Fairness as the extension of the 80% rule to protect multi-dimensional intersectional categories, with respect to output values by measuring the fairness cost of mechanism $M(x)$ with a parameter ϵ. A mechanism $M(x)$ is $\epsilon-$ deferentially fair (DF) with respect to $A, (\Theta)$ if for all $\theta \in \Theta$ with $x \in \theta$ and $y \in Range(M)$:

$$e^{-\varepsilon} \leq \frac{P_{M,\theta}(M(x)=y|s_i,\theta)}{P_{M,\theta}(M(x)=y|s_j,\theta)} \leq e^{\varepsilon}, \tag{3.18}$$

$$\text{for all } (s_i, s_j) \in A \times A \text{ where } P(s_i|\theta) > 0, P(s_j|\theta) > 0.$$

In this equation, the $s_i, s_j \in A$ are tuples of all the protected attributes values (gender, race, veteran status, marital status, etc) and Θ is a set of distributions θ which could plausibly generate each instance x. The example given is θ is a Gaussian distribution of credit scores per value of protected attributes, with mean and standard deviation in a certain range (Foulds and Pan, 2018).

Differential Fairness imposes a limitation on the comparative performance

across every pair of groups. The significance of this constraint is gauged in relation to the performance disparity between the highest and lowest performing groups (Maheshwari et al., 2023). This is evaluated using the following performance metric:

$$DF(h_\theta) \equiv \max_{g,g' \in \mathcal{G}} \log \left(\frac{m(h_\theta, \mathcal{T}_g)}{m(h_\theta, \mathcal{T}_{g'})} \right) \leq \epsilon \tag{3.19}$$

where $m(h_\theta, \mathcal{T}_g) = TPR(h_{\theta,\mathcal{T}_g})$ is the TPR for equalized opportunity and FPR is defined as $m(h_\theta, \mathcal{T}_g) = 1 - FPR(h_{\theta,\mathcal{T}_g})$

A leveling down effect occurs when a situation arises where a model that improves the performance across all groups is deemed more unfair by DF. In the example given by (Maheshwari et al., 2023), let h_θ and $h_{\hat{\theta}}$ be two models. Let the worst and best off group-wise performance be 0.4 and 0.5 respectively for h_θ. For $h_{\hat{\theta}}$, let it be 0.75 and 0.95. According to DF, h_θ is more fair than $h_{\hat{\theta}}$ since the two groups are closer even though $h_{\hat{\theta}}$ has better performance over the two groups. Thus, h_θ is leveling down to $h_{\hat{\theta}}$, but is deemed more fair. To combat this leveling down issue, (Maheshwari et al., 2023) came up with the $\alpha-$intersectional Fairness which is a convex combination of two components (i) the relative performance $\triangle_{rel}$ and (ii) $\triangle_{abs}$ which captures the leveling down effect by accounting for absolute performance of the worst-off group.

$$I_\alpha(g, g', h_\theta, m) = \alpha\Delta_{\text{abs}} + (1-\alpha)\Delta_{\text{rel}}, \tag{3.20}$$

where $\alpha \in [0, 1]$ and

$$\Delta_{\text{abs}} = \max\left(1 - m(h_\theta, \mathcal{T}_g), 1 - m(h_\theta, \mathcal{T}_{g'})\right),$$

$$\Delta_{\text{rel}} = \frac{1 - \max\left(m(h_\theta, \mathcal{T}_g), m(h_\theta, \mathcal{T}_{g'})\right)}{1 - \min\left(m(h_\theta, \mathcal{T}_g), m(h_\theta, \mathcal{T}_{g'})\right)}.$$

We then take the maximum value of I_α over all pairs of groups to get the $\alpha-$Intersectional Fairness.

$$IF_\alpha(h_\theta, m) \equiv \max_{g,g' \in G} I_\alpha(g, g', h_\theta, m) \leq \gamma \tag{3.21}$$

3.10 Worst-case Ratios

Worst case ratios are used when we are looking at the intersection of multiple protected attributes such as gender and race or other groups at the same time (Ghosh et al., 2022). The worse case ratios we will explore are demographic parity ratio, disparate impact ratio, conditional statistical parity ratio, equal opportunity ratio multiclass, and equal odds ratio multiclass.

3.10.1 Demographic Parity Ratio

Demographic parity compares the pass rate of (positive outcome) of two groups where sg_i is the subgroup i. It is satisfied if:

$$P\left(\hat{Y}\middle|A \in \text{sg}_i\right) = P\left(\hat{Y}\middle|A \in \text{sg}_i\right) \forall i, j \in \mathbb{N},\ i \neq j \tag{3.22}$$

Where N is the total number of subgroups. When using the worst-case ratio formula, it is also known as demographic parity ratio (Ghosh et al., 2022):

$$\text{DPR} = \frac{\min\left\{P\left(\hat{Y}\middle|A \in \text{sg}_i\right) \forall i \in \mathbb{N}\right\}}{\max\left\{P\left(\hat{Y}\middle|A \in \text{sg}_i\right) \forall i \in \mathbb{N}\right\}} \tag{3.23}$$

3.10.2 Disparate Impact Ratio

Disparate impact compares the pass rate of one group versus another with the four-fifths rule stating the ratio of the pass rate of group 1 to the pass rate of group 2 has to be greater than 80%. For the worst case definition, all intersectional disparate impact is defined as the minimum disparate impact between all possible pairs of subgroups sg (Ghosh et al., 2022).

$$\text{DI} = \min\left\{\frac{P(\hat{Y}|A \in \text{sg}_i)}{P(\hat{Y}|A \in \text{sg}_i)}; \forall i, j \in N, i \neq j\right\} \tag{3.24}$$

3.10.3 Conditional Statistical Parity Ratio

This uses the worst-case min-max ratio, with predictor $\hat{Y}$, member A, and legitimate attribute L (Corbett-Davies et al., 2017):

$$P\left(\hat{Y} \mid L = 1, A \in \text{sg}_i\right) = P\left(\hat{Y} \mid L = 1, A \in \text{sg}_j\right) \quad \forall i, j \in N, i \neq j \tag{3.25}$$

Using the worse case min-max ratio definition for conditional statistical parity ratio would be defined as (Ghosh et al., 2022):

$$CSPR = \frac{\min\left\{P(\hat{Y} \mid L = 1, A \in \text{sg}_i)\ \forall i \in N\right\}}{\max\left\{P(\hat{Y} \mid L = 1, A \in \text{sg}_i)\ \forall i \in N\right\}} \tag{3.26}$$

The multiclass case can be seen in the following equation where the minimum of all the ratios is taken.

$$CSPR_{\text{Multiclass}} = \min \left\{ \frac{\min \left\{ P(\hat{Y} \mid L = 1, A \in \text{sg}_i) \ \forall i \in N \right\}}{\max \left\{ P(\hat{Y} \mid L = 1, A \in \text{sg}_i) \ \forall i \in N \right\}} \right\} \tag{3.27}$$

3.10.4 Multiclass Equal Opportunity Ratio

The TPR is compared for the protected and unprotected group. The minimum of these in general is taken for the multiclass case (Ghosh et al., 2022).

$$\begin{aligned} &P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1) \\ &\quad = P(\hat{Y} = y_k | A \in \text{sg}_j, Y = 1) \quad \forall i, j \in N, \forall k \in K, i \neq j \end{aligned} \tag{3.28}$$

$$\text{EOppR} = \frac{\min \left\{ P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1) \forall i \in N \forall k \in K \right\}}{\max \left\{ P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1) \forall i \in N, \forall k \in K \right\}} \tag{3.29}$$

(3.30)

$$\text{EOppR}_{\text{Multi}} = \min \left\{ \frac{\min \left\{ P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1), \forall i \in N, \forall k \in K \right\}}{\max \left\{ P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1), \forall i \in N, \forall k \in K \right\}} \right\}$$

where K is the set of all possible output classes.

3.10.5 Multiclass Equalized Odds Ratio

K is the set of all possible output classes. The closer the value is to 1, the lower the disparity is. For this case we calculate the worst-case odds for each output class y and then take the minimum of all those values (Ghosh et al., 2022). This is the multiclass case for equalized odds ratio.

$$P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right) = P\left(\hat{Y} = y_k \middle| A \in \text{sg}_j\right) \quad \forall i, j \in N, i \neq j, \forall k \in K \tag{3.31}$$

$$\text{EOddR}_{\text{Multi}} = \min \left\{ \frac{\min \left\{ P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right) \right\}}{\max \left\{ P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right) \right\}} \right\} \forall i \in N, \forall k \in K \quad (3.32)$$

where K is the set of all possible output classes.

In the binary case, we check to see if the model's TPR and FPR are similar across different demographic groups. If the EOddsR is closer to 0, it indicates better fairness in terms of equalized odds. Since this is the binary case we only care about $\hat{y} = 1$. We calculate the min and max probabilities for each subgroup i. Calculate the ratio and then take the minimum to get the worse case.

$$\text{EOddR}_{\text{Binary}} = \min_{i \in N} \left\{ \frac{\min \left\{ P(\hat{Y} = 1 | A \in \text{sg}_i) \right\}}{\max \left\{ P(\hat{Y} = 1 | A \in \text{sg}_i) \right\}} \right\} \tag{3.33}$$

3.10.6 Fairness Through Awareness

Fairness through awareness means that similar individuals should have similar predicted outcomes, i.e. if two individuals perform equally on a predicted similar metric, the predictor should predict the same outcome for both of them (Mukherjee, 2023b). In the classification problem, if any two individuals are similar respect to a particular task, they should be classified similarly. We can do this by assuming a distance metric and a Lipschitz condition that requires that any two individuals x,y that are a distance $d(x, y) \in [0, 1]$ map to distribution $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$ (Dwork et al., 2011).

$$D(Mx, My) \leq d(xy) \tag{3.34}$$

An example of this could be a set of applicants V with a distance metric between applicants $k : V \times V \rightarrow \mathbb{R}$, a mapping for a set of applicants to probability distributions over outcomes $M : V \rightarrow \delta A$, and a distance D metric between distribution of outputs. Then fairness is achieved iff $D(M(x), M(y)) \leq k(x, y)$ (Verma and Rubin, 2018). In an example given by Verma and Rubin a possible distance metric k could be defined as the distance between two applicants i and j to be 0 if the attributes X (all attributes other than gender) are identical and 1 if some attributes in X are different. D could be defined as 0 if the classifier resulted in the same prediction and 1 otherwise.

3.10.7 Fairness Through Unawareness

The key idea in Fairness through unawareness is that no protected attribute should be used by the predictor to decide the outcome (Mukherjee, 2023b). A classifier satisfies this definition when no sensitive attributes are explicitly used in the decision-making process, i.e. in our example gender related features are not used for training the classifier so decisions cannot rely on

these features and classification outcome should be the same for applicants i and j who have the same attributes $X : X_i = X_j \rightarrow d_i = d_j$ (Verma and Rubin, 2018). One approach might be to ignore all protected attributes such as race, color, religion, gender, disability, marital status, veterans status, etc. However, this idea of fairness through unawareness is ineffective due to the existence of redundant encoding, ways of predicting protected attributes from other features(Hardt et al., 2016). This results in an essentially equivalent classifier, in the sense that it implements the same function because other features are correlated with the sensitive attribute (Barocas et al., 2023a). (Chen et al., 2023) investigate the concept of fairness under unawareness by employing probabilistic proxy models. These models are designed to estimate missing protected attribute values using other observed variables. A proxy model refers to any model that infers these missing values, and when this inference is specifically based on the prediction of conditional class membership probabilities, it is termed a probabilistic proxy model (Chen et al., 2018).

3.11 Conclusion

In conclusion, we have explored various bias and discrimination types and how they effect the machine learning algorithms. The different fairness measures have been reviewed in detail at both the binary and multiclass levels. Fairness through awareness and unawareness has been explored as well as intersectional fairness. Statistical and psychological phenomenon has been touched on to see how biases occur in that aspect. There is room for growth in these areas of research especially in the multiclass cases. Chapter 8 uses these techniques for fairness measures to predict how many times an addict goes to a rehabilitation program and completes it and how fairness measures such as gender, race, marital status among other variables play a role in prediction. We also use the method of intersection fairness to look at how multiple variables play a role together.

Chapter 4

Fairness in Optimization

4.1 Introduction

In this chapter, we we will explore pre-processing, in-processing, and post-processing techniques for debiasing and fair algorithms such as fair regression, fair clustering, fair SVM both linear and nonlinear, and fair classification, and fair embedding (Mukherjee, 2023a). We also will look at equity in optimization with use of social welfare functions (Chen and Hooker, 2023). Finally we will look at how AI and ethics play a role in law enforcement and privacy (Mukherjee, 2023c).

The primary objectives of the survey is to review various fairness measures used in algorithmic decision making, optimization of their loss function, and reduction of bias in protected classes. A focus on demographic data is important because those are the data where the bias lies when it comes to sensitive attributes such as gender, age, religion, marital status, veterans status, race, pregnancy, disability, and sexual orientation. Our unique contribution is to explore how fairness measures, loss function optimization, and bias reduction are related. We want to show how these aspects work together in a way that is easy to understand.

4.2 Research Questions

- **How can bias in machine learning algorithms be reduced across protected classes, and what are the key bias types to consider?**
  - The research looks into biases such as confirmation bias, sampling bias, and interaction bias, offering strategies to reduce bias across protected classes.

- **How do fairness, bias reduction, and data privacy contribute to the development of ethical AI?**
  - The research addresses how fairness, bias reduction, and data privacy work together to develop ethical AI, particularly in sensitive domains like healthcare and law enforcement.
- **How can social welfare functions, like the Gini coefficient, Hoover index, and McLoone index, be incorporated into optimization to promote fairness?**
  - The research explores how social welfare functions can be used in optimization problems to enhance fairness and reduce inequality in resource distribution.
- **What are the ethical concerns associated with AI, particularly in law enforcement and data privacy, and how can they be addressed?**
  - The research highlights ethical concerns in AI, especially in law enforcement and online data tracking, and how fairness measures and bias mitigation strategies can address these concerns.

4.3 Key Terms for Bias Mitigation Algorithms

Before diving into algorithms designed to enhance fairness, it's essential to familiarize ourselves with key terms. These terms include favorable label, protected attribute, privileged value (of a protected attribute), fairness metric, and discrimination/unwanted bias. A favorable label denotes a classification whose value aligns with an outcome advantageous to the recipient, such as securing a job, loan approval, or avoiding termination. The protected attribute refers to a subset of a population whose outcomes should exhibit parity, encompassing factors like gender, age, religion, veterans status, sexual orientation, or education level. The privileged value represents a historical advantage for a specific group within a protected attribute. A fairness metric quantifies undesirable bias in training data or models. Discrimination or unwanted bias arises when certain groups, whether privileged or unprivileged, face systematic advantages or disadvantages in a fair or unfair manner. This phenomenon is often linked to attributes like age, gender, sexual orientation, marital status, veterans status, and education level (Mahoney et al., 2020b).

4.4 Pre-Processing Algorithms

We can attempt to manipulate the training data before training the algorithm by reducing the bias in the pre-processing algorithms. Four main types of pre-processing methods are used to mitigate bias in the training data (Kamiran and Calders, 2011). These are suppression, massaging, reweighting, and sampling. Covariance Analysis is a part of suppression and is a method explored by (Ángel Pavón Pérez et al., 2023). The following are in the AIF360 package as of early 2020: Reweighting pre-processing, optimized pre-processing, learning fair representations, and disparate impact remover (Mahoney et al., 2020c). Those are discussed in the following sections.

4.4.1 Suppression

Suppression is the removal of the sensitive attribute and those highly correlated with it (Valentim et al., 2022). This involves eliminating the attributes that have the highest correlation with the sensitive attribute S, as well as S itself, in order to reduce bias between the class label and that attribute (Kamiran and Calders, 2011). This can reduce discrimination in downstream tasks in some cases (Ekstrand et al., 2021). In (Verma et al., 2021), the following process is done to remove bias in the training data: (1) generate a pool of pairs of similar individuals and find the discriminating pairs with them and determine if the two individuals were not treated fairly. (2) Rank the training data points according to their contribution to the unfair decisions. (3) Remove some of the highest ranked training data points and retrain a model with a lower individual discrimination.

4.4.1.1 Covariance Analysis

In (Ángel Pavón Pérez et al., 2023), the following process is described for the pre-processing algorithm: (1) create a ML model to predict sensitive attributes. (2) Discard attributes related with the sensitive attribute based on their significance. (3) Create a ML model with the remaining attributes. (4) Evaluate the model on both accuracy and fairness. For step (2), this is where the covariance analysis comes in. Features selection is done using Chi-Squared Independence test and Mann-Whitney U test. The Chi-Squared test uses a significance value to test for independence of the categorical target to the categorical attribute. The Mann-Whitney U test analyzes whether the numerical attributes are distributed differently across the categorical target.

4.4.2 Reweighting Process

Reweighting involves carefully assigning weights to certain inputs to reduce discrimination (Ekstrand et al., 2021). We can correct past incorrect cases by weighting correct cases more heavily and weighting incorrect (discriminatory) cases less (Nielsen, 2020). When reweighting, we generate a weight for the training sample in each group label combination differently to ensure fairness before classification (Mahoney et al., 2020c). The example given in (Krasanakis et al., 2018), considers objects with $X(S) = b$ and $X(Class) = +$ that will get higher weights than objects with $X(S) = b$ and $X(Class) = -$ and objects with $X(S) = w$ and $X(Class) = -$. We restrict ourselves to one binary sensitive attribute S with domain b, w and a binary classification problem target attribute Class with domain $-, +$, where "+" is the desirable class for the data subject to satisfying $S = b$ and $S = w$ that represents the deprived and favored community, respectively.

The expected probability:

$$P_{\text{exp}}(S = b \wedge \text{Class} = +) := \frac{|\{X \in D \mid X(S) = b\}|}{|D|} \times \frac{|\{X \in D \mid X(\text{Class}) = +\}|}{|D|} \tag{4.1}$$

The observed probability:

$$P_{\text{obs}}(S = b \wedge \text{Class} = +) := \frac{|\{X \in D \mid X(S) = b \wedge X(\text{Class}) = +\}|}{|D|} \tag{4.2}$$

The assigned weight:

$$W(X) := \frac{P_{\text{exp}}(S = X(S) \wedge \text{Class} = X(\text{Class}))}{P_{\text{obs}}(S = X(S) \wedge \text{Class} = X(\text{Class}))} \tag{4.3}$$

Note the weights in the following figure that are calculated by $X(S) = f$ is 60% and for $X(Class) = +$ is 50% for the expected values, while the observed value is 40% for F+ ; thus, $W(X) = \frac{0.6 \times 0.5}{0.4} = 0.75$ is the first entry in the following table. The expected values are shown to the right of the table.

Sex	CL	Weight		
F	+	0.75	F	0.6
F	+	0.75	M	0.4
F	+	0.75	M+	0.1
F	-	1.50	M-	0.3
M	+	2.00	F+	0.4
M	-	0.67	F-	0.2
M	-	0.67	+	0.5
M	-	0.67	-	0.5
F	+	0.75		
F	-	1.50		

Figure 4.1: Reweighting Example

The study by (Krasanakis et al., 2018) employs a convex underlying label error perturbation technique in training the model. This method aims to transform classification errors into a probability measure, facilitating the adjustment of estimated labels towards the desired underlying labels. Subsequently, this probability measure is utilized to infer training weights based on classifier outputs. The process involves iteratively retraining the classifier using these updated weights.

4.4.3 Resampling

In (Kamiran and Calders, 2011), sampling is described as a process by which the data set is first divided into four parts, DP (deprived positive), DN (deprived negative), FP (favored positive), and FN (favored negative). Then preferential sampling is introduced as a method to focus on data objects near the decision boundary, which are more likely to be affected by discrimination. This involves sorting data objects by their probability of belonging to the positive class and adjusting the group sizes by either decreasing or increasing the data set by duplicating objects close to the boundary or removing them, depending on whether the group needs to be enlarged or reduced. This iterative process ensures that adjustments reflect an effort to counteract discrimination, aiming for a fairer representation of all groups in the training dataset, ultimately leading to the learning of a less biased classifier.

4.4.4 Massaging

Massaging the data is done by altering class labels from negative to positive for sensitive groups and vice versa until discrimination is minimized (Ekstrand et al., 2021). Going back to our example from (Kamiran and Calders, 2011), with $X(S) \in \{b, w\}$ and $X(Class) \in \{-, +\}$ we can massage our data by changing some objects X with $X(S) = b$ from - to $+$, and some number of objects with $X(S) = w$ from $+$ to -. In this way we can decrease the discrimination so that the overall class distribution stays the same. The promotion candidates pr will be X with $X(S) = b$ and $X(Class) = -$ while demotion candidates dem will be $X(S) = w$ and $X(Class) = +$. The ranks will be according to their positive class probability and how it is learned and is not random. A higher score indicates a higher chance to be in the positive class, so the promotion candidates will be sorted in descending order, while the demotion candidates will be sorted in ascending order by rank score; thus, objects closer to the decision border will be selected first to be relabeled, leading to minimal effect on the overall accuracy. M is defined as the number of pairs that need to be modified to be discriminate free. The following formulas are given in (Kamiran and Calders, 2011) to compute the M that provides zero discrimination:

$$\frac{p_w - M}{|D_w|} - \frac{p_b + M}{|D_b|} = \text{disc}(D) - M\left(\frac{1}{|D_b|} + \frac{1}{|D_w|}\right) = \text{disc}(D) - \frac{M|D|}{|D_w||D_b|} \tag{4.4}$$

In order to reach zero discrimination the following must be achieved where D_b and D_w are data with $S = b$ and $S = w$, respectively, and p_b and p_w are the number of positive objects w.r.t $S = b$ and $S = w$. Note that D is the given data set and

$$M = \frac{\text{disc}(D) \times |D_b| \times |D_w|}{|D|} \tag{4.5}$$

4.4.5 Optimized Pre-Processing

Optimized pre-processing uses a probabilistic transformation that modifies the features and labels in the data set, adhering to constraints and objectives related to group fairness, individual distortion, and preservation of data fidelity (Mahoney et al., 2020c).

4.4.6 Learning Fair Representation

Learning fair representation discovers a hidden representation that effectively encodes the data while simultaneously concealing details about

protected attributes (Mahoney et al., 2020c).

4.4.7 Disparate Impact Remover

Modifies the features values to improve fairness between groups while maintaining the rank order within each group (Mahoney et al., 2020c).

4.5 In-Processing Algorithms

During the in-processing stage of algorithm development, we can mitigate undesirable bias by incorporating loss functions into the training process. This approach effectively adjusts the training algorithm to reduce bias (Mahoney et al., 2020c).

4.5.1 Loss Function Optimization in Classification Problems

In fair classification we can minimize disparate impact to maximize fairness with accuracy being the constraint with the following optimization problem with θ^* being the optimal loss function obtained by training the unconstrained classifier.

$$\begin{aligned} \min \quad & \left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{Z}_i - \bar{\mathbf{Z}}) \theta^T \mathbf{x}_i \right| \\ \text{subject to} \quad & L(\theta) \leq (1+\gamma) L(\theta^*) \end{aligned} \tag{4.6}$$

$\gamma \geq 0$ is the additional loss allowed over the unconstrained classifier. If $\gamma \to 0$ the optimal loss of the non-constrained classifier should almost equal the constrained classifier and reach the maximum achievable accuracy (Mukherjee, 2023a). We improve accuracy while adhering to fairness constraints by optimizing a specific loss function in the training data set to estimate θ. As c approaches zero, the classifier becomes more equitable. This is attributed to c serving as a limiting factor in the covariance between sensitive attributes and the signed distance from the feature vector to the decision boundary

(Mukherjee, 2023a):

$$\begin{aligned} &\min\ L(\theta) \\ &\text{subject to} \\ &\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \bar{\mathbf{Z}})\theta^T\mathbf{x}_i \leq \mathbf{c}, \\ &\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \bar{\mathbf{Z}})\theta^T\mathbf{x}_i \geq -\mathbf{c}. \end{aligned} \tag{4.7}$$

4.5.1.1 Logistic Regression

When using logistic regression, the loss function given is the maximum likelihood, as discussed in (Mukherjee, 2023a):

$$L(\theta) = -\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i, \theta) \tag{4.8}$$

Thus, the optimization problem becomes:

$$\begin{aligned} &\min -\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i, \theta) \\ &\text{subject to} \\ &\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \bar{\mathbf{Z}})\theta^T\mathbf{x}_i \leq \mathbf{c}, \\ &\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \bar{\mathbf{Z}})\theta^T\mathbf{x}_i \geq -\mathbf{c}. \end{aligned} \tag{4.9}$$

Recall $y_i \in \{1, -1\}$, x_i is the feature vector and z_i being the sensitive attributes.

4.5.1.2 Linear SVM

For linear SVM we use the following SVM loss function where $C\sum_{i=1}^{N}\xi_i$ is the adjustment to penalize the data points falling inside the margin. See

(Mukherjee, 2023a):

$$\begin{aligned} L(\theta) &= \min ||\theta||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i\theta^T\mathbf{x}_i \geq 1-\xi_i, \quad \forall i \in \{1,\ldots,N\}, \\ & \xi_i > 0, \quad \forall i \in \{1,\ldots,N\}. \end{aligned} \tag{4.10}$$

The full optimization problem then becomes:

$$\begin{aligned} \min \quad & ||\theta||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i\theta^T\mathbf{x}_i \geq 1-\xi_i, && \forall i \in \{1,\ldots,N\}, \\ & \xi_i \geq 0, && \forall i \in \{1,\ldots,N\}, \\ & \frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \bar{\mathbf{Z}})\theta^T\mathbf{x}_i \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \bar{\mathbf{Z}})\theta^T\mathbf{x}_i \geq -\mathbf{c}. \end{aligned} \tag{4.11}$$

4.5.1.3 Non-Linear SVM

Non-linear SVM can be achieved with the use of a kernel function to map the feature vectors x into a higher dimensional space using some operator $\phi(\cdot)$. The decision boundary is $\theta^T\phi(qx) = 0$. We must use the dual of the optimization problem since the features space is complex due to the transformation making the quadratic problem hard to solve (Mukherjee, 2023a). The following formulation can be seen where α are the dual variables, $g_{\boldsymbol{\alpha}}(\mathbf{x}_i) = \sum_{j=1}^{N}\boldsymbol{\alpha}_j y_j k(\mathbf{x}_i, \mathbf{x}_j)$ is the newly signed distance from the decision boundary in the transformed space and the function $h_{\boldsymbol{\alpha}}(\mathbf{x}_i) = \sum_{j=1}^{N}\boldsymbol{\alpha}_j y_j \frac{1}{c}\delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Note that $k(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function.

$$
\begin{aligned}
&\min \left(\sum_{i=1}^{N} \alpha_i + \sum_{i=1}^{N} \alpha_i y_i \left(g_i(\mathbf{x}_i) + h_\alpha(\mathbf{x}_i) \right) \right) \\
\text{subject to} \quad &\alpha_i \geq 0, \quad \forall i \in \{1, \ldots, N\}, \\
&\sum_{i=1}^{N} \alpha_i y_i = 0, \\
&\frac{1}{N} \sum_{i=1}^{N} (\mathbf{Z}_i - \bar{\mathbf{Z}}) \theta^T \mathbf{x}_i \leq \mathbf{c}, \\
&\frac{1}{N} \sum_{i=1}^{N} (\mathbf{Z}_i - \bar{\mathbf{Z}}) \theta^T \mathbf{x}_i \geq -\mathbf{c}.
\end{aligned}
\tag{4.12}
$$

4.6 Post-Processing Algorithms

Post-processing algorithms may be used when users have access to only black-box models such that bias can be reduced by manipulating the output predictions after training the algorithm (Mahoney et al., 2020c). One such method is through thresholding which uses different decision boundaries for different groups to ensure non-discriminatory outcomes under some definition leaving the data and model alone while post-processing the model output to achieve fairness (Ekstrand et al., 2021).

4.6.1 Fair Classification

For a binary classification problem, we find a function $f(x)$ that maps a characteristic vector x to a binary label $y \in \{-1, 1\}$ using a set of training data points N, which can be represented as $\{(x_i, y_i)\}_{i=1}^{N}$. This is the margin-based classifier's decision boundary, expressed as a set of parameters denoted by θ. The objective is to find θ^* that corresponds to the decision boundary resulting in the best classification accuracy in the test set by minimizing the loss function L in the training set (Mukherjee, 2023a).

As described by (Xian et al., 2023), the objective in fair classification is to develop a randomized classifier that is attribute-aware, denoted as $h : \chi \times A \to Y$, that minimizes the classification error rate over μ, while adhering to the constraints imposed by specified fairness criteria. The classifier associated with a particular group a is represented as $h_a : \chi \to Y$, where χ is

the input space and h_a is equivalent to $h(x, a)$. The error rate is defined as:

$$\begin{aligned} \text{err}(h) &:= P(h_A(X) \neq Y) \\ &= \sum_{a \in [m]} w_a P(h_A(X) \neq Y | A = a) \\ &= \sum_{a \in [m]} w_a \int_{X \times Y} P(h_a(x) \neq y) \, d\mu_a(x, y), \end{aligned} \tag{4.13}$$

The decomposition presented in the final line emphasizes the stochastic nature of the classifier h. Note this is a classification scenario involving k classes. This scenario is characterized by a joint distribution, μ, which encompasses the input X from a set $\mathcal{X}$, a demographic group membership—often referred to as the sensitive attribute—A, within a set $\mathcal{A}$ that is defined as $[m] = \{1, \ldots, m\}$, and the class labels represented as one-hot vectors Y within a set $\mathcal{Y} = \{e_1, \ldots, e_k\}$. The marginal distribution for the input X is given by μ_X. For a specific group $A = a$, the distribution μ conditioned on A is denoted by μ_a. Furthermore, the proportion of the population belonging to group a is indicated by w_a, defined as $w_a := P_\mu(A = a)$.

An example given is approximate demographic parity in which $\alpha \in [0, 1]$ a classifier $h : \chi \times A \to Y$ is to satisfy $\alpha - DP$ if $\Delta_{DP}(h) \leq \alpha$ is defined as:

$$\begin{aligned} \Delta DP(h) &:= \max_{a,a' \in [m]} \max_{y \in Y} |P(h_A(X) = y | A = a) - P(h_A(X) = y | A = a')| \\ &= \max_{a,a' \in [m]} \left| h_a \sharp \mu_{X_a} - h_{a'} \sharp \mu_{X_{a'}} \right|_\infty, \end{aligned} \tag{4.14}$$

where

$$P(h_A(X) = y | A = a) = \int_X P(h_a(x) = y) d\mu_{X_a}(x). \tag{4.15}$$

is the proportion of outputs with class assignment y on group α and

$$\|p - q\|_\infty := \max_{z \in Z} |p(z) - q(z)|, \tag{4.16}$$

between two distributions p,q. Note that the push forward measure is used, denoted as $f \sharp p$ and involves a randomized function f that maps from a measurable space (X, S) to another space (Y, T), utilizing a Markov kernel K. For any given measure p on (X, S) and for every T in the target space $\mathcal{T}$, the push-forward measure on Y is succinctly defined as:

$$f \sharp p(T) = \int_X K(x, T) dp(x), \tag{4.17}$$

where $K(x, T)$, a function or kernel defined over $X \times \mathcal{T}$, enables the measure p to transition from its original domain X to the new domain Y. This expression concisely describes how the measure p is projected to the space Y through the mapping facilitated by f and articulated by the kernel K (Xian et al., 2023).

4.6.2 Fair Regression

In (Mukherjee, 2023a), fair regression is accomplished by identifying a function $f(x)$ that maps a feature vector x to a continuous label $y \in [0, 1]$ through the utilization of a set of N training data points, represented as $\{(x_i, y_i)\}_{i=1}^{N}$, where it is noted that all y_i are real-valued and continuous, i.e., $y_i \in [0, 1]$. The performance of this mapping is quantified by a loss function L. Specifically, the paper models the GPA of first-year college students as a regression problem, where each y_i corresponds to a GPA normalized between 0 and 1. The loss function used is the squared loss, defined as $L(y_i, f(x_i)) = \frac{(y_i - f(x_i))^2}{2}$. Next they ensure that statistical parity is achieved by setting $P[f(x_i) \geq s|z = 1] = P[f(x_i) \geq s]$ where $s \in [0, 1]$ and z is the protected attribute thus we have the optimization problem:

$$\begin{aligned} &\min_{f \in F} \mathbb{E}[L(y_i, f(x_i))] \\ &\text{subject to} \quad |P[f(x_i) \geq s \,|\, z = 1] - P[f(x_i) \geq s]| \leq c \end{aligned} \tag{4.18}$$

where F in the family of function form which f is to be searched and different values of c can be chosen to vary in the strength of the constraint in favor of the protected group where c is the fairness accuracy knob like the parameter used in equation 7.

4.6.3 Fair Clustering

One of the most popular fairness metrics used in fair clustering is balance. It assumes that there are N protected groups and K different clusters with ρ and ρ_c be the proportion of samples of dataset belonging to the protected group z and the proportion of samples in cluster $C \in [K]$ belonging to protected group z. Let $\rho_{C,z} = \frac{\rho}{\rho C}$j be based on the balance can be formulated over all the clusters and the protected groups as follows (Mukherjee, 2023a):

$$\min_{C \in [K], z \in [N]} \min \left(\rho_{C,z}, \rho_{C,z}^{-1}\right) \tag{4.19}$$

Notice that the balance metric ranges from 0 to 1, where a higher balance value signifies a more equitable clustering outcome. Consequently, a fair clustering algorithm aims to maximize the balance metric to enhance fairness. Another method shown in (Mukherjee, 2023a) is social fairness which uses K-means clustering. This uses the notion of group fairness. The cost function uses U as the centers for the K different clusters and the input dataset x defining the cost of K-means clustering as:

$$O(U, x) = \sum_{x_i \in x} \min_{U_i \in U} \|x_i - U_i\|^2 \tag{4.20}$$

let x_z denote the samples of x that belong to the protected group z then the social fairness is defined as:

$$max_{z\in[N]}O\frac{(U,x)}{x_z} \tag{4.21}$$

In this equation, contrary to balance, which requires maximization, social fairness ought to be minimized as it represents a distance function.

4.6.4 Fair Embedding

Fair embedding involves vector representation of words with a distance function to measure the closeness of one word to another in Natural Language Processing (NLP). An example given in (Mukherjee, 2023a) related to word similarities shows bias in word embeddings. Therefore we need embeddings that are fair. Word vectors are normalized where $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ are words and a measure of similarity between two word vectors is as follows:

$$\cos(\overrightarrow{w_1},\overrightarrow{w_2}) = \frac{\overrightarrow{w_1}\cdot\overrightarrow{w_2}}{\|\overrightarrow{w_1}\|\|\overrightarrow{w_2}\|} \tag{4.22}$$

4.6.4.1 Debiasing Algorithm

(Mukherjee, 2023a) discusses the debiasing algorithm for fair embedding. The algorithm operates through three key stages: locating the gender subspace, achieving neutralization, and facilitating equalization. The initial stage involves identifying a singular axis within the embedding that represents the overarching gender dimension. The neutralization stage is aimed at ensuring that words which are gender-neutral exhibit no bias within this gender subspace. Finally, in the equalization stage, the goal is to adjust the spatial arrangement of certain word groups outside the subspace, so they are uniformly distant from gender-neutral terms, such as ensuring the term "sewing" has the same proximity to both members of the gender pair male, female. To identify to gender subspace the following three formulas are used:

The projection of $\overrightarrow{w}$ on B is defined as:

$$\overrightarrow{w_B} = \sum_{j=1}^{K}(\overrightarrow{w}\cdot\overrightarrow{b_j})\overrightarrow{b_j} \tag{4.23}$$

where B is a subspace of k orthogonal unit vectors $\{\overrightarrow{b_1},\overrightarrow{b_2},...,\overrightarrow{b_k}\}\subset\mathcal{R}^d$. If $k=1$ the subspace reduces to a single direction.

Next we need defining sets $D_1, D_2, D3, ..., D_n \subset D$, the value of the integer k, and the word embedding $\overrightarrow{w} \in \mathcal{R}^4$.

The means of the defining sets is defined as:

$$\overrightarrow{\mu_i} = \sum_{w \in D_i} \frac{\overrightarrow{w}}{D_i} \tag{4.24}$$

The average displacement of individual word vectors is calculated by $\overrightarrow{w} - \overrightarrow{\mu_i}$. Our objective is to identify the dictionary or subspace that accounts for the greatest variance within the space. This is achieved by determining the principal singular vectors from the covariance matrix X, which is defined as follows:

$$\mathbf{C} = \sum_{i=1}^{n} \sum_{\overrightarrow{w} \in D_i} \frac{(\overrightarrow{w} - \overrightarrow{\mu_i})^T (\overrightarrow{w} - \overrightarrow{\mu_i})}{D_i} \tag{4.25}$$

where the gender subspace is defined through the first k rows of the $\mathbf{SVD}(C)$

Next we neutralize the words by re-embedding them as follows:

$$\overrightarrow{w} \Longleftarrow \frac{\overrightarrow{w} - \overrightarrow{w_B}}{||\overrightarrow{w} - \overrightarrow{w_B}||} \tag{4.26}$$

where $\overrightarrow{w_B}$ is the projection of w on the gender subspace B. Essentially, this removes the gender bias that might be present in the embedding w. An example of this can be seen in the following figures.

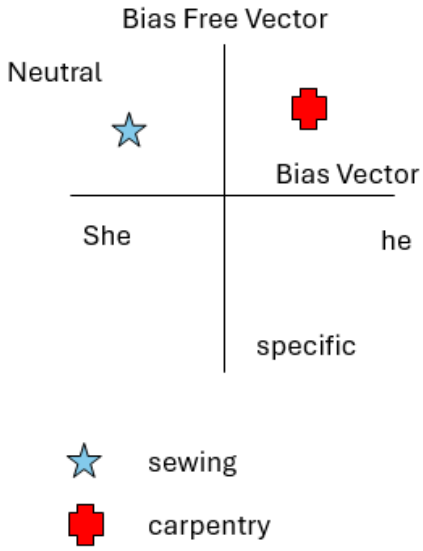


Figure 4.2: Before Neutralization

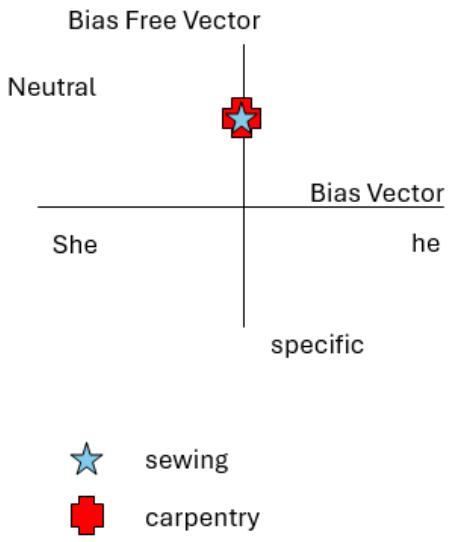


Figure 4.3: After Neutralization

The next step is equalization in which the family of equality sets are represented by $\Theta = \{E_1, E_2, ..., E_m\}$ where $E_i \in W$ and for each equality set $E_i \in \Theta$ let :

$$\overrightarrow{\mu} = \sum_{\overrightarrow{w} \in E} \frac{\overrightarrow{w}}{E} \tag{4.27}$$

Every word can be represented as the sum of its projection onto the gender subspace, $\overrightarrow{w_B}$, and its component orthogonal to this subspace, $\overrightarrow{w_\perp}$. To remove bias, we shift everything to mean zero within the bias subspace, replacing $\overrightarrow{w_B}$ with $\overrightarrow{w_B} - \overrightarrow{\mu_B}$. If we denote $\overrightarrow{w_\perp}$ as $\overrightarrow{v}$, the corrected representation of a word is $\overrightarrow{w} = \overrightarrow{v} + (\overrightarrow{w_B} - \overrightarrow{\mu_B})$. Since these are unit vectors, $||w||^2 = 1$, thus $||\overrightarrow{v} + (\overrightarrow{w_B} - \overrightarrow{\mu_B})|| = 1$ or $||\overrightarrow{w_B} - \overrightarrow{\mu_B}|| = \sqrt{1 - ||\overrightarrow{v}||^2}$. This leaves the following equation:

$$\overrightarrow{w} = \overrightarrow{v} + \sqrt{1 - ||\overrightarrow{v}||^2} \frac{(\overrightarrow{w} - \overrightarrow{w_B})}{||\overrightarrow{w} - \overrightarrow{w_B}||} \tag{4.28}$$

The plots for before and after equalization with the word sewing can be seen in the following figures. Notice that in 4.5 sewing is spaced out more equally than in 4.4.

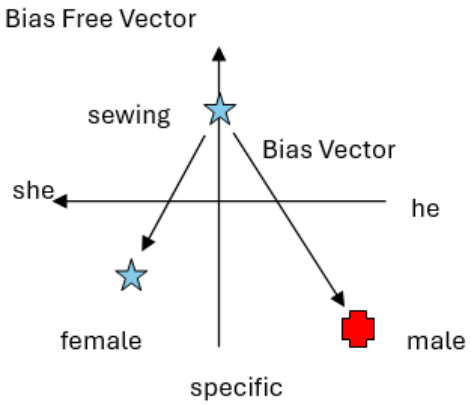


Figure 4.4: Before Equalization

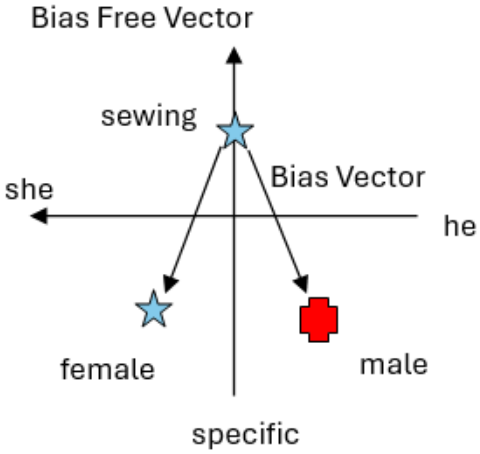


Figure 4.5: After Equalization

4.6.5 Equalized Odds

Equalized Odds are discussed in (Hardt et al., 2016). According to (Lohia et al., 2018), the equalized odds post-processing algorithm ensures that both privileged and unprivileged groups exhibit identical false negative and false positive rates. This goal is attained by solving an optimization problem to calculate the probabilities for inverting predictions, indicated as $\hat{y}_1 = 1 - \hat{y}_i$, across four defined scenarios:

- $(d_i = 0, \hat{y} = 0)$,
- $(d_i = 0, \hat{y} = 1)$,
- $(d_i = 1, \hat{y} = 0)$,
- $(d_i = 1, \hat{y} = 1)$.

Utilizing these probabilities, the modification of predictions for specific data points occurs through a random selection process.

4.7 Equity in Optimization

Fairness is measured through inequality metrics, which aim to be minimized and are often referred to as characteristics of deviation type dispersion (Ogryczak et al., 2014). For outcomes to be considered perfectly equal, these metrics should approach zero. Basic inequality metrics include those based on

deviations, like the mean absolute difference, the maximum absolute difference, the standard deviation, and the mean absolute deviation. Additionally, there are relative inequality metrics that are adjusted by the mean outcome, such as the Gini Coefficient. According to (Chen and Hooker, 2023), a generic optimization problem that incorporates equity in the form of $max_{\text{inf}}\{f(x)|x \in S_x\}$ can be stated as follows:

$$\max_{u,x}\{W(u)|u = U(x), x \in S_x\} \tag{4.29}$$

where $u = (u_1, u_2, ..., u_n)$ is a vector of utilities distributed across parities $1, ..., n$. These utilities can be profit, cost, or whatever is appropriate to the problem. By replacing the original objective function with the social welfare function $W(u)$ that measures the distribution of the utility u, then the problem becomes:

$$\max_{u,x}\{W(u)|(u, x) \in S_x\} \tag{4.30}$$

where $(u, x) \in S$ if and only if $u = U(x)$ and $x \in S_x$

They also look at fairness as a constraint with the social welfare function being bounded by some lower bound:

$$\max_{u,x}\{f(x)|W(u) \geq LB, (u, x) \in S\} \tag{4.31}$$

Chen and Hooker investigate different social welfare functions and how they evaluate over the optimization problems. Some of these are Hoover index, Gini coefficient, McLoone Index, and Max-Min Fairness.

4.7.1 Gini Coefficient

In economics, the Gini Coefficient also known as the Lorenz measure (a relative measure of absolute difference) is recognized as a prominent measure of relative inequality, which is normalized by the mean outcome. (Ogryczak et al., 2014). According to (Chen and Hooker, 2023) the Gini coefficient is related to the Lorenz curve and diagonal line representing perfect equality and therefore vanishes under perfect equality, letting the social welfare function $W(u) = -G(u)$ we have:

$$G(u) = \frac{1}{2\bar{u}n^2}\sum_{i,j}|u_i - u_j| \tag{4.32}$$

Then we linearize the model using a change of variables by introducing a

scalar variable t and write $u = u'/t$ and $x = x'/t$ which gives:

$$\min_{x',u',V,t} \left\{ \frac{1}{2n^2} \sum_{i,j} v_{ij} \;\middle|\; -v_{ij} \leq u'_i - u'_j \leq v_{ij}, \forall i, j, \;\; u' = 1, \, t \geq 0, \, (u', x') \in S' \right\} \tag{4.33}$$

4.7.2 Hoover Index

The Hoover index is also related to the Lorenz curve in that it is proportional to the maximum vertical distance between the Lorenz curve and the diagonal line that represents perfect equality (Chen and Hooker, 2023) and its social welfare function is as follows:

$$W(u) = -\frac{1}{2n\bar{u}} \sum_i |u_i - \bar{u}| \tag{4.34}$$

4.7.3 McLoone Index

The McLoone index compares the total utility of individuals at or below the median utility to the utility if they were brought up to the median utility with the index being 1 if nobody's utility is strictly below the median and it approaches 0 if the utility distribution has a very long lower tail or all the utilities are assumed positive (Chen and Hooker, 2023). The social welfare function is thus:

$$W(u) = \frac{1}{|I(u)|\tilde{u}} \sum_{i \in I(u)} u_i \tag{4.35}$$

where $\tilde{u}$ is the median of utilities u and $I(u)$ is the set of indices of utilities at or below the median such that $I(u) = \{i | u_i \leq \tilde{u}\}$.

4.7.4 Max-Min Fairness

According to (Chen and Hooker, 2023), the Rawlsian difference principle is used in the maximin criterion which states that inequality should exist only to the extent that it is necessary to improve the lot of the worst-off. The maximin criterion maximizes the social welfare function $w(u) = \min_i\{u_i\}$ which is linearized as the following optimization model:

$$\max_{x,u,w} \{w | w \leq u_i, \forall i; (u, x) \in S\} \tag{4.36}$$

4.7.5 Price of Fairness

According to (Ogryczak et al., 2014), the 'price of fairness' (POF) manifests when a fair solution results in a decrease in system efficiency as opposed to an optimization that only considers efficiency. The POF is measured by the relative difference from an optimally efficient solution that maximizes the sum of all performance functions (total outcomes). The price of fairness for a fairness concept F within a feasible set A is given by:

$$POF(F, A) = \frac{\sum_{i=1}^{m} y_i^T - \sum_{i=1}^{m} y_i^F}{\sum_{i=1}^{m} y_i^T} \tag{4.37}$$

where y^T is the outcome vector maximizing the total outcome $T(y)$ over A, and y^F is the outcome vector that maximizes the fairness concept $F(y)$ over A.

4.8 AI and Ethics

There are ethical concerns when using AI algorithms in areas such as law enforcement or online through collection methods used for recommendation algorithms. These are discussed in detail in the next section.

4.8.1 Use in Law Enforcement

AI is used in law enforcement in areas such as fraud detection, traffic accident monitoring, child pornography tracking, identifying anomalies in public spaces and among other things in an idea of crime management called predictive policing (Mukherjee, 2023c). Ethics comes into play when we are tracking individuals and if we are being bias when doing so as such in the COMPAS software. Stereotyping is an issue causing over-policing in certain areas or over certain people.

4.8.2 Data Protection and User Privacy

Privacy is an issue on various online outlets now with social media and website tracking and in apps tracking data for AI algorithms to make recommendations. Some of these are password sharing, online shopping, use of IoT, browsing the web, and posting on social media (Mukherjee, 2023c). Password sharing is mitigated by two factor authentication and fingerprint scanning. With online shopping, our credit card information is shared on

websites. IoT technology allows all devices to be able to be connected on one network, which is great to be able to access all your information from anywhere. However, if you lose one of the devices, it could be hacked on all of the devices. HTTP cookies for browsing history can be easily compromised when browsing the web. Social media is probably one of the larger areas for data breaches. Geo-tagging gives people your location if you are posting photos from home or tells if you are on vacation and for how long. Also, personal information is readily available on social media if you don't choose to lock your account down.

4.9 Conclusion

In conclusion, exploring fairness measures as well as algorithms used in the pre-processing, in-processing, and post-processing steps are vital to machine learning. The examination of algorithms across pre-processing, in-processing, and post-processing stages reveals the breadth of approaches available to mitigate bias. Techniques such as reweighting, massaging, and the optimization of loss functions during model training, as well as post-processing adjustments such as thresholding, offer targeted strategies for enhancing fairness. In addition, the adaptation of these methods to specific applications, such as fair classification, regression, clustering, and embedding, demonstrates the adaptability of fairness measures to various domains within machine learning. We also explored how ethics plays a role in AI through law enforcement and online algorithms. There is room for growth in all these areas. The various optimization methods are used in Chapter 5 to look at how the Gini coefficient and Hoover index can be applied as fairness measures to different distance metrics to optimize the placement of treatment facilities based off supply and demand using drug court data in Oklahoma. The concept of the Gini index is also used in Chapter 7 when looking at fairness in an epidemic model using poverty by county.

Chapter 5

Optimizing Treatment Facility Locations in Oklahoma Using Haversine, Euclidean, Manhattan, and Chebyshev Distance Optimization

5.1 Introduction

Computational Criminology (CC) is a field in which computer simulations and models are applied to study topics of interest for criminologists (Berk, 2008). In this chapter, we develop a CC application using a Facility Location Problem (FLP) model (Cantlebary and Li, 2020). FLP is a classic optimization problem that determines the best location for a factory or warehouse to be placed based on geographical demands, facility costs, and transportation distances. In our work we develop a multi-objective optimization model to optimally place rehabilitation treatment facilities in Oklahoma by minimizing cost, total of various distance metrics and maximizing the total number of facilities located in each region, while not exceeding the number of facilities allowed to be located according to the maximum covering location problem.

Prisons in Oklahoma are over capacity resulting in a higher cost to the taxpayers. There is also a shortage of available beds in treatment centers. Many people reoffend because they are not given the opportunity to go to treatment before they are sent to prison. Also, many rehab facilities require insurance that most addicts and alcoholics don't have the necessary treatment when they are at the point in their addiction that requires treatment. Fewer crimes would occur if an offender had not been suffering from addiction.

More access to treatment centers may help decrease reoffending as well as drug related co-occurring crimes from happening. Statistically speaking, the United States has some of the highest recidivism rates in the world with almost 44% of the criminal released returning to prison within the first year out of prison. 68% of drug offenders are rearrested within 3 years of release from prison (Belenko et al., 2013b). Oklahoma itself ranks 3rd if looking at the world's incarceration rate considering every state as a country (Wildra and Herring, 2022). For these reasons, a focus on treatment facilities and their locations is used to try to meet the demand of drug offenders' rehabilitation in Oklahoma as opposed to a prison focused approach. Several methods are being incorporated in this research such as use of the Haversine distance formula, Gini Coefficient and Hoover Index fairness measures. The Haversine distance formula (the great circle distance) is focused on in (Akkihal, 2006). We use the Haversine formula because it looks at the spherical shape of the earths distance with longitude and latitude ignoring valley depth and hill height (Maria et al., 2020). Other distance metrics such as the Manhattan distance, Euclidean distance, and the Chebyshev distance are compared in great detail in (Mora et al., 2016). The Manahattan distance looks at a rectangular or right angle travel metric (Batta et al., 1990). The Euclidean distance is based on direct distance without obstacles such as to get the value of the length of a diagonal line on a triangle (Maria et al., 2020). The use of many fairness measures as constraints in optimization were explored in (Chen and Hooker, 2023). In this chapter's research we focused specifically on Gini Coefficient and Hoover Index. We also look a multi-criteria optimization approach. Multi-criteria location problems in three categories of bi-objective, multi-objective and multi-attribute problems are explored in the survey (Farahani et al., 2010). The goal of this Chapter is to minimize cost and distance while maximizing the coverage of demand by placing rehabs in optimal regions in Oklahoma to best help addicts and alcoholics across the state, while also comparing different distance metrics in conjunction with two fairness measures. The two fairness measures used are the Hoover Index and Gini Coefficient which will be applied to distribute the facilities more equally. Gurobi optimizer was used in Python and there was no weight set on the objective functions, instead the priorities were set at 1,2, and 3. This leaves the weight at a default value of 1 for all objectives for their given priority level. They were solved in the order in which the priority was set (Gurobi Optimization, LLC, 2023).

In this case we are looking at minimizing distance and cost while maximizing number of facilities. We combined this with the method of Maximal Covering Location Problem (MCLP) model (Rader, 2010) by covering as many demand sites as possible using only P facilities. The P-Center Problem is defined as the minimization of the maximum distance between demand site and its closest facility while determining which facility is used to meet that demand, (Rader, 2010). For our suggested model we will be using the center

of each county as the starting and ending latitude and longitude location minimizing the distance with use of the squared Haversine distance function for each region while covering all demand with the closest facility without exceeding the total number of facilities built.

We can utilize the concepts from the literature in the previous chapters by applying the mentioned fairness measures with each of the distance metrics we have chosen and see how the final layout of our facilities is affected.
The Chapter is organized as follows: in section 2 we discuss the problem statement. Sections 3 and 4 give an overview of the research questions and our unique contributions. In section 5 we briefly discuss the methodology and assumptions of the model. In section 6 the necessary background is developed. In section 7 we provide an analysis of our models. Section 8 discusses two fairness measures that expands our methods considering fairness issues. In section 9 our results are discussed. Finally section 10 concludes our Chapter.

5.2 Problem Statement

Oklahoma faces significant challenges with overcrowded prisons and a shortage of rehabilitation treatment facilities. This issue exacerbates recidivism rates, particularly among drug offenders, who often lack access to necessary treatment. The current distribution of rehabilitation centers is uneven, leading to inefficiencies and increased travel distances for individuals seeking treatment. This study aims to optimize the placement of treatment facilities across Oklahoma's eight regions. The objective is to minimize costs and travel distances while maximizing the number of facilities and ensuring fair distribution using various distance metrics (Haversine, Euclidean, Manhattan, and Chebyshev) and fairness measures (Hoover Index and Gini Coefficient). By addressing these factors, the goal is to improve access to treatment, reduce recidivism, and alleviate the burden on the state's prison system.

5.3 Research Questions

- **How do different distance metrics (Haversine, Euclidean, Manhattan, and Chebyshev) influence the optimal placement of rehabilitation facilities across Oklahoma's regions?**
  - Comparison of four distance metrics to determine the most efficient for minimizing travel distance.
- **How do fairness constraints such as the Hoover Index and Gini**

Coefficient affect the distribution and placement of rehabilitation facilities?

  - Analysis of fairness constraints to ensure equitable distribution compared to models without fairness.

- **Which distance metric, when combined with fairness constraints, provides the most effective balance between minimizing distance and equitable access?**

  - Evaluation of fairness and distance metrics to balance distance minimization and equitable facility placement.

5.4 Our Contributions

In this paper, we make the following contributions:

1. **Development of a Multi-Criteria Optimization Model**: We formulate an integer multi-criteria nonlinear programming model to optimize the placement of rehabilitation treatment facilities across Oklahoma. The model aims to minimize costs, travel distances, and maximize the number of facilities while adhering to the constraints of the maximum covering location problem.

2. **Application of Multiple Distance Metrics**: We explore and compare the performance of four distance metrics—Haversine, Euclidean, Manhattan, and Chebyshev—in the context of the facility location problem. This comparative analysis helps determine the most effective metric for optimizing travel distances in real-world geographical scenarios.

3. **Incorporation of Fairness Measures**: To ensure equitable distribution of facilities, we integrate fairness measures such as the Hoover Index and Gini Coefficient into our model. We evaluate the impact of these measures on the placement of facilities and demonstrate how they can be used to achieve a fairer allocation.

4. **Use of Real-World Data**: Our model is based on actual drug court data from Oklahoma counties, ensuring that our findings are grounded in real-world conditions and are relevant to policy-making and implementation.

5. **Computational Analysis and Solver Utilization**: We employ the Gurobi solver to handle the complex optimization tasks, providing a detailed analysis of the model's performance across different scenarios. We present the results of our computational experiments, highlighting the efficiency and effectiveness of our approach.

6. **Policy Implications and Recommendations**: Based on our findings, we provide practical recommendations for policymakers in Oklahoma. Our

study offers insights into how optimized placement of treatment facilities can reduce recidivism, lower costs, and improve access to rehabilitation services.

5.5 Methodology and Assumptions

Four distance metrics were applied in the optimization models to evaluate their impact on total travel distance (Deza and Deza, 2009):

- **Euclidean Distance:** Measures the straight-line distance between two points on a flat plane. This was used as the baseline distance measure.
- **Haversine Distance:** Accounts for the curvature of the earth by measuring the shortest path between two points on a sphere (great-circle distance). This metric is commonly used for calculating distances between geographic coordinates.
- **Manhattan Distance:** Measures the distance along a grid-based path, often referred to as the "taxicab" distance, as it represents movement through a series of right-angle turns.
- **Chebyshev Distance:** Determines the greatest distance in any single dimension, making it a useful measure in situations where diagonal movement is allowed.

The models were solved using each of these distance metrics to identify which metric led to the most efficient facility placement in terms of total travel distance. Applying a Haversine distance metric instead of the Euclidean distance metric gives a more precise distance when using latitude and longitude between two points on a map. The Haversine distance was chosen over a linear distance metric because it is the best measure of the spherical distance that is typically used in the airline industry. Although this is not needed in Oklahoma, this could be beneficial for future work when this model is extended over the entire United States. The Euclidean metric measures the distance between two points on a flat 2D plane while Haversine measures the distance in a 3-dimensional space. Google maps utilized the Haversine distance formula as shown in (Maria et al., 2020). The longitudinal and latitudinal coordinates can be converted to 2-dimensional coordinates for the Euclidean formula by multiplying by 111.32 to convert degrees to kilometers. This was done to compare the Euclidean distance to the nonlinear Haversine distance in our models.

To ensure an equitable distribution of treatment centers across the eight Oklahoma regions, we incorporated two fairness measures (Cowell, 2011):

- **Hoover Index:** This measure evaluates the inequality in the distribution of facilities by comparing the actual distribution to a perfectly equal distribution. It was applied at thresholds of 0.1, 0.2, and 0.3 to determine its impact on facility placement.
- **Gini Coefficient:** A widely-used measure of inequality, the Gini coefficient was also applied at thresholds of 0.1, 0.2, and 0.3. Like the Hoover Index, it was used to promote a fairer distribution of treatment centers.

Both models were ran with the Gini Index and Hoover index at various thresholds to compare which models gave the fairest output with the least distance traveled along with the lowest cost to the tax payer.

The following assumptions were made to simplify the model and ensure its feasibility:

- Each treatment center has a capacity of 30 beds.
- The cost to the taxpayer is 10% of the per capita income averaged across the region.
- The demand for each region is based solely on the number of individuals in drug court, aggregated at the county level.
- The 8 Oklahoma regions used in this study correspond to the regions defined by the Oklahoma Office of Homeland Security (Figure 3).
- The distance between regions was calculated from the central county of each region to the central county of another.
- Travel within the same region is assumed to have zero distance, prioritizing treatment facilities within an individual's home region.

This methodology allowed us to evaluate the optimal placement of treatment centers using different distance metrics and fairness measures, while addressing the constraints and objectives of the problem.

Gurobipy was used as the optimizer. In this optimizer, prioritization was set pre-emptively using priority settings as follows: 1) Maximize Coverage 2) Minimize distance 3) Minimize Cost According to Gurobi Optimizer Reference Manual (Gurobi Optimization, 2024), a preemptive hierarchical or lexicographic multi-objective optimization ensures that lower-priority objectives are only optimized within the solution space that does not degrade higher-priority objectives with larger priority values indicating higher priorities. (Section 12.2.2). Thus, the first priority result is found, and then the next priority result is found such that the higher priority is not hurt, and so on down in

the hierarchy. With the priorities set in this order, the coverage stays the same in each of the optimization problems with 38 facilities being placed. The distance increases on all the models except for the model using the Manhattan distance metric and the cost decreases. So, while adding the fairness measure may make the model better in one objective, it makes it worse in another objective in all but one model. With fairness being added as a constraint, its priority is set before the objective functions by bounding the new solution space, setting it at a higher priority than that of the objective functions in preemptive prioritization, making it solve for fairness first. For this reason, the situation can occur in which the model can improve in the one scenario that we see taking place with the Manhattan distance with Hoover index. The fairness constraint altered the feasible region, which allowed different optima within the same maximum coverage level. This means that the model found a different layout of facilities that still covered everyone just as well, but also happened to lower the total travel cost because the fairer placement spread facilities more evenly across the state. Also, the lower cost is due to the spread happening over the spread including the lower income counties multiplied by the corresponding fixed income. With the model not including fairness all facilities places in Pottawatomie County made people travel from all areas of Oklahoma to one region of Oklahoma calling for a greater distance and with a higher population size leading to higher fixed income causing the cost to be higher than in the fair models also with cost being solved as the last priority objective function.

5.6 Background

Data were obtained from the Drug Court Performance and Outcome Report for 2002-2005 for Oklahoma (Oklahoma Department of Mental Health and Substance Abuse Services, 2006). Since demand is based on the number of people in the drug court, the limitations are based only on the counties that have a drug court program. In Figure 5.1, the data rolled up to the county can be seen. The region each county falls in is listed, as well as the number of people in drug court, the number of rehabilitation centers, and the number of beds available (number of rehabs multiplied by 30 beds), and the per capita income for that county. The counties highlighted in yellow are larger counties, such as Oklahoma and Tulsa Counties. Notice that the higher populated counties such as Tulsa and Oklahoma Counties have a significantly higher number of people in drug court and the number of treatment centers available than the number of people in drug court and the number of available treatment centers in the smaller counties such as Lincoln and Ottawa Counties in orange.

County	Region	# DC	total county beds	total county centers	per capita incom	population
Cherokee	4	99	30	1	16084	48098
Cleveland	6	58	30	1	25831	299587
Comanche	3	14	120	4	20778	123046
Craig	2	41	30	1	18784	14123
Creek	4	210	30	1	21891	72699
Custer	1	11	30	1	22003	27886
Delaware	2	23	30	1	20142	41413
Garfield	1	4	90	3	22812	61920
Grady	3	9	60	2	21687	56658
Hughes	5	61	30	1	18083	13407
Jackson	3	40	90	3	21249	24556
LeFlore	5	151	30	1	17357	48907
Lincoln	6	26	30	1	20774	34188
Mayes	2	88	30	1	19975	39589
McClain	6	61	30	1	23556	45306
McCurtain	5	40	30	1	17456	30931
Muskogee	4	78	30	1	19161	66354
Oklahoma	8	559	480	16	25723	802559
Okmulgee	4	22	30	1	19071	36990
Ottawa	2	27	30	1	17638	30338
Payne	2	193	60	2	19540	82794
Pottawatomie	6	93	90	3	19437	73533
Rogers	2	146	30	1	25358	98836
Stephens	3	11	30	1	22790	43710
Tulsa	7	711	300	10	26769	677358
Wagoner	4	177	30	1	24049	86644

Figure 5.1: Data Roll Up by County

Next, we can see the data rolled up by region in Figure 5.2. The cost to taxpayers can be seen in the 10 percent rehab cost. This is for one person to stay 30 days in a rehabilitation center. As previously mentioned, this is assumed to be 10% of the average per capita income for that region. To better understand the data, a beds needed column was added which takes the supply (number of available beds) – the demand (the number of people in drug court). Regions 1 and 3 are green meaning that they have an overage of available beds that are not being used. All other regions in red show how many beds are needed to cover the demand of addicts and alcoholics for that region.

Region	# DC (Demand)	# IR - Beds (Supply)	# IR - Centers (Supply)		Beds needed	average per capita income	10 percent rehab cost	population
1	15	120	4	105	105 Beds not used	22407.5	2241	89806
2	518	210	7	-308	308 Beds needed	20239.5	2024	307093
3	74	300	10	226	226 Beds not used	21626	2163	247970
4	586	150	5	-436	436 Beds needed	20051.2	2005	310785
5	252	90	3	-162	162 Beds needed	17632	1763	93245
6	238	180	6	-58	58 Beds needed	22399.5	2240	452614
7	711	300	10	-411	411 Beds needed	26769	2677	677358
8	559	480	16	-79	79 Beds needed	25723	2572	802559

Figure 5.2: Data Roll Up by Region

In Figure 5.3, the regions of Oklahoma based on the Oklahoma Office of Homeland Security can be seen. Since this analysis will be based on minimizing the various distance metrics, this will be a fully connected graph with every arc between every region being considered. Let each region be a node and each arc be the possible paths to be traveled between bordering regions. The arcs and corresponding longitudes and latitudes of each can be seen in Figure 5.4. For this model, we will assume that the distance between

region i and region j is from the center of counties of each region. We will also just focus on distance traveled in km for the various distance functions and not consider the time to travel between regions. However, the objectives could easily be modified to base the analysis off of time instead of distance.

Region	Center County	latitude	longitude
1	Dewey	36.0175	98.9245
2	Washington	36.6771	95.9410
3	Stephens	34.5203	97.8722
4	Muskogee	35.7111	95.3103
5	Pittsburg	34.9879	95.8143
6	pottawatomie	35.2754	97.0068
7	Tulsa	36.1593	95.9410
8	Oklahoma	35.6038	97.3517

Figure 5.3: regions, latitude/longitude, center county

5.7 Analysis

5.7.1 The Network

The network model representation of this model can be seen in Figure 5.4. The numbered nodes correspond to the number name of the region from the map above. The number along the arcs represents the distance in miles from region i to region j. In parenthesis, the numbers represent the supply and demand at region i and the fixed cost at region j respectively. They are represented as (s_i, d_i, f_j). For this model, we allow for a facility to be built in every region. The fully connected graph can be seen in the following figure, and every node has the potential to be connected to every other node.

5.7.2 The Euclidean Model

The decision variables, parameters, and model are shown on the next page. The priorities for this multi-criteria optimization model were set as follows:

- P_1: Minimize Total Cost
- P_2: Minimize Total Euclidean Distance Travelled

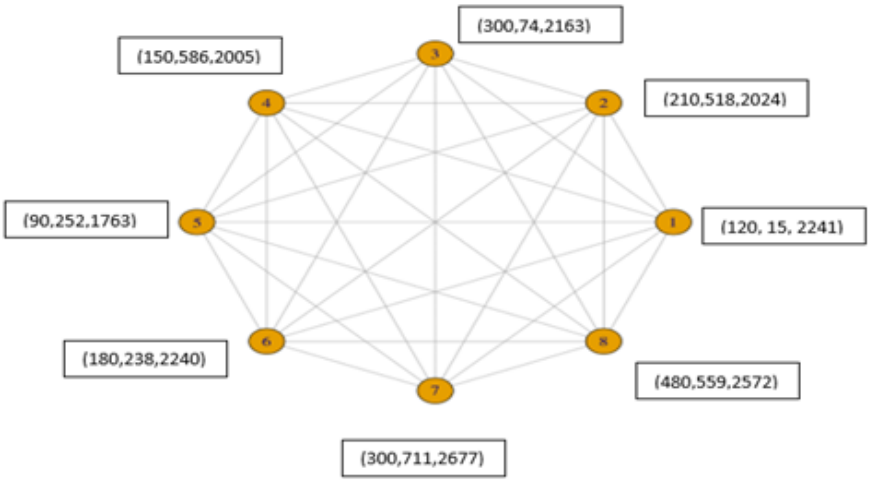


Figure 5.4: The Network

- P_3: Maximize the Total Facilities Built

Parameters were set to not exceed the total number of facilities P, needed to cover the entire demand of the state. This is a method used in the P-centered and maximal covering location problems mentioned in the introduction. It is important to note that this model is an integer model as we are not allowed to have a fraction of people or facilities built.

Let Y_j be the number of facilities at region j, and X_{ij} be the number of people traveling from region i to region j.

Decision Variables

X_{ij} = The number of people traveling from node i to node j
Y_j = The number of facilities built at node j

Parameters

V = Set of demand nodes (Regions)
A = Set of arcs (from region i to region j)
s_i = supply at region i (people - number of beds available)
d_i = demand at region i (people in drug court)
f_j = fixed cost of facility at node j (10% per avg per capita income)
E_{ij} = Euclidean distance between region i and j (km per person)
P = number of facilities to locate
β = The number of beds in each rehab
ϵ_j = The number of days in rehab j

Several versions of the model were run. One version was run without using any fairness measures. We also looked at changing the distance parameter to either Euclidean, Haversine, Manhatten, or Chebyshev distance. Next, the Hoover index (*1) was used at values of 0.1, 0.2, and 0.3. Finally, the Gini index (*2) was used, also at the values of 0.1, 0.2, and 0.3. In the model that does not include the constraint, all facilities end up being built in one region, the one with the smallest Euclidean distance and smallest fixed cost.

In the model, Z_3 corresponds to priority 1, Z_2 corresponds to priority 2, and Z_1 corresponds to priority 3. The first constraint is the flow balance constraint. This allows for the number of people traveling in and out of the node to not be more than the supply minus the demand for that region. The second constraint ensures that the total number of facilities built y_j does not exceed the maximum number of facilities allowed to allocate P. P in this model is assumed to be the total demand minus the total supply divided by 30 beds, rounded up to the next highest integer. For this model, that number ends up being 38 as the maximum number of facilities allowed to locate.

Figures 8 and 9 show the Python code used to solve this model. Gurobi was the solver used. The `setObjectiveN()` function was used to prioritize the three objective functions in this model. The Euclidean distance function is used in the Z_3 objective function and its explanation can be seen on the following page (Maria et al., 2020).

$$P = \text{round}\left(\left|\frac{\left(\sum_{i=1}^{8} d_i - \sum_{j=1}^{8} s_j\right)}{\beta}\right|, 0\right) \tag{1}$$

$$\text{minimize } Z_1 = \sum_{j=1}^{8} \epsilon_j f_j Y_j \tag{2}$$

$$\text{minimize } Z_2 = \sum_{i=1}^{8} \sum_{j=1}^{8} E_{ij} X_{ij} \tag{3}$$

$$\text{maximize } Z_3 = \sum_{j=1}^{8} Y_j \tag{4}$$

subject to:

$$\sum_{(i,j)\in A} X_{ij} - \sum_{(k,i)\in A} X_{ki} \geq s_i - d_i \tag{5}$$

$$\sum_{j\in V} Y_j \leq P \tag{6}$$

$$\frac{1}{2n\bar{Y}} \sum_{i} \left|Y_i - \bar{Y}\right| \leq \{0.1, 0.2, 0.3\} \tag{*1}$$

$$\frac{1}{2n^2\bar{Y}} \sum_{i,j} |Y_i - Y_j| \leq \{0.1, 0.2, 0.3\} \tag{*2}$$

$$X_{ij}, Y_j \geq 0, \forall (i,j) \in A, \forall j \in V \tag{7}$$

$$\text{where } E_{ij} = \Delta\varphi^2 + \Delta\gamma^2 k$$

$$\Delta\varphi = \varphi_j - \varphi_i \text{ (latitude)}$$

$$\Delta\gamma = \gamma_j - \gamma_i \text{ (longitude)}$$

$$k = 111.319 \text{ (conversion to km)}$$

$$\text{where } \bar{Y} = \text{ mean of } Y$$

Note that we assume $\epsilon_j = 30$ for all rehabilitation centers, meaning each rehab has a 30-day stay. Additionally, we assume that each rehab has only 30 beds, i.e., $\beta = 30$.

5.7.3 The Haversine Parameter

$h_{ij} =$ Haversine distance between region i and region j (km per person)

The changes made to this model were in the second objective function in which the Haversine distance formula was used instead of the Euclidean distance formula. It can be seen in (3) as follows.

$$\text{minimize } Z_2 = \sum_{i=1}^{8} \sum_{j=1}^{8} h_{ij}^2 X_{ij} \tag{3}$$

$$\begin{aligned} \text{where } c &= 2 \cdot \operatorname{atan2}\left(\sqrt{a}, \sqrt{1-a}\right) \\ a &= \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_j) \cdot \cos(\varphi_i) \cdot \sin^2\left(\frac{\Delta\gamma}{2}\right) \\ \Delta\varphi &= \varphi_j - \varphi_i \text{ (latitude)} \\ \Delta\gamma &= \gamma_j - \gamma_i \text{ (longitude)} \\ r &= 6377 \text{ km (Earth's equatorial radius)} \end{aligned} \tag{9}$$

5.7.4 The Manhattan Parameter

M_{ij} = Manhattan distance between region i and region j (km per person)

The changes made to this model were again made in the second objective function in which the Manhattan distance formula was used instead of the Euclidean distance formula. It can be seen in (3) as follows.

$$\text{minimize } Z_2 = \sum_{i=1}^{8} \sum_{j=1}^{8} k M_{ij} X_{ij} \tag{3}$$

$$\begin{aligned} \text{where } M_{ij} &= \Delta\varphi + \Delta\gamma \\ \Delta\varphi &= \varphi_j - \varphi_i \text{ (latitude)} \\ \Delta\gamma &= \gamma_j - \gamma_i \text{ (longitude)} \\ k &= 111.319 \text{ (conversion to km)} \end{aligned}$$

5.7.5 The Chebyshev Parameter

C_{ij} = Chebyshev distance between region i and region j (km per person)

Finally, the changes made to this model were in the second objective function again in which the Chebyshev distance formula was used instead of the Euclidean distance formula. It can be seen in (3) as follows.

$$\text{minimize } Z_2 = \sum_{i=1}^{8} \sum_{j=1}^{8} kC_{ij}X_{ij} \tag{3}$$

$$\begin{aligned} \text{where } C_{ij} &= \max(\Delta\varphi, \Delta\gamma) \\ \Delta\varphi &= \varphi_j - \varphi_i \text{ (latitude)} \\ \Delta\gamma &= \gamma_j - \gamma_i \text{ (longitude)} \\ k &= 111.319 \text{ (conversion to km)} \end{aligned}$$

5.8 Fairness Measures

Two fairness measures were used in this analysis. First, the Hoover index was used. This corresponds to constraint (∗1) above. This is also known as the Robin Hood index. It is the fractional total utility to be redistributed to achieve perfect equality (Chen and Hooker, 2023). The Hoover index or Robin Hood index can be seen in Figure 5.5 as the maximum vertical distance between the Lorenz curve and the 45-degree line (Kennedy et al., 1996).

The Gini coefficient is similar in its relation to the Lorenz curve, being proportional to the area between the Lorenz curve and the diagonal line and the total area between that and the area below the 45-degree line, represented as Gini $= \frac{A}{A+B}$ as shown in Figure 5.6 (De Maio, 2007). These Lorenz curves can be seen in the following figures. For both, a value of 0 is considered completely equal and a value of 1 is considered completely unequal. Both were tried at the values of 0.1, 0.2, and 0.3.

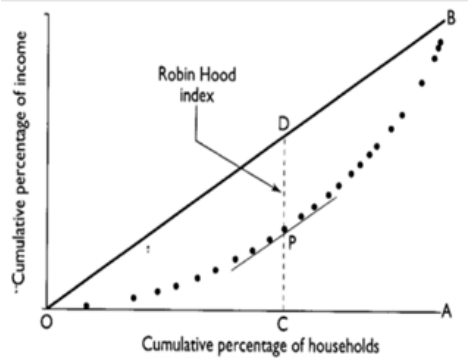


Figure 5.5: Hoover Index

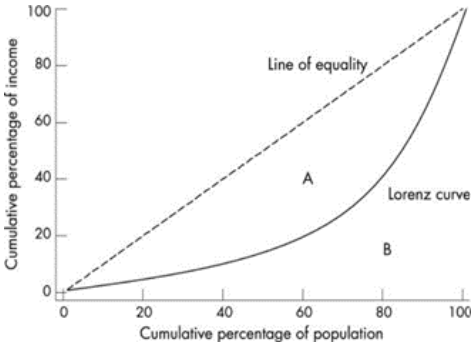


Figure 5.6: Gini Coefficient

Auxiliary variables were created for both the Hoover index and Gini index. This was done to linearize the constraints for use of absolute values. The explanation of how to linearize these constraints can be seen on the next page and the code can be seen in the following figures.

Let $y[i]$ be the number of facilities at node i and n be the total number of nodes.

$$\bar{y} = \frac{1}{n} \sum_{j \in \text{nodes}} y[j]$$

Define $u[i]$ as the auxiliary variable for each node i and $u[i,j]$ for each pair of nodes i and j.

The linearization of the absolute value constraints for the Hoover Index can be expressed as follows:

$$u[i] \geq y[i] - \bar{y} \quad \forall i \in \text{nodes} \tag{10}$$
$$u[i] \geq \bar{y} - y[i] \quad \forall i \in \text{nodes} \tag{11}$$

The linearization of the absolute value constraints for the Gini index can be expressed as follows:

$$u[i,j] \geq y[i] - y[j] \quad \forall i,j \in \text{nodes}, i \neq j \tag{12}$$
$$u[i,j] \geq y[j] - y[i] \quad \forall i,j \in \text{nodes}, i \neq j \tag{13}$$

Where:

- $u[i]$ is the auxiliary variable representing the absolute value of the difference between $y[i]$ and the average number of facilities $\bar{y}$.

- $u[i, j]$ is the auxiliary variable representing the absolute value of the difference between $y[i]$ and $y[j]$.
- $y[i]$ is the number of facilities at node i.
- $y[j]$ is the number of facilities at node j.
- $\bar{y}$ is the average number of facilities.

These constraints ensure that the auxiliary variables are always greater than or equal to the absolute value of the respective differences.

```
### FAIRNESS MEASURES
n = len(nodes)
avg_facilities = quicksum(y[j] for j in nodes) / n

# Create auxiliary variables
u = m.addVars(nodes, vtype=GRB.CONTINUOUS, name='u')

# Linearize constraints for absolute value
m.addConstrs((u[i] >= y[i] - avg_facilities for i in nodes), name='abs_lower')
m.addConstrs((u[i] >= avg_facilities - y[i] for i in nodes), name='abs_upper')

# Calculate Hoover index
# Calculate Hoover index
hoover_sum = quicksum(u[i] for i in nodes)
m.addConstr(hoover_sum <= 0.3 * 2 * n * avg_facilities, name='hoover_index')

# Set Objective and Priorities
z1 = quicksum(30 * fixed_cost[j] * y[j] for j in nodes)
z3 = quicksum((haversine(lat[i], lon[i], lat[j], lon[j]) ** 2) * x[i, j] for (i, j) in arc)
z2 = quicksum(y[j] for j in nodes)
m.update()
m.NumObj = 3
m.setObjectiveN(z1, GRB.MINIMIZE, priority=0)
m.setObjectiveN(z3, GRB.MINIMIZE, priority=1)
m.setObjectiveN(-z2, GRB.MINIMIZE, priority=2)
```

Figure 5.7: Hoover Index Code

```
n = len(nodes)
avg_facilities = quicksum(y[j] for j in nodes) / n

# Create auxiliary variables
u = m.addVars(nodes, vtype=GRB.CONTINUOUS, name='u')

# Linearize constraints for absolute value
m.addConstrs((u[i] >= y[i] - y[j] for i in nodes for j in nodes if i != j), name='gini_lower')
m.addConstrs((u[i] >= y[j] - y[i] for i in nodes for j in nodes if i != j), name='gini_upper')

# Calculate Hoover index
gini_sum = quicksum(u[i] for i in nodes)
m.addConstr(gini_sum <= 0.3 *2* n ** 2 * avg_facilities, name='gini_index')

# Set Objective and Priorities
z1 = quicksum(30 * fixed_cost[j] * y[j] for j in nodes)
z3 = quicksum((haversine(lat[i], lon[i], lat[j], lon[j]) ** 2) * x[i, j] for (i, j) in arc)
z2 = quicksum(y[i] for i in nodes)
```

Figure 5.8: Gini Coefficient Code

5.9 Results

First, we ran the model without using any of the fairness constraints. Figure 5.9 shows the model's solution with $x[\text{node}_i, \text{node}_j]$ being the number of people traveling on arc (i, j) and $y[\text{node}_j]$ being the number of facilities located in region j. Notice that the flow for all models is the same for all four metrics without using fairness, which in the first model places all facilities in the center-most lowest cost region, region 6, which corresponds to Pottawatomi County.

Figure 5.10 shows the total cost associated with placing the facilities at that node as well as how many facilities were placed in total.

Variable	X
x[node6,node8]	79
x[node2,node7]	847
x[node1,node6]	169
x[node3,node5]	162
x[node6,node2]	1155
x[node7,node4]	436
x[node3,node1]	64
y[node6]	38

Figure 5.9: Flow No Fairness

Notice the total cost to build these facilities was roughly 2.55 million dollars with a total distance traveled of 406,745 km for the Euclidean, 373,191 km

for the Haversine, 538,108 for the Manhattan, and 346,543 for the Chebyshev model. All these correspond to the models without fairness measures applied. Figures 5.11, 5.12, 5.13, 5.14 shows the flow for all four metrics with no fairness measures implemented. The placement of facilities and total cost was the same. The difference was in the distance for each metric. The resulting outputs are shown with the distances for each metric and the number of facilities placed at each region and total cost based on 10 percent per capita income for each region in the following figures. Detailed information is provided in Table 5.1.

Variable	X
x[node6,node8]	79
x[node2,node7]	411
x[node2,node4]	436
x[node1,node2]	1155
x[node6,node3]	824
x[node3,node1]	1050
x[node6,node5]	162

Figure 5.10: Flow of all 4 Metrics with Fairness Added

```
Facilities:
node1 -0.0 -0.0
node2 -0.0 -0.0
node3 -0.0 -0.0
node4 -0.0 -0.0
node5 -0.0 -0.0
node6 38.0 2553600.0
node7 -0.0 -0.0
node8 -0.0 -0.0
38.0
2553600.0
Total Distance Traveled: 406745.37
```

Figure 5.11: Total Cost Euclidean No Fairnes

```
Facilities:
node1 -0.0 -0.0
node2 -0.0 -0.0
node3 -0.0 -0.0
node4 -0.0 -0.0
node5 -0.0 -0.0
node6 38.0 2553600.0
node7 -0.0 -0.0
node8 -0.0 -0.0
38.0
2553600.0
Total Distance Traveled: 373191.06
```

Figure 5.12: Total Cost Haversine No Fairnes

```
Facilities:
node1 -0.0 -0.0
node2 -0.0 -0.0
node3 -0.0 -0.0
node4 -0.0 -0.0
node5 -0.0 -0.0
node6 38.0 2553600.0
node7 -0.0 -0.0
node8 -0.0 -0.0
38.0
2553600.0
Total Distance Traveled: 538108.85
```

Figure 5.13: Total Cost Manhattan No Fairnes

```
Facilities:
node1 -0.0 -0.0
node2 -0.0 -0.0
node3 -0.0 -0.0
node4 -0.0 -0.0
node5 -0.0 -0.0
node6 38.0 2553600.0
node7 -0.0 -0.0
node8 -0.0 -0.0
38.0
2553600.0
Total Distance Traveled: 346543.2
```

Figure 5.14: Total Cost Chebyshev No Fairness

Next, thresholds of .1, .2, and .3 were tried for both the Hoover and Gini index. This is done for all four distance metrics and results are recorded in Table 5.1. All four metrics led to the same results except when it came to distance and can be seen in Figues 5.15 - 5.38.

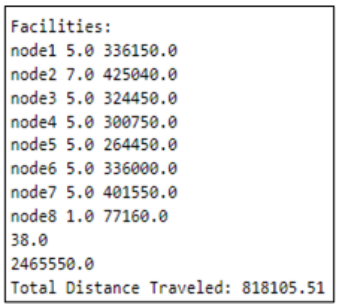


Figure 5.15: Euclidean Hoover .1

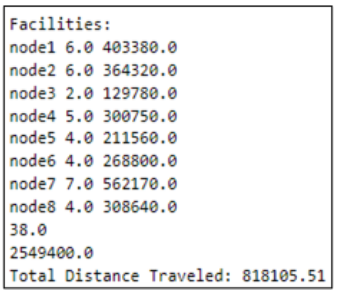


Figure 5.16: Euclidean Hoover .2

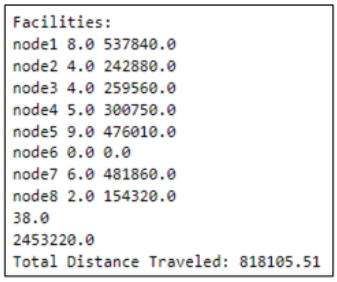


Figure 5.17: Euclidean Hoover .3

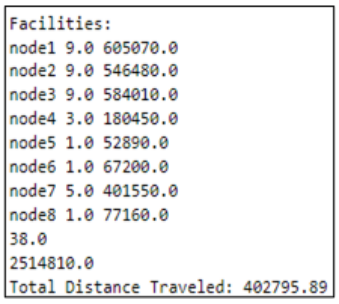


Figure 5.18: Euclidean Gini .1

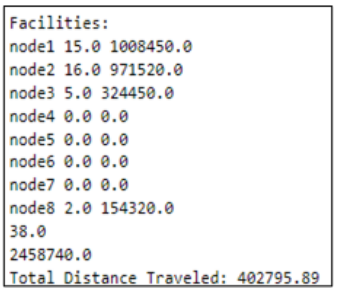


Figure 5.19: Euclidean Gini .2

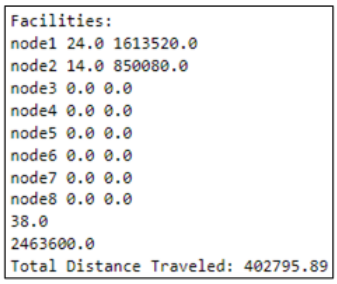


Figure 5.20: Euclidean Gini .3

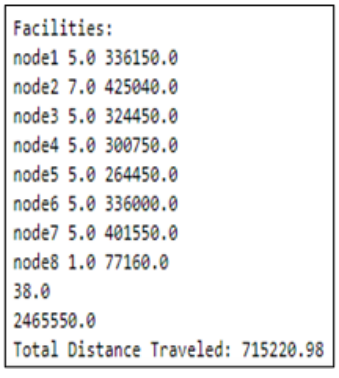


Figure 5.21: Haversine Hoover .1

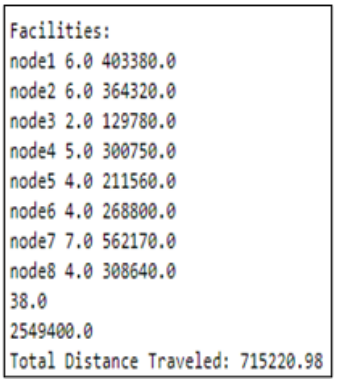


Figure 5.22: Haversine Hoover .2

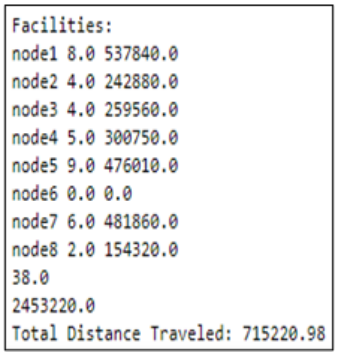


Figure 5.23: Haversine Hoover .3

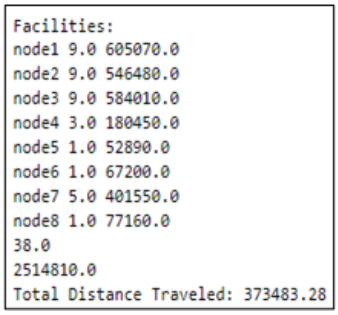


Figure 5.24: Haversine Gini .1

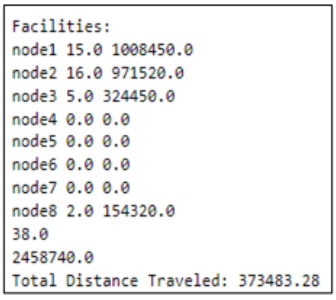


Figure 5.25: Haversine Gini .2

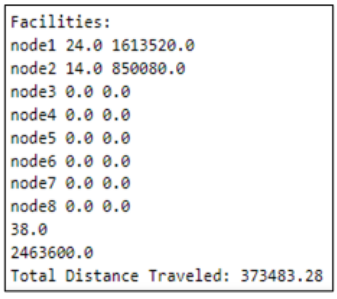


Figure 5.26: Haversine Gini .3

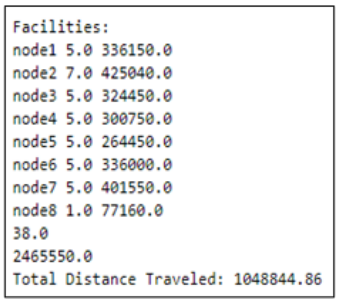


Figure 5.27: Manhattan Hoover .1

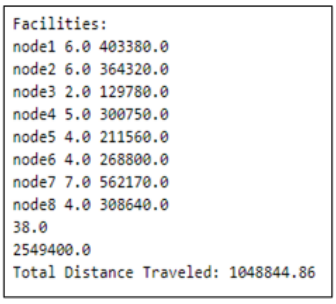


Figure 5.28: Manhattan Hoover .2

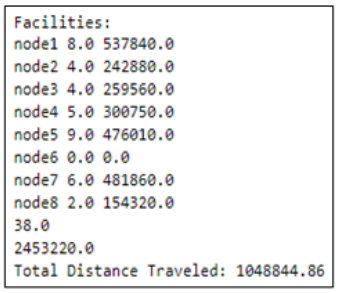


Figure 5.29: Manhattan Hoover .3

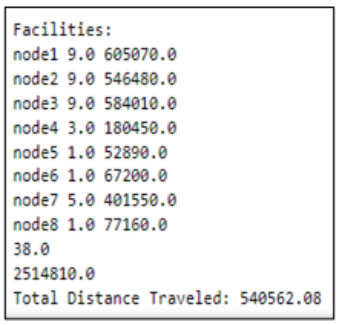


Figure 5.30: Manhattan Gini .1

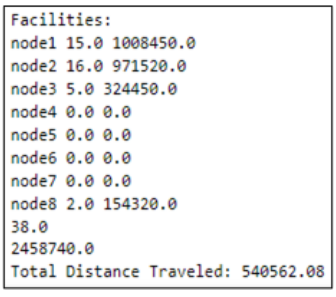


Figure 5.31: Manhattan Gini .2

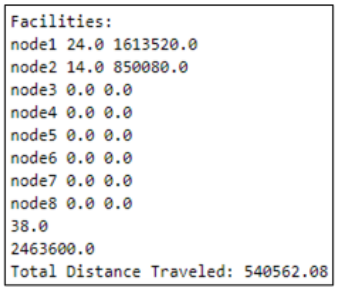


Figure 5.32: Manhattan Gini .3

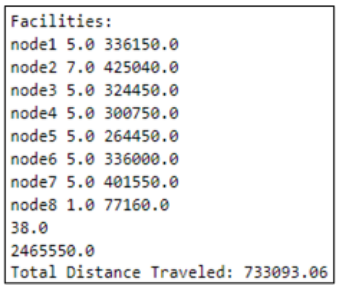


Figure 5.33: Chebyshev Hoover .1

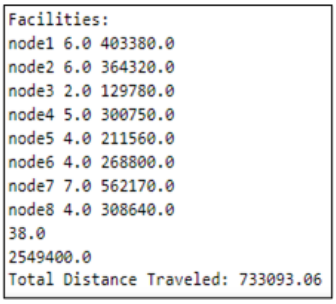


Figure 5.34: Chebyshev Hoover .2

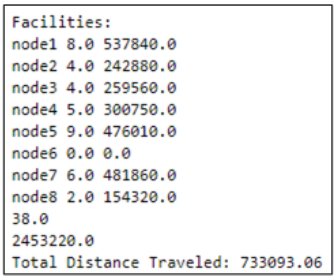


Figure 5.35: Chebyshev Hoover .3

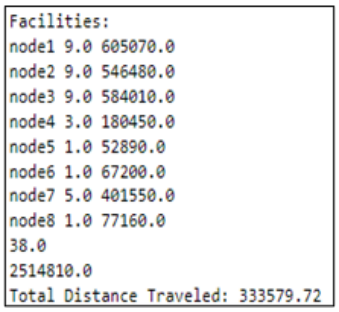


Figure 5.36: Chebyshev Gini .1

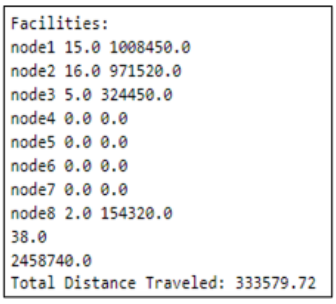


Figure 5.37: Chebyshev Gini .2

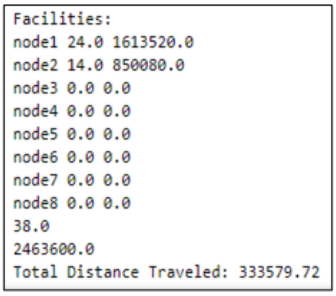


Figure 5.38: Chebyshev Gini .3

Table 5.1: Comparison of different methods and their performance metrics

Methods	Threshold	Facilities	# Regions w Facilities	Total Cost (#Facility*10% Per Capita Income by Region)	Total Distance (km) Euclidean	Total Distance (km) Haversine	Total Distance (km) Manhattan	Total Distance (km) Chebyshev
Hoover	0.3	38	7	$2,453,220	818,105	715,220	108,844	733,093
Gini	0.2	38	4	$2,458,740	402,795	373,483	540,562	333,579
Gini	0.3	38	2	$2,463,600	402,795	373,483	540,562	333,579
Hoover	0.1	38	8	$2,465,550	818,105	715,220	108,844	733,093
Gini	0.1	38	8	$2,514,810	402,795	373,483	540,562	333,579
Hoover	0.2	38	8	$2,549,400	818,105	715,220	108,844	733,093
No Fairness Measure	None	38	1	$2,553,600	406,745	373,191	538,108	346,543

5.10 Conclusions

A Facility Location Model with Capacity Covering Constraints was used with multiple objective functions that were prioritized to minimize cost while helping the most people by placing an optimal number of treatment centers in each region to meet demand while choosing one of four distance metrics (Haversine, Euclidean, Manhattan, or Chebyshev) minimizing the distance to the treatment center from the starting location. In the first scenario that does not look at a fairness constraint, all facilities were located in Region 6, Potawatomi County, in all four distance models. This is the cheapest of the two center most regions. The total cost in the first scenario was $2.55 million, however having all facilities located in one region is not helpful to patients throughout the state. The difference between these models was in the total distance traveled; for this particular model, Chebyshev traveled the shortest distance overall with 346,543 km, Haversine was next with 373,191 km, then Euclidean with 406,745 km, and lastly Chebyshev with 538,108 km. The distance delta between the best model and the worst model is 191,565 km. The overall optimal model when it comes to fairness is the Hoover index of 0.3 with a total cost of $2.45 million with a Manhattan distance. This

model had a total distance traveled of 108,844 km, Haversine had 715,220 km, Chebyshev had 733,093 km, and Euclidean had 818,105 km. For the best Gini index model, model 0.2 was the best threshold with a total cost of $2.46 million, Chebyshev having the best distance of 333,579 km, Haversine coming in second with 373,483 km, Euclidean next with 402,795 km, and Manhattan last with 540,562 km. This model had a distance delta of 206,983 km. Compared to not using fairness at all, we can save a total cost of $100,380 by using a Hoover index fairness of ≤ 0.3 with 7 regions having facilities and a total cost savings of $94,860 by implementing a Gini index of ≤ 0.2 and getting a spread of 4 regions having facilities. Overall, facilities are spread out more evenly using the Hoover index than the Gini Index when coupled with any of the four distance metrics. Taking the cost, number of facilities located in the regions overall, and total distance traveled into consideration, we believe that using a threshold of 0.3 and Hoover index fairness with the Manhattan distance is the best model when looking at distance. When looking at all the optimal models, without fairness, the model leans toward what happens along the Oklahoma City/Tulsa corridor, where interstates and turnpikes make long diagonal trips feel quicker and easier. When we add fairness, rural and under served counties matter more, and travel there is mostly 'right angle' driving on sectional roads with few diagonal shortcuts. In that setting, a simple grid like measure fits the reality better and reduces those really long trips. In short: Metro corridors make diagonals feel natural; equity shifts attention to places where people actually drive on grids. A central takeaway is that equity can improve efficiency. By requiring more even access across counties, the model relocates capacity to places that eliminate the longest trips. The result is lower total travel cost and a smaller tail of extreme distances. This runs counter to the idea that fairness necessarily raises costs. In our setting, it rebalanced the network away from over served metro corridors toward rural demand, where each mile saved matters more. Practically, planners can adopt fairness targets with confidence that they may save money and improve access at the same time.

Chapter 6

A Treatment Focused Compartmental Model Approach to Drug Offense Mitigation in Oklahoma

6.1 Introduction

The United States has some of the highest recidivism rates in the world with almost 44% of the criminal released returning to prison within the first year out of prison (rec), 68% of drug offenders are rearrested within 3 years of release from prison (Belenko et al., 2013a), and Oklahoma itself ranks 3rd if looking at the worlds incarceration rate considering every state as a country (Wildra and Herring, 2021b). In 2018, Oklahoma's recidivism rate has improved; however, it still stands at 24 8%, reflecting continued challenges to break the cycle of re-offending among released individuals (World Population Review, 2025).The recidivism rate in Oklahoma greatly exceeds the national average for drug offenders, resulting in prisons being at or overcapacity resulting in higher cost to taxpayers. This results in a shortage of available beds in the existing treatment centers with the increased number of addicts/alcoholics. The majority of people who re-offend do so because they are not given the opportunity to go to treatment centers instead of being incarcerated. Many rehabilitation centers require insurance or financial assistance not available to a large percentage of addicts. There are crimes that most likely would not have occurred had an offender not been suffering from an addiction; these co-occurring crimes—such as theft committed to fund drug use or assault occurring under the influence—are secondary offenses that are closely linked to the underlying substance abuse. More accessibility to treatment centers may help not only decrease re-offending, but could

also decrease the occurrence of drug related or co-occurring crimes. By investigating a treatment center model approach, the hope is to show a decrease in incarceration over time for drug crime offenses in comparison to a punishment focused system that aims to incarcerate or put offenders on probation instead of or before sending them to rehabilitation.

As mentioned in the previous chapter, this chapter builds on the idea of compartmental models in computational criminology and adds two aspects, one is a poverty component and the other is the aspect of fairness, i.e. the Gini index which in this case goes right along with poverty for each county.

6.2 Research Questions

- **Would shifting Oklahoma's drug offense mitigation from a punishment-first system to a treatment-first system reduce incarceration rates?**
  - Would utilizing treatment before incarceration or probation make a difference?
- **Would a treatment-focused approach lower public costs compared to the current system?**
  - Would this in turn lead to less people incarcerated which would save the tax payers dollars?
- **Which system parameters most influence costs and outcomes of incarceration?**
  - Using sensitivity analysis and the sobol index, can we determine which parameters influence these financial out incarceration outcomes?
- **Do results hold across different county sizes (large vs. small counties in Oklahoma)?**
  - Does a big county like Oklahoma or Tulsa county have similar results to a small county like Beckham county?
- **Can we identify what parameters contribute most to incarceration flow?**
  - Using Sobol in Python we can look at unknown parameters instead of using a coin flip and float between values to see how different values affect the incarceration flow.

Glossary

- **Deferred Sentence** – A judicial disposition where sentencing is postponed pending successful completion of probation, potentially resulting in dismissal of charges. (bla, 2019)
- **Suspended Sentence** – A judgment imposed but not executed, allowing the defendant to serve probation; violation may result in imposition of the original sentence. (Ame, 2017)
- **Treatment** – An inpatient drug and alcohol rehabilitation program.
- **Incarceration** – Confinement in prison (excluding jail detention) as penal sanction.
- **Recidivism** – Reversion to criminal behavior measured by return to prison within a specified period. (of Justice Statistics, 2018)

6.3 Assumptions

For the purpose of this research, drug court data for the state of Oklahoma was used to obtain the rates the flow between each compartment of the models. Both alcohol and drug offenders were lumped together in this study as both were members of drug court and both attend 12 step meetings at an equal rate. The rates that were not available in the drug court data were set to a value of .50 or 50%. The boundary of this problem is limited to just the state of Oklahoma including only the counties that have a drug court program available. The data is based off a report from fiscal year 2002 to fiscal year 2005. The birth and death rates along with the population numbers are from the year 2005. We assumed that no person would go to a rehab more than once in a year; however, relapse rates were considered.

6.4 Background

6.4.1 Current State

Compartmental models show how individuals within a closed population are separated into mutually exclusive groups, or compartments, based on their disease status. Each individual is considered to be in 1 compartment at a given time, but can move from one compartment to another based on the parameters of the model. (Galea et al., 2020) There have been several

applications of compartmental models similar to a SIR model being used to model behavior of various systems. One of the models that was researched was the SEIR model and different variations in relation to the spread of an epidemic. A simple model explained was one containing Susceptible, Infected, Exposed, and Recovered and how the flow of people is throughout each compartment over time. The compartments being S, E, I, and R. The simple model and its system of differential equations can be seen in Figure 6.1.

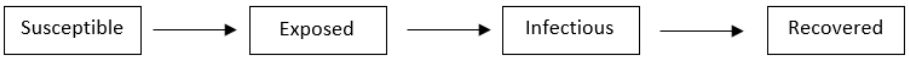


Figure 6.1: Basic SEIR Model

Another article looks at how adding a reform program to the model lowers the recidivism rate. The compartmental model can be seen in Figure 6.2. There are classes or compartments shown in the model. Those 8 classes and their descriptions can be seen in Figure 6.3. Finally, the rates are listed along the arcs of the model. These include starting population, per capita prison and incarceration rates, proportions of released individuals that were involved in or completed the inside out program, re-incarceration rates for those that either complete or don't complete the program, a social influence parameter, and the proportion of unsuccessful rate of the reentry program, and finally proportion of people that go back to prison. They also show how modifying a model can drastically simplify it. These can be seen in Figure 6.4. (Alvarez et al., 2012)

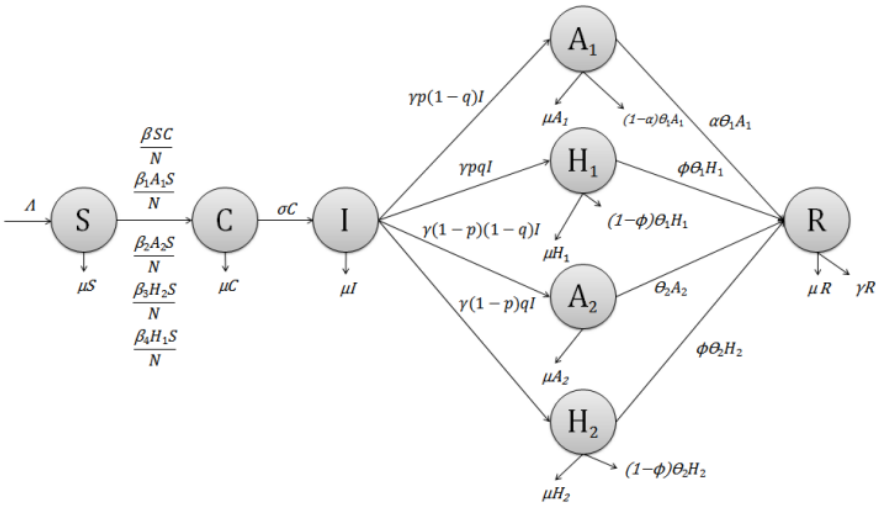


Figure 6.2: General Model Dynamics of Inmates Program Completion

Class	Description
S	Population that is susceptible to becoming criminals
C	Individuals that commit crimes and are not imprisoned
I	Imprisoned criminal population
A_1	Released inmates that completed the inside program, but did not complete or were not involved in outside program
A_2	Released inmates that did not complete or were not involved the inside program within prison and also did not complete or were not involved in the outside program
H_1	Released inmates that completed the inside program, and completed or were involved in the outside program (control group)
H_2	Released inmates that did not complete or were not involved in the inside program, but completed or were involved the outside program
R	Released inmates that go back to prison a second time irrespectively of their past experience with the reform programs

Figure 6.3: Eight Classes and Their Descriptions

Parameters	Description
N	Starting population
Λ	Rate of entry of individuals into core-group population as susceptibles
σ	Per capita incarceration rate
γ	Per capita prison release rate
μ	Per capita mortality rate
q	Proportion of released individuals that were involved in outside program
p	Proportion of released individuals that completed the inside program
θ_1	Per capita reincarceration rate for those who completed the inside program
θ_2	Per capita reincarceration rate for not completing the inside program
β	Social influence parameter related to individuals in the C class
$\beta_i = \epsilon_i \beta$	Social influence parameter related to individuals in the A_1, A_2, H_1, H_2 class where ϵ_i is the reduction in the social influence of A_1, A_2, H_1, H_2 class individuals as compared to C class individuals
ϕ	Proportion of unsuccess rate of the reentry program
α	Proportion of individuals that go back to prison

Figure 6.4: Parameters Inmate Program Model

Next, an article that was researched looks at the economics of prison in particular modeling the dynamics or recidivism. This model looks at prison population both with and without an education component. In this case, that is obtaining a GED. The flow diagram, variables used, and table of parameters can be seen in Figures 6.5, 6.6, and 6.7 respectively. (Lopez et al., 2018)

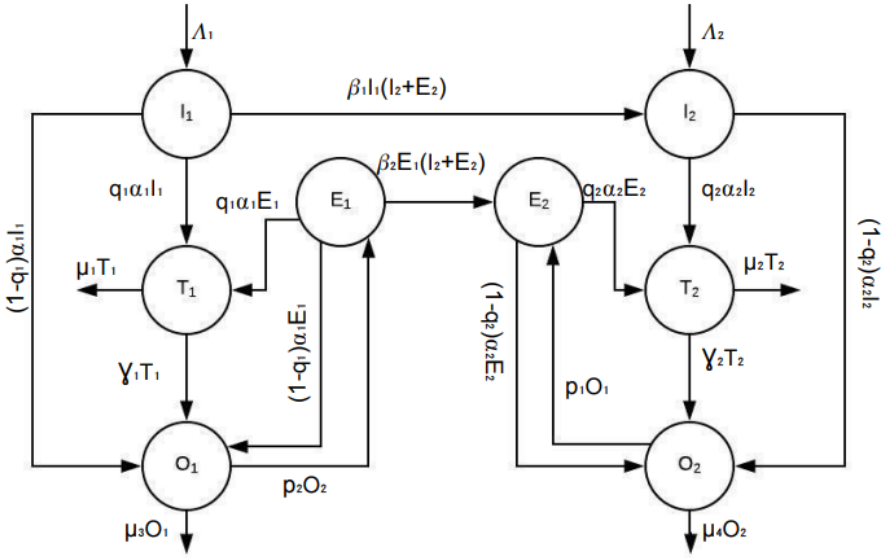


Figure 6.5: Prison Dynamics with Eduction Model

Class	Description
I_1	Inmates with no GED
I_2	Inmates with GED or more
E_1	Returning offenders without GED
E_2	Returning offenders with GED or more
T_1	Released inmates without GED attending the transition program
T_2	Released inmates with GED attending the transition program
O_1	Inmates without GED who are still at risk to recidivate
O_2	Inmates with GED who are still at risk to recidivate

Figure 6.6: Variables with Education Model

Parameters	Description	Units
Λ_1	Incoming population without GED	People per time
Λ_2	Incoming population with GED or more	People per time
β_1	Rate of influence between I_2 and I_1	Per person per time
β_2	Rate of influence between E_2 and E_1	Per person per time
p_1	Rate of people leaving O_1	1/time
p_2	Rate of people leaving O_2	1/time
q_1	Proportion of released inmates without GED who go to the transition program	Dimensionless
q_2	Proportion of released inmates with GED or more who go to the transition program	Dimensionless
μ_1	Per capita rehabilitation rate from T_1	1/time
μ_2	Per capita rehabilitation rate from T_2	1/time
μ_3	Per capita rehabilitation rate from O_1	1/time
μ_4	Per capita rehabilitation rate from O_2	1/time
γ_1	Rate of people leaving T_1	1/time
γ_2	Rate of people leaving T_2	1/time
α_1	Rate of people leaving I_1	1/time
α_2	Rate of people leaving I_2	1/time

Figure 6.7: Parameters Education Model

Finally, the last article represents several models that increase in complexity while looking at criminal activity and incarceration as a disease spread model similar to how diseases such as influenza are modeled in epidemiology. Using a three-dimensional model with compartments such as: not criminally active (X), criminally active (C), and incarcerated (I). The five-dimensional models capture the process of prisoner reentry (or returning from jail) by introducing the distinction between states C1 and C2 (based on whether one has been incarcerated before) and introducing state R to represent the state one is in immediately after release from prison/jail. (McMillon et al., 2014). These diagrams can be seen in Figures 6.8 and 6.9, respectively.

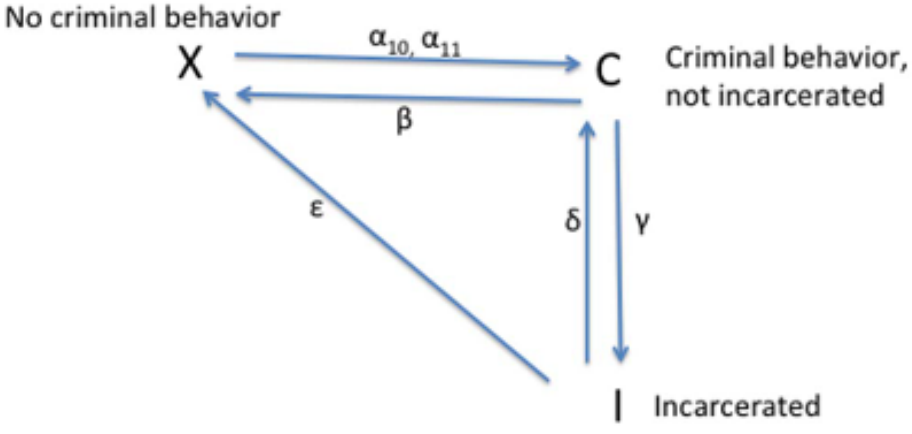


Figure 6.8: 3D Model Crime

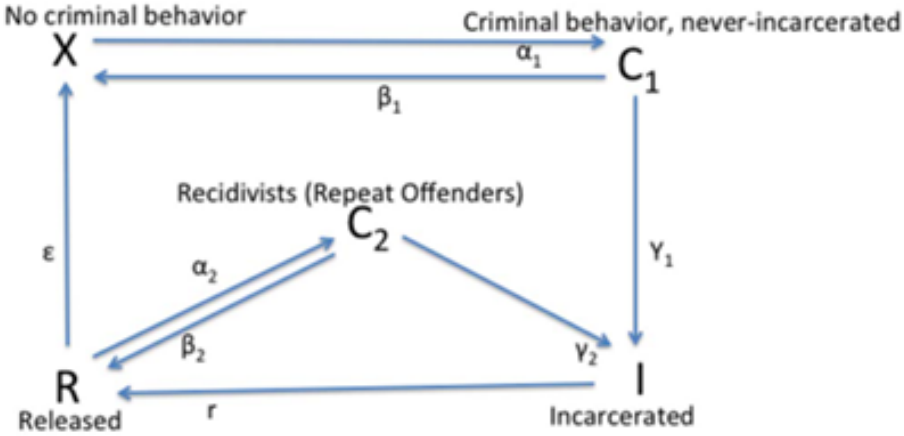


Figure 6.9: 5D Model Crime

6.5 Methodology

Based on all these examples, a very simple model was drawn out with only 4 compartments. These were free, probation, incarceration, and treatment. Probation was then broken out to include both deferred and suspended sentences. The flows were entered based on available statistics in the drug court data report. The initial model (Figure 6.10), the more complex model with rate parameters (Figure 6.11), the parameters table with descriptions (Table 6.1), and the system of differential equations (Equation 6.1) captures flows between free, deferred, suspended, treatment, and incarcerated populations. This is the current probation/punishment-based system.

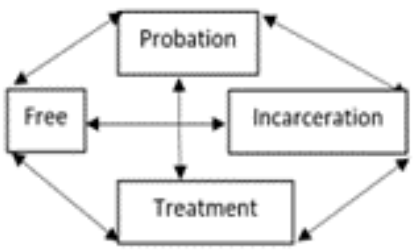


Figure 6.10: Simple Model

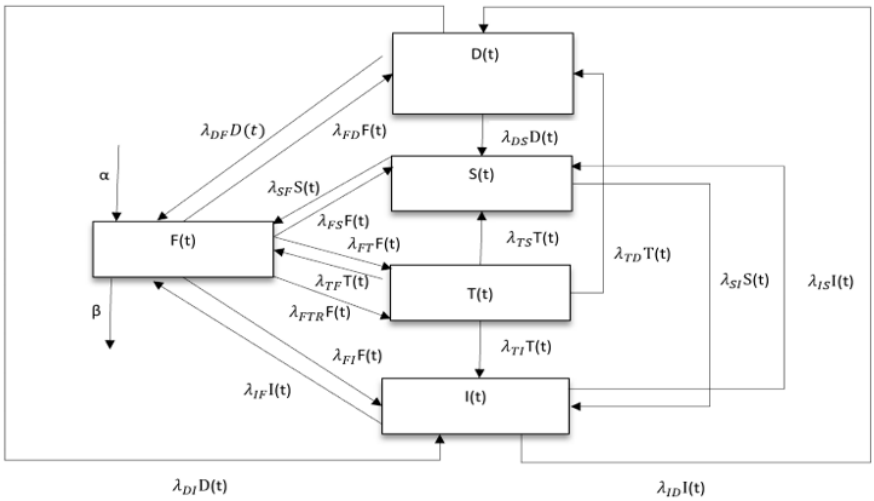


Figure 6.11: More Complex Model with Parameters Added

$$
\begin{aligned}
\frac{dF}{dt} &= \alpha + \lambda_{DF}D(t) + \lambda_{IF}I(t) + \lambda_{TF}T(t) + \lambda_{SF}S(t) - \beta - \lambda_{FI}F(t) - \lambda_{FTR}F(t) - \lambda_{FT}F(t) - \lambda_{FS}F(t) - \lambda_{FD}F(t) \\
\frac{dD}{dt} &= \lambda_{FD}F(t) + \lambda_{TD}T(t) + \lambda_{ID}I(t) - \lambda_{DF}D(t) - \lambda_{DI}D(t) - \lambda_{DS}D(t) \\
\frac{dS}{dt} &= \lambda_{DS}D(t) + \lambda_{FS}F(t) + \lambda_{TS}T(t) + \lambda_{IS}I(t) - \lambda_{SF}S(t) - \lambda_{ST}S(t) \\
\frac{dT}{dt} &= \lambda_{FT}F(t) + \lambda_{FTR}F(t) - \lambda_{TF}T(t) - \lambda_{TI}T(t) - \lambda_{TS}T(t) - \lambda_{TD}T(t) \\
\frac{dI}{dt} &= \lambda_{DI}D(t) + \lambda_{FI}F(t) + \lambda_{TI}T(t) + \lambda_{SI}S(t) - \lambda_{IF}I(t) - \lambda_{ID}I(t) - \lambda_{IS}I(t)
\end{aligned}
\tag{6.1}
$$

Parameter	Description	Source	Rate Oklahoma State	Rate Oklahoma County	Rate Beckham County
α	Birth Rate Oklahoma 2005	Oklahoma.gov (okl, 2008)	14.6%	17.2%	16.1%
β	Death Rate Oklahoma 2005	Oklahoma.gov (okl, 2008)	10.2%	9.34%	11.96%
λ_{FD}	Free to Deferred Rate (Sentence Type – Deferred)	Drug Court Report (odm, 2005)	17%	2.3%	40.9%
λ_{FS}	Free to Suspended Rate (Sentence Type – Suspended)	Drug Court Report (odm, 2005)	8.7%	0.6%	0.0%
λ_{FT}	Free to Treatment Rate (Felony Drug Charges)	Drug Court Report (odm, 2005)	54%	80.8%	99.9%
λ_{FI}	Free to Incarcerated Rate	Drug Court Report, Google.com	$n_0 \cdot \frac{993}{N_0}$	same	same
λ_{DS}	Deferred to Suspended Rate	Assumed	50%	50%	50%
λ_{TS}	Treatment to Suspended Rate	Drug Court Report (odm, 2005)	20.3%	0.2%	39.1%
λ_{FTR}	Free to Treatment Relapse Rate	Google.com	50%	50%	50%
λ_{ID}	Incarcerated to Deferred Rate	Drug Court Report (odm, 2005)	8.8%	0.4%	4.3%
λ_{IS}	Incarcerated to Suspended Rate	Drug Court Report (odm, 2005)	20.3%	0.2%	39.1%
λ_{IF}	Incarcerated to Free Rate (Charges Dismissed)	Drug Court Report (odm, 2005)	58.1%	90.8%	26.1%
λ_{DF}	Deferred to Free Rate	Assumed	50%	50%	50%
λ_{SF}	Suspended to Free Rate	Assumed	50%	50%	50%
λ_{DI}	Deferred to Incarcerated Rate	Assumed	50%	50%	50%
λ_{TD}	Treatment to Deferred Rate	Assumed	50%	50%	50%
λ_{TI}	Treatment to Incarcerated Rate	Assumed	50%	50%	50%
λ_{TF}	Treatment to Free Rate (Completion Rate)	Drug Court Report (odm, 2005)	64.9%	64.9%	64.9%
λ_{SI}	Suspended to Incarcerated Rate	Assumed	50%	50%	50%
N_0	Initial Population Size (2005)	Google.com	3,543,442	684,192	18,810
n_0	Number Entering Drug Court	Drug Court Report (odm, 2005)	3532	559	23

Table 6.1: Data Rates and Sources for State of Oklahoma and Selected Counties. NOTE: Rates from 2005 may not reflect current trends.

6.5.1 Future State

By comparing the compartmental model simulation results of the current state model that is more complex and probation/prison centered to the proposed model that is more treatment focused, the hope is to see a decrease in the number of people in prison over time. A proposed treatment focused compartmental model can be seen in Figure 6.12 and Equations 6.2 respectively. The rates and data for these models were found at various locations. The rates that are still in the model are the same as in the previous model. The flows from compartments is what changes from the previous model to the proposed model. As you can see, this model looks at sending the offender to treatment first. You can only enter incarceration from the treatment or suspended compartments. This means if you fail treatment, you can go to either suspended, deferred, or incarcerated. From Deferred you can only go to free or suspended, and from suspended you can only go to free or incarcerated.

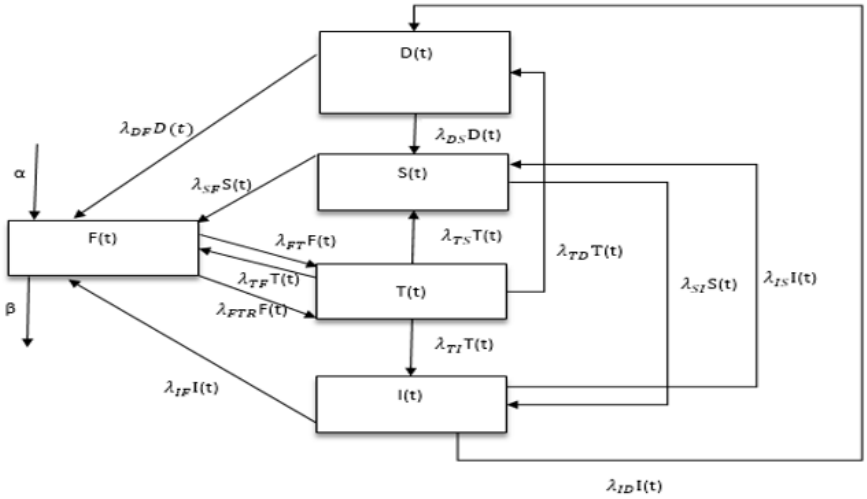


Figure 6.12: Treatment Focused Model

$F(t)$ = Free population
$D(t)$ = Deferred population
$S(t)$ = Suspended population
$T(t)$ = Treatment population
$I(t)$ = Incarcerated population

$$\frac{dF}{dt} = \alpha + \lambda_{DF}D(t) + \lambda_{IF}I(t) + \lambda_{TF}T(t) + \lambda_{SF}S(t) - \beta - \lambda_{FTR}F(t) - \lambda_{FT}F(t) \tag{6.2a}$$

$$\frac{dD}{dt} = \lambda_{TD}T(t) + \lambda_{ID}I(t) - \lambda_{DF}D(t) - \lambda_{DS}D(t) \tag{6.2b}$$

$$\frac{dS}{dt} = \lambda_{DS}D(t) + \lambda_{TS}T(t) + \lambda_{IS}I(t) - \lambda_{SF}S(t) - \lambda_{ST}S(t) \tag{6.2c}$$

$$\frac{dT}{dt} = \lambda_{TF}F(t) + \lambda_{FTR}F(t) - \lambda_{TF}T(t) - \lambda_{TS}T(t) - \lambda_{TI}T(t) - \lambda_{TD}T(t) \tag{6.2d}$$

$$\frac{dI}{dt} = \lambda_{TI}T(t) + \lambda_{SI}S(t) - \lambda_{IF}I(t) - \lambda_{ID}I(t) - \lambda_{IS}I(t) \tag{6.2e}$$

6.6 Analysis

A simulation model was run in Vensim software to compare the current state to the future state by breaking the output down into dollars. Vensim is an industrial-grade system dynamics simulation software designed to enhance the performance of real-world systems. It supports the development of robust and reliable models, ranging from causal loop diagrams to fully tested simulations, while integrating both qualitative and quantitative data. With custom interfaces for analysis, communication, and implementation. (Ventana Systems, Inc., 2024). It can be downloaded for free at https://vensim.com/free-download/.

The system dynamics was used in this analysis. System Dynamics (SD) helps us to understand non-linear behavior by modeling a system using feedback loops, stocks, and flows to represent and analyze relationships within a system (Systems Complex, 2023). In the diagrams that follow, you can see that the boxed lines are the parameters between each compartment with the curved lines being the input and outputs of each compartment.

For this research, a 10-year period was considered. The variable N in this model is the initial population of the region in the free compartment. As mentioned previously, all other compartments are assumed to start with zero. The rate at which offenders go from free to incarcerated states is the number of people entering the drug court for that region divided by the total population. This, along with the other rates in the table, will change depending on which county is being looked at in the simulation. For the first simulation, Oklahoma as a state was run with the number listed previously. This simulation model

can be seen in Figure 6.13. Notice that on the right-hand side there are several totals and cumulative totals calculated. This is how to model is validated by seeing the increase or decrease in dollars spent/saved over the 10-year period. The treatment cost of $1583 per person (assuming a person can only enter treatment once a year) per year was obtained from (Wagener, 2022). This is the amount it costs the taxpayer for a person to be in a state funded rehab over 30 days. This number is assuming the average per person. This number is multiplied by the total number of people in treatment and then integrated over the entire 10 years to get a cumulative treatment cost. Similarly, the cost to the tax payers for an inmate to be incarcerated for one year is $16,497 (ver, 2015) is also calculated by multiplying by the number of people in the incarcerated compartment and then integrated over the 10-year period. The total cumulative cost is then the sum of the two cumulative costs. The plots of incarcerated over time, treatment over time, and cumulative cost for Oklahoma can be seen in Figure 6.18 and Figure 6.20.

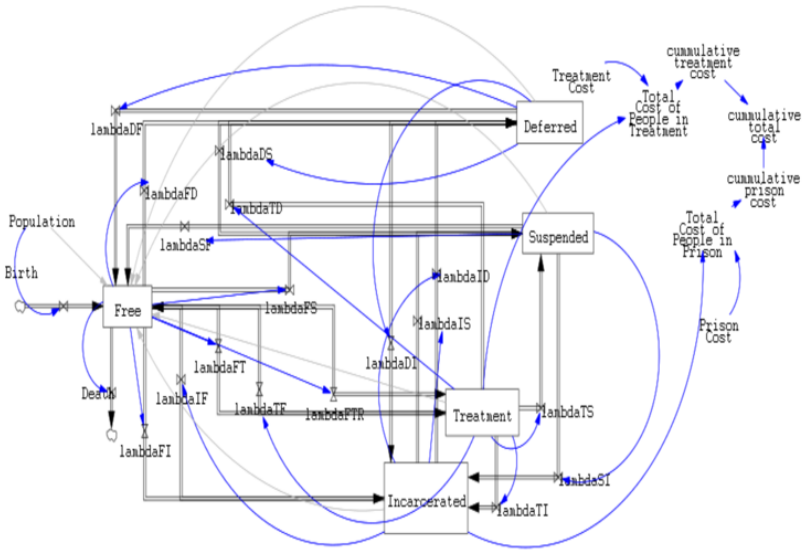


Figure 6.13: Vensim Punishment Focused Model

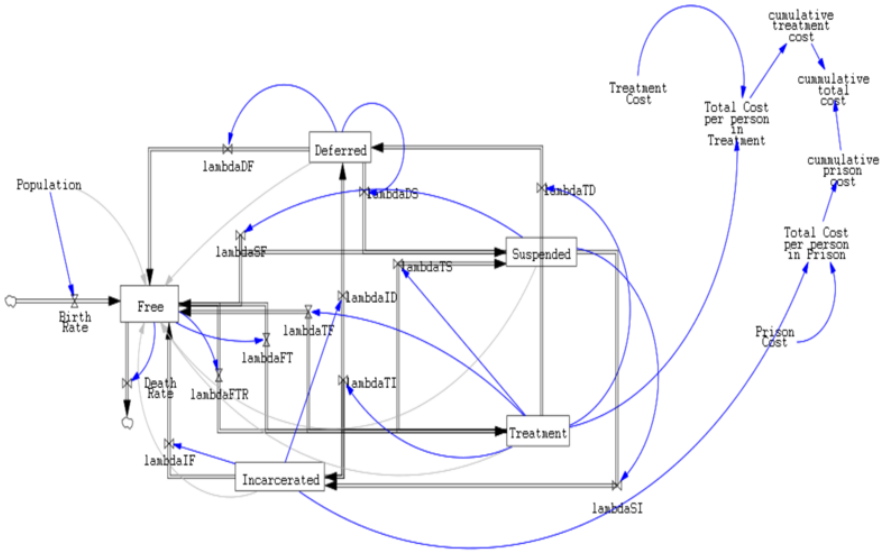


Figure 6.14: Vensim Treatment Focused Model

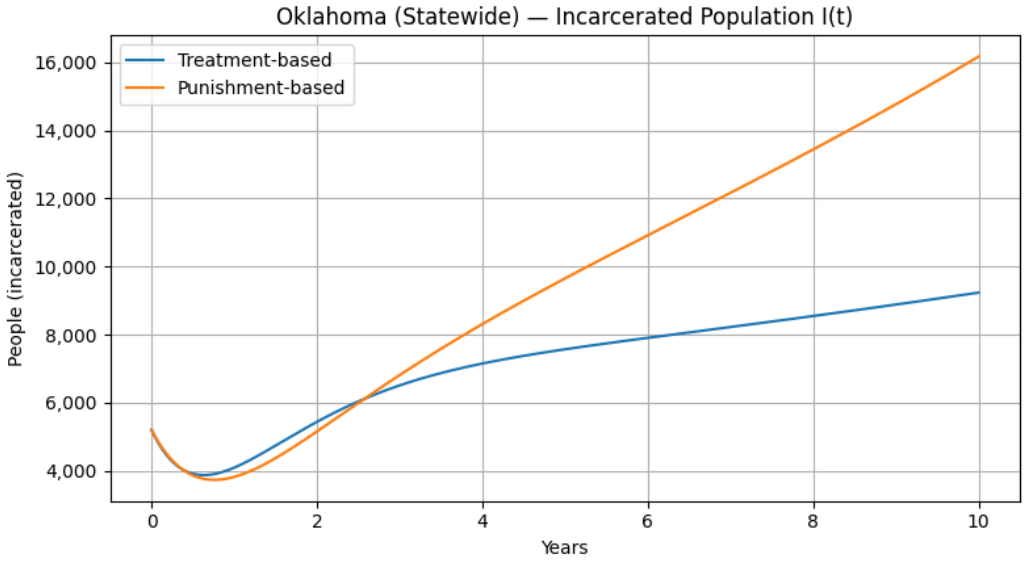


Figure 6.15: Model Comparison Incarcerated Over Time for Oklahoma

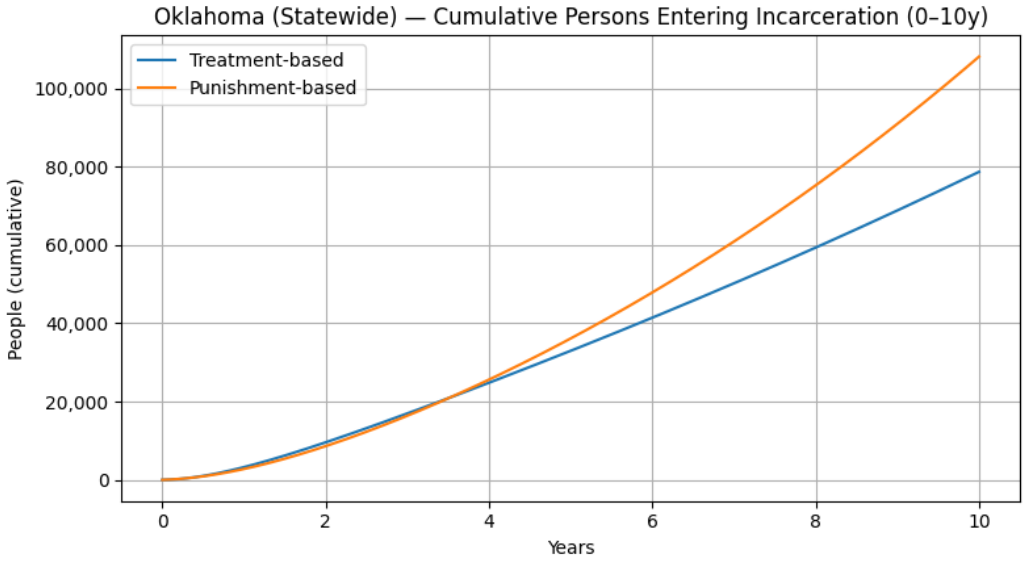


Figure 6.16: Model Comparison Over Time Incarcerated Cumulative for Oklahoma

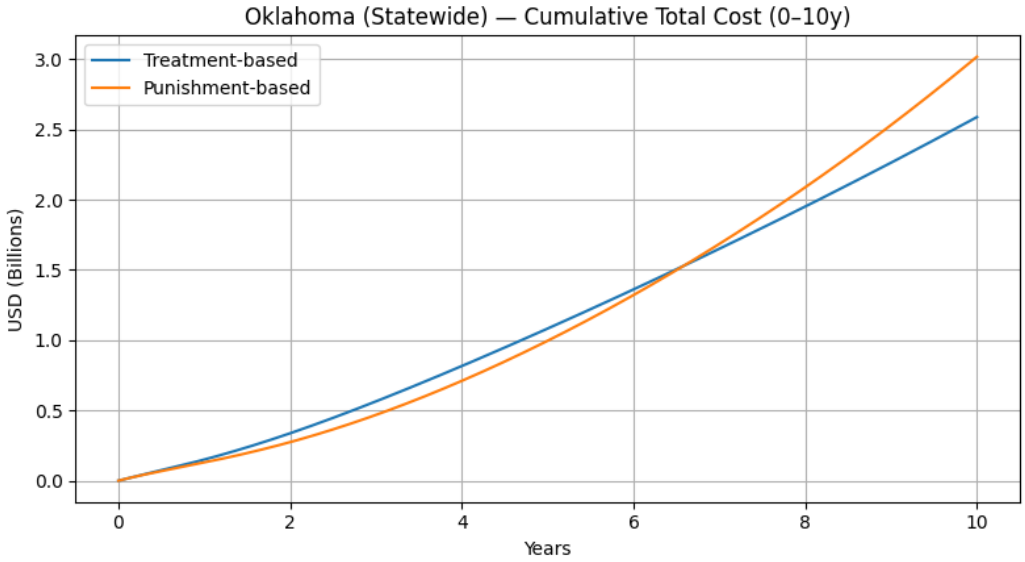


Figure 6.17: Model Comparison Over Time Cost for Oklahoma

6.7 Results

By comparing the two simulations for the state of Oklahoma, we can see that the number of people incarcerated over the 10 years for the first

punishment-based model is around 110,000 and 80,000 cumulative incarcerated persons for treatment-based model and punishment-based models respectively over a 10 year period. There were peaks of 65,000 people, with just roughly 39,000 with a peak of around 30,000 people between two and four years in both models, a cumulative total cost of three billion for punishment model and around 2.6 billion treatment-based model over a 10 year period. This is a savings of 400 million to choose the treatment first model.

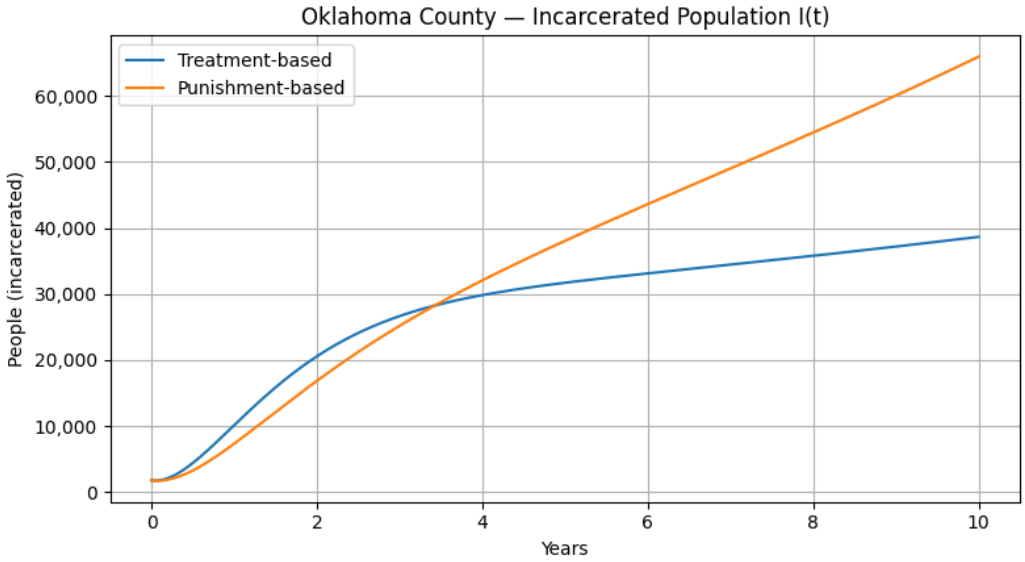


Figure 6.18: Model Comparison Incarcerated Over Time for Oklahoma County

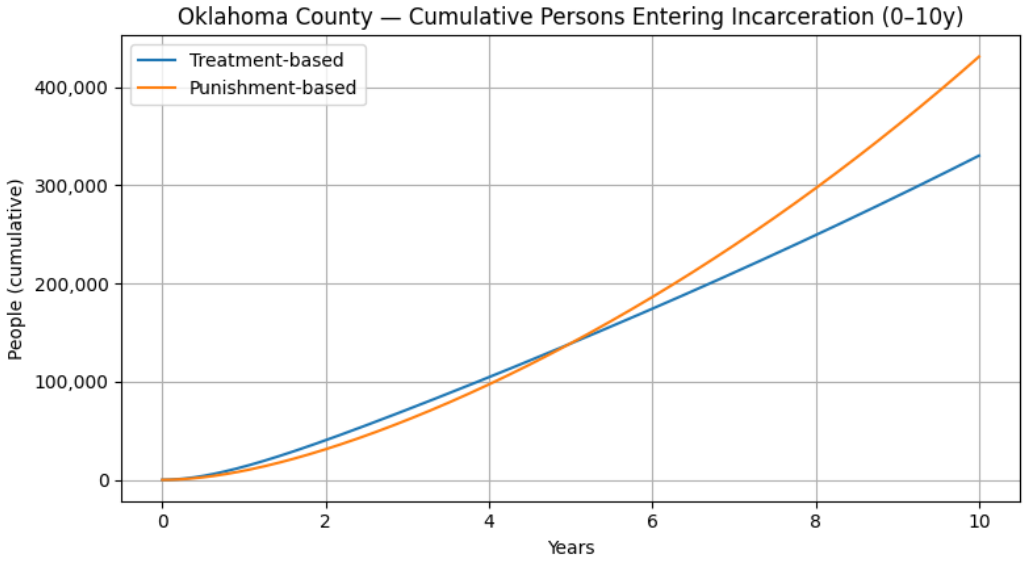


Figure 6.19: Model Comparison Over Time Incarcerated Cumulative for Oklahoma County

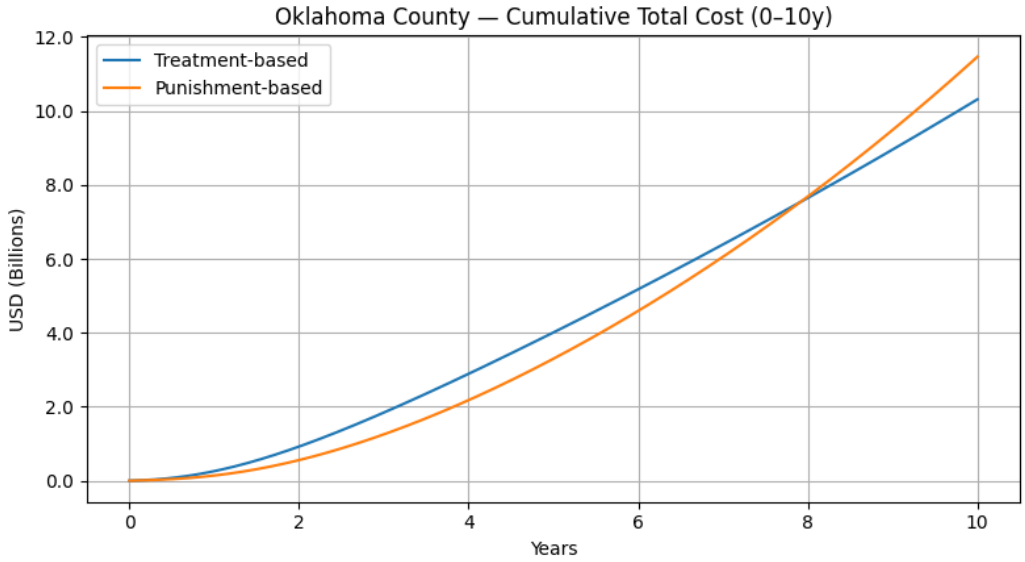


Figure 6.20: Model Comparison Over Time Cost for Oklahoma County

All of the above plots (Figures 6.18, 6.19), and 6.20 are for Oklahoma County data only. The plots for Beckham County data are shown in Figures 6.21, 6.22, and 6.23

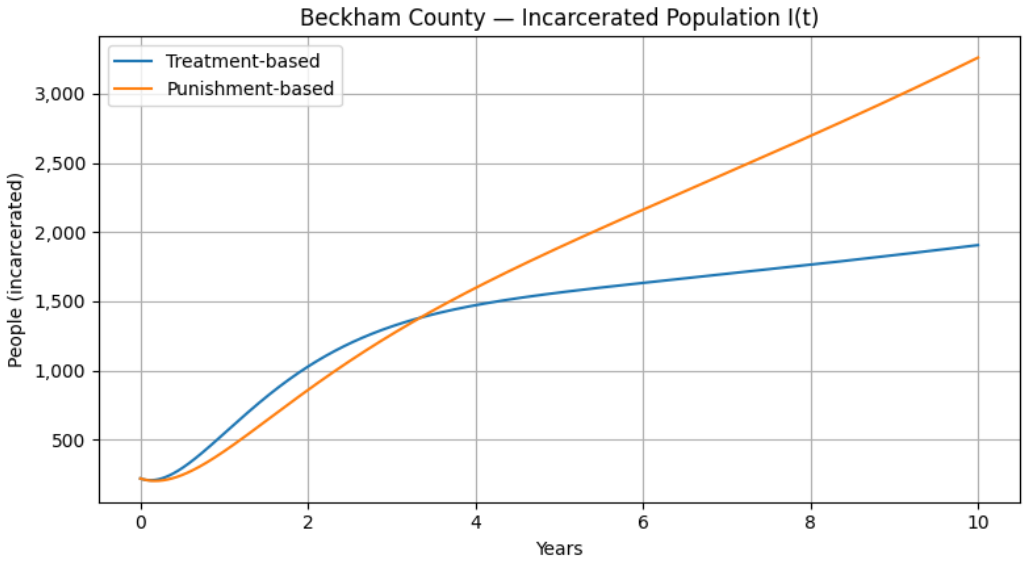


Figure 6.21: Model Comparison Over Time for Beckham County

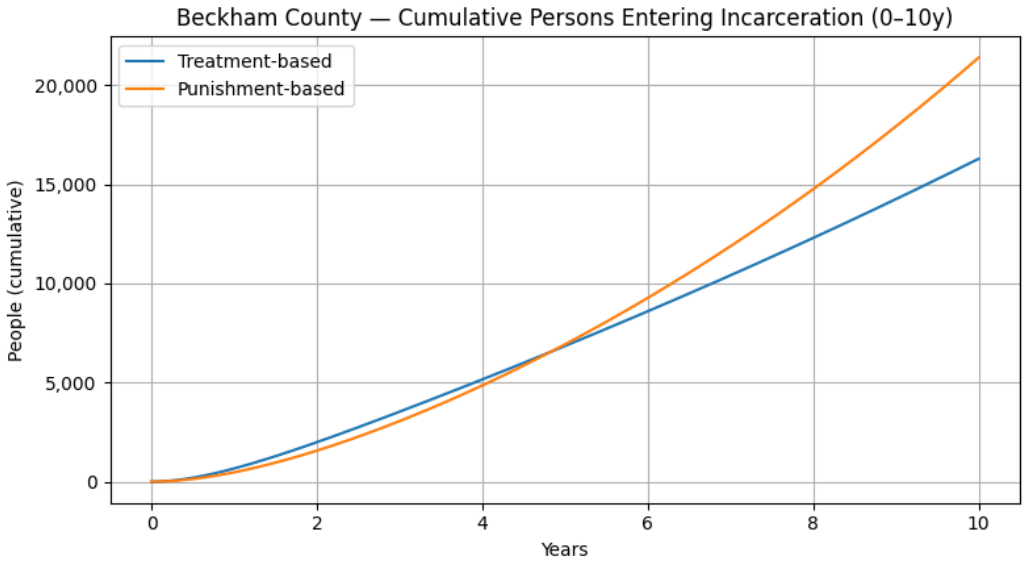


Figure 6.22: Model Comparison Over Time Incarcerated Cumulative for Beckham County

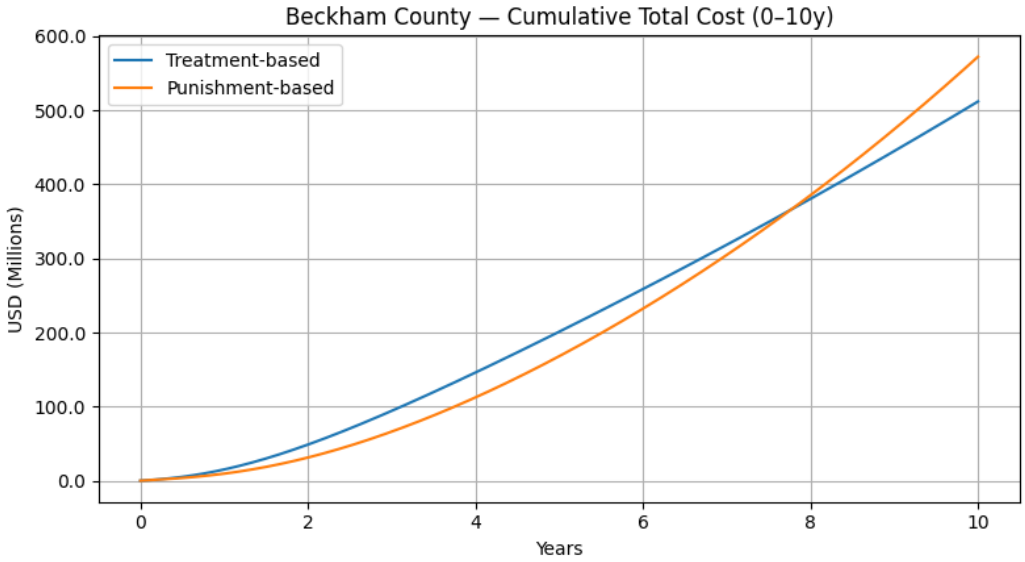


Figure 6.23: Model Comparison Over Time Cost for Beckham County

6.8 Sensitivity Analysis via Python Translation

To perform a detailed sensitivity analysis, the original Vensim models (treatment-focused and punishment-focused) were translated into Python. This allowed the use of the SALib package, which supports Sobol variance-based sensitivity analysis across multiple parameters.

What is Sobol Sensitivity Analysis?

Sobol sensitivity analysis is a global variance-based method used to quantify the contribution of individual input parameters and their interactions to the overall output variance of a model. It is particularly well-suited for nonlinear and non monotonic systems such as those found in epidemiological or justice system models.

The analysis decomposes the output variance into portions attributed to each parameter (first-order effects) and combinations of parameters (higher-order effects). The key metrics include:

- **First-order Sobol index (S_1):** Measures the proportion of output variance explained by varying one input alone, holding all others fixed.
- **Total Sobol index (S_T):** Captures the total contribution of a parameter, including both its individual effect and its interactions with other parameters.

Using Sobol analysis, we can identify which parameters most influence model outcomes (e.g., incarceration rates or total cost) and prioritize them for data refinement or policy intervention. This method is robust, model-independent, and widely used in uncertainty quantification for complex systems.

For the **treatment-centered model**, the following seven were evaluated:

- λ_{DS}: Deferred $\rightarrow$ Suspended
- λ_{TS}: Treatment $\rightarrow$ Suspended
- λ_{SF}: Suspended $\rightarrow$ Free
- λ_{SI}: Suspended $\rightarrow$ Incarcerated
- λ_{TI}: Treatment $\rightarrow$ Incarcerated
- λ_{ID}: Incarcerated $\rightarrow$ Deferred

- λ_{IF}: Incarcerated → Free

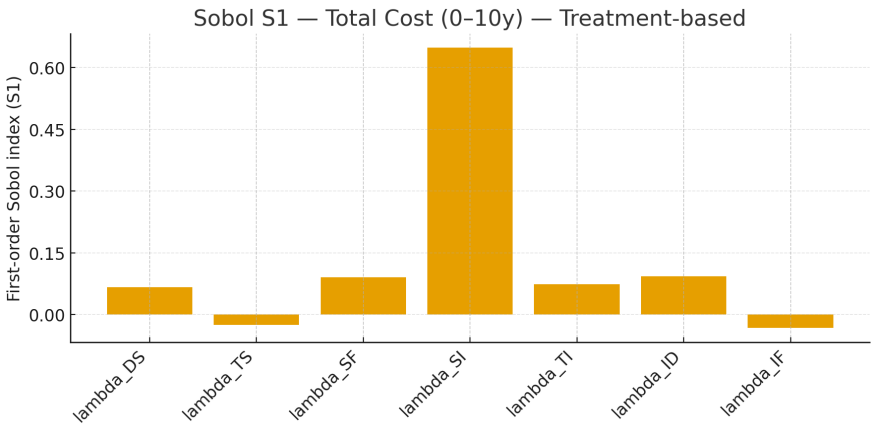


Figure 6.24: Sobol Sensitivity for Total Cost (Treatment Model)

Interpretation:

- The number 1 cost driver is how often people on suspension/supervision end up incarcerated (the S→I rate).
- If fewer supervised people are revoked back to prison, total costs drop the most.
- A second tier of smaller factors: how quickly people leave supervision to freedom, leave prison to deferred, go from treatment to prison, and move from deferred to supervision. These matter, but much less than S→I.
- We did not vary the per-person treatment cost in this sensitivity run, so the results are about flows between states, not price tags.

For the **punishment-focused model**, the seven parameters evaluated were as follows:

- λ_{DS}: Deferred → Suspended
- λ_{TS}: Treatment → Suspended
- λ_{SF}: Suspended → Free
- λ_{SI}: Suspended → Incarcerated

- λ_{TI}: Treatment → Incarcerated
- λ_{ID}: Incarcerated → Deferred
- λ_{IF}: Incarcerated → Free

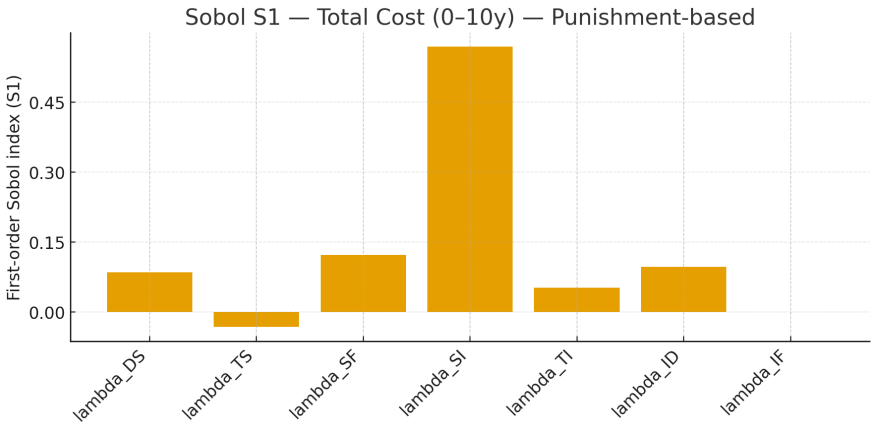


Figure 6.25: Sobol Sensitivity for Total Cost (Punishment Model)

Interpretation:

- Again, the biggest driver of total cost is supervision → incarceration (S→I).
- Changing that one pipeline explains about half (or more) of why total cost goes up or down.
- Smaller contributors include supervision → free, incarcerated → deferred, deferred → supervision, and treatment → incarceration.
- Routine releases from prison to free (I→F) and treatment → supervision (T→S) had little effect.

6.9 Validation

To validate that the suggested treatment-based model performs better, the rates and populations were modified for a larger county and a smaller county to see if the same results occur. Oklahoma County was chosen because it had a large population; in contrast, Beckham County was chosen because it had a small population. The smallest counties in Oklahoma did not have a

drug court program in place to obtain data however; Beckham County had a smaller population in relation to the other counties. The following rates and initial values were changed, as can be seen in the last two columns of Table 6.2.

Parameter	Meaning	Baseline (State)	Sobol Lower	Sobol Upper	Oklahoma Co.	Beckham Co.
λ_{DS}	Deferred → Suspended	0.50 (assumed)	0.00	1.00	0.50	0.50
λ_{TS}	Treatment → Suspended	0.203	0.00	1.00	0.002	0.391
λ_{SF}	Suspended → Free	0.50 (assumed)	0.00	1.00	0.50	0.50
λ_{SI}	Suspended → Incarcerated	0.50 (assumed)	0.00	1.00	0.50	0.50
λ_{TI}	Treatment → Incarcerated	0.50 (assumed)	0.00	1.00	0.50	0.50
λ_{ID}	Incarcerated → Deferred	0.088	0.00	1.00	0.004	0.043
λ_{IF}	Incarcerated → Free	0.581	0.00	1.00	0.908	0.261

Table 6.2: Total Cost — Punishment Model: Sobol parameter settings and county overrides used in validation. Baselines come from the state-level table; county overrides use Oklahoma and Beckham county values where available, else assumptions.

6.10 Conclusions

From the two simulations being run on Oklahoma state data, we can see a significant cost difference between the two models. Similarly, by comparing the plots for both the large county and the small county for punishment versus treatment-focused models, we can see that they both perform better for the treatment-focused model regardless of the population size. The trends look roughly the same just proportional to the population size. For this reason, we can conclude that giving the drug offenders the opportunity to attend treatment first and then if failed go to a deferred sentence, then if that fails, accelerate to a suspended sentence and finally if they fail that, then that is when they are incarcerated. In all model comparisons run there were a lot less people incarcerated over the 10-year time period for the treatment-focused models. Another thing to note is that according to the flow rates and relapse rates, it looks like the number of people in the beginning of both models has a spike in the first few years in the number of people in treatment and then goes back down over time. This could be due to the success rates for the completion of treatment or in this case the drug court. A limitation of these models is that there were no warm-up periods used, which causes some of the compartments to start out empty, making the results slightly skewed. In the next chapter, we look at a different model in which we use a warm-up period in each section to have a more accurate view of what is going on. This could explain why there is such a large delta between the cost of prison vs treatment models.

Although none of the methods from the main literature review was used in this chapter, a new literature review was performed to look at compartmental models and how they can be used to model criminological data. This was

then applied to this chapter. The next chapter builds on this idea however it adds the concept of fairness to it.

Chapter 7

Poverty-Structured $SIRV_2$ Model with Fairness Modeling Methamphetamine Addiction in Oklahoma

7.1 Introduction

Methamphetamine abuse has become a huge problem in the United States. There are 43 per 1,000 estimated meth users of the age 18 and older in Oklahoma in 2021, 1 per 1,000 between the ages 12 and 17, 43 per 1,000 of the age 12 and older, 1.3% of adults used meth between the years of 2018 and 2019, with a 53% change in meth death rate between the years 2015 and 2019 (World Population Review, 2024). Meth users are more susceptible to catching illnesses such as HIV and Hepatitis C due to risky behaviors as well as the use of "dirty" needles. We have noticed that the majority of people who become users tend to have come from a background of poverty or lived in a high poverty area while growing up. For this reason, we plan to modify the White and Comiskey SIR ODE model (Battista, 2015) to incorporate a second independent variable of which will be average federal poverty level for the county. We are interested in how people in each poverty level bracket fall into each class of the SIR model, whether it be susceptible to using, actively using, or recovered. We then want to add a dual vaccine component which in our case would be a trip to detox and then potentially a trip to a rehabilitation facility. Once this model is done, we want to map out the poverty level by county for the state of Oklahoma and then compare it with the location of Alcoholics Anonymous (AA)/ Narcotics Anonymous (NA) meetings. Finally, we will compare these plots to see if there are hot spots of "infected" addicts in areas of lower poverty as well as hot spots of recovered addicts in areas that

have a higher frequency of AA/NA meetings taking into account location of the meetings as well as available meetings per area in square feet.

7.2 Research Questions

- **How does poverty influence the dynamics of methamphetamine addiction?**
  - Examines how addiction spreads in various socioeconomic groups, focusing on those in poverty.
- **What impact do poverty levels have on recovery and relapse rates?**
  - Analyzes how individuals transition between addiction, recovery, and relapse.
- **How does the spatial distribution of recovery resources correlate with poverty?**
  - Assesses geographic disparities in access to AA/NA meetings across Oklahoma counties.
- **Can spatial modeling (e.g., hotspot analysis) enhance understanding of addiction patterns?**
  - Investigates spatial distribution of addicts and recovery resources.
- **Future Work: Incorporating fairness as a treatment vaccine component**
  - Explore fairness in access to treatment using a vaccine framework to ensure equitable recovery chances using the Gini Index.

7.3 Methods

7.3.1 The Model

In the SIRV2 model, there are five compartments used. S is the susceptible class, I is the infected class, R is the recovered/removed class, V_1 is entering a detox treatment class, and V_2 is entering a 30-day rehabilitation class. For this model, there will be two independent variables, t for time (in years) and p for the percentage of the county in poverty. In this model, $S(p,t)$ will be

the susceptible people or people who are not using methamphetamines at poverty level p, at time t. $I(p,t)$ will be the infected people or people who are actively using methamphetamines at poverty level p, time t. $R(p,t)$ will be the recovered people or people who have ceased using methamphetamines due to going into drug court at poverty level p, time t. We have $V_1(p,t)$ which is entering a detox treatment facility (first vaccine) at poverty level p, time t. Finally, $V_2(p,t)$, which is entering a 30-day rehabilitation facility in poverty level treatment p, time t. The transition diagram of this model can be seen in Fig 7.1.

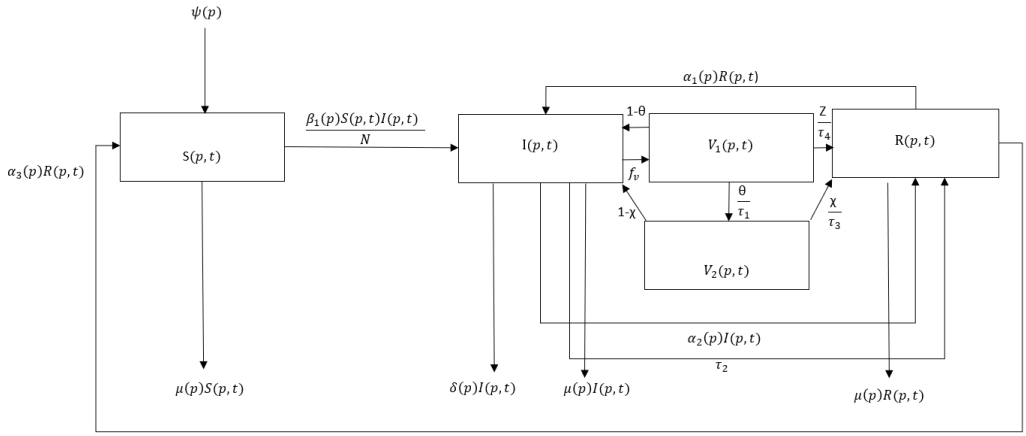


Figure 7.1: Transition Diagram

The flow of this model can be described as follows. The initial population is born into the model at the susceptible class. For simplicity, we are going to ignore the case in which a person is born addicted to methamphetamines due to their mother using drugs during her pregnancy. Taking this into account, we will let $\psi(p)$ represent the number of people entering the susceptible population (entering poverty within their county). Once in the susceptible class, they can either die naturally at a rate of $\mu(p)$ or enter the infected class (users of methamphetamines). Once they enter the infected class, they can die naturally, die of an overdose, or enter the recovered class (enter drug court). They can also enter the first vaccination class. Once in the first vaccination class, they can go back to the infected class, to recovered class, or to second vaccination class. After reaching the second vaccination class, they can go back to the infected class or to the recovery class. Once in the recovery class, they can relapse back into the infected class (relapse back to active using), go die naturally, or simply return to the susceptible class after treatment (drug court). They could also leave any of these classes if they leave poverty. This will be taken into account in the boundary conditions. This model can be described as a dynamical system of 5 coupled, non-linear partial differential equations and can be seen in the following formulas:

To clarify how socioeconomic inequality influences addiction outcomes in our model, we introduce a conceptual pathway linking key variables. Figure 7.2 illustrates how county-level poverty rates inform the Gini index, which we use as a proxy for income inequality. This Gini index, in turn, governs the allocation of intervention resources—modeled as vaccine compartments representing detox and rehabilitation services. The fairness of this allocation directly affects recovery outcomes by modulating access to treatment. Counties with lower Gini coefficients, reflecting more equitable income distributions, are assumed to have more balanced access to these critical services, thereby improving recovery rates and reducing relapse risk.

Figure 7.2: Conceptual pathway linking county-level poverty, income inequality (Gini index), and fairness-based allocation of detox and rehabilitation resources. Lower Gini values correspond to more equitable treatment access and reduced relapse.

System of Equations

$$\frac{\partial S(p,t)}{\partial t} - D_S \frac{\partial S(p,t)}{\partial p} = \psi(p) + \alpha_3(p)R(p,t) - \frac{\beta_1(p)S(p,t)I(p,t)}{N} - \mu(p)S(p,t), \quad t > 0, \quad 0 < p < 1$$

$$\frac{\partial I(p,t)}{\partial t} - D_I \frac{\partial I(p,t)}{\partial p} = \alpha_1(p)R(p,t) + \frac{\beta_1(p)S(p,t)I(p,t)}{N} - (\mu(p) + \delta(p))I(p,t) - \alpha_2(p)I(p,t) - \tau_2(p)I(p,t), \quad t > 0, \quad 0 < p < 1$$

$$\frac{\partial R(p,t)}{\partial t} - D_R \frac{\partial R(p,t)}{\partial p} = \alpha_2(p)I(p,t) - \alpha_1(p)R(p,t) - \mu(p)R(p,t) - \alpha_3(p)R(p,t) + \tau_2(p)I(p,t) + \frac{Z}{\tau_4}V_1(p,t), \quad t > 0, \quad 0 < p < 1$$

$$\frac{\partial V_1(p,t)}{\partial t} - D_{V_1} \frac{\partial V_1(p,t)}{\partial p} = f_v(p)I(p,t) - (1-\theta)V_1(p,t)$$

$$- \frac{\theta}{\tau_1} V_1(p,t) - \frac{\delta}{\tau_4} V_1(p,t), \qquad t > 0, \quad 0 < p < 1$$

$$\frac{\partial V_2(p,t)}{\partial t} - D_{V_2} \frac{\partial V_2(p,t)}{\partial p} = \frac{\theta}{\tau_1} V_1(p,t) - (1-\chi) V_2(p,t) - \frac{\chi}{\tau_3} V_2(p,t), \qquad t > 0, \quad 0 < p < 1$$

Boundary Conditions

$$\begin{array}{lll} S(0,t) = 0, & S(1,t) = f(t), & t > 0 \\ I(0,t) = 0, & I(1,t) = g(t), & t > 0 \\ R(0,t) = 0, & R(1,t) = h(t), & t > 0 \\ V_1(0,t) = 0, & V_1(1,t) = w(t), & t > 0 \\ V_2(0,t) = 0, & V_1(1,t) = y(t), & t > 0 \end{array}$$

Initial Conditions

$$\begin{aligned} S(p,0) &= \phi(p) - \beta_0(p)\phi(p), && 0 \le p \le 1 \\ I(p,0) &= \beta_0(p)\phi(p), && 0 \le p \le 1 \\ R(p,0) &= \alpha_1(p)\beta_0(p)\phi(p), && 0 \le p \le 1 \\ V_1(p,0) &= \sum_{t=-30}^{0} f_v(p) I(p,t), && 0 \le p \le 1 \\ V_2(p,0) &= \sum_{t=-30}^{0} \left(\frac{\theta}{\tau_1} V_1(p,t) + (1-\chi) V_2(p,t) \right), && 0 \le p \le 1 \end{aligned}$$

7.3.2 Parameter Definitions

The parameters of this model consider the rates of death/exit from a class based on a person's poverty level. This is important because if they are at a high poverty level, they may not have access to or income available to go into a treatment program (afford the fees of drug court). At a high poverty level, they may also have more access to drugs because of the area they live in. On the other hand, the higher their poverty level, the bigger the chances are that

they will end up incarcerated because they may not be able to afford legal representation. This could lead to varying results which can explain how they enter and leave the infected and recovered classes. These parameters are all explained in Table 7.1

Parameter	Description	Units
$\psi(p)$	The rate at which people enter poverty level p	1/year
$\phi(p)$	The entire population of people in poverty level p	People
$\beta_0(p)$	The rate at which people in poverty level p get on drugs (infected)	1/year
$\beta_1(p)$	The interaction coefficient between susceptible and infected	1/year
$\alpha_1(p)$	The rate at which people in poverty level p enter drug court	1/year
$\alpha_2(p)$	The rate at which people in poverty level p fail drug court (relapse)	1/year
$\alpha_3(p)$	The rate at which people in poverty level p graduate drug court (recover)	1/year
$\mu(p)$	The natural death rate of people in poverty level p	1/year
$\delta(p)$	The enhanced (overdose) death rate of people in poverty level p	1/year
f_v	The Gini Coefficient (Fraction of population to receive first dose - Go to detox)	Dimensionless
θ	Detox efficiency	Dimensionless
χ	Treatment efficiency	Dimensionless
Z	Transition rate from detox to recovered	Dimensionless
τ_1	Time to arrival between detox and treatment	365/Days
τ_2	Time to arrival between infected and recovered	365/Days
τ_3	Time to arrival between treatment and recovered	365/Days
τ_4	Time to arrival between treatment one and recovered	365/Days
D_S	Diffusion coefficient for susceptible individuals	Poverty^2/year
D_I	Diffusion coefficient for infected individuals	Poverty^2/year
D_R	Diffusion coefficient for recovered individuals	Poverty^2/year
D_{V_1}	Diffusion coefficient for detox group	Poverty^2/year
D_{V_2}	Diffusion coefficient for treatment group	Poverty^2/year
N	The total population in the model	People

Table 7.1: Model Parameters, Descriptions, and Units

7.3.3 Assumptions

There have been several assumptions made for this model to keep it as simple as possible. These assumptions are listed as follows: We will assume that the only population existing is the one that is shown thus,

$$N = S\,(p,t) + I\,(p,t) + R(p,t) + V_1\,(p,t) + V_2\,(p,t) \tag{7.1}$$

Like mentioned previously, we will assume that no person is born straight into the infected (addicted) class. This means that every person must start in the susceptible class. The natural death rate will not be affected by the use of or stopping the use of drugs. We are only going to look at the population of people in the state of Oklahoma. Reasons for this will be seen later when we compare locations of 12-step meetings, areas of low poverty, numbers of

arrests, and densities of "infected" addicts. We will assume that the only way to drug court is through a court order, and the entrance, exits, completion, and failure rates between detox and the 30-day treatment facility are the same for all facilities as the ones obtained from phone interviews collected from the Referral Center (TRC) (Communication, 2022) and Firstep for Women (Janis, 2022). This helps to see if the two vaccine components work. We will assume that poverty levels for each county remain constant over time. We will assume that meeting locations are constant, meetings will not be created nor destroyed. This part will be used in later comparisons. Finally, we will assume that AA meetings and NA meetings are equivalent in what they provide in recovery and will only look at location of AA meetings as many addicts go to AA meetings over NA meetings due to availability and "wellness" of the group (wellness in the context of amount of long term sobriety among the members). We will use the Gini coefficient as a fairness aspect towards poverty for each county in the vaccine (or treatment 1 or 2) compartments. This value is pre-calculated and obtained from (AIDSVu, 2025). Since this is a nonlinear model and we are lacking the data for the interaction coefficient parameters, we chose the interaction coefficients that produced the best models. The results of this paper will focus on the building of the model itself and how the different calculations were determined using the drug court data. After these calculations are shown, we can simplify the model by adding in an example of some of the numbers for a few specific counties, and then look at the steady-state model. Then we will create a matrix containing the numbers for each of the 77 counties to show the model with the data that we do have available.

7.3.4 The Data

To get an idea of what data we compiled to use in this model, we have created Table 7.2, showing the fields in our data set, how they were calculated (if calculated from other fields).

Variable Name	Explanation	Calculation
County	County name	Constant by County
FIPS	County FIPS Code, used for choropleth plots and geocoding in R	Constant by County
Population	Population within the county	Constant by County
Rural Population	Population within the county thats rural	Constant by County
Area	Area of county	Constant by County
Poverty	Number of people in poverty	Constant by County
Poverty Percentage	Percentage of the county that is in poverty	$Poverty/(Population * Year)$
Death Rate	Natural rate at which people die	(Deaths)/(Population)
Meetings	Number of available meetings	Constant by County
Meeting Density	Number of Meetings available by area 2017	Meetings/Area
Deaths	Number of Deaths in 2016	N/A
Drug Deaths	Number of drug overdose deaths in 2016	Constant Given Over All Counties
Drug Death Rate	The rate at which drug overdoses occur per year	(Drug Deaths)/Population
Number in DC	Number of people in drug court	Constant by County
Percent Meth in DC	Percent of people in drug court that are on meth	Constant by County
Relapse Rate	Rate at which people relapse (fail drug court)	1-DC Grad %age Meth
DC Grad %age	Percentage of people that graduate drug court	Constant Given by County
DC Grad %age Meth	Rate at which people in drug court on meth graduate	(DC %age on Meth)/(DC Grad %age)
DC Entrance Rate	Rate at which people enter drug court	(Number in DC)/(Population)
Per Capita Income	County level per capita income	Constant by County
Average Monthly Income	County level per monthly income in drug court	Constant by County
Percent Rural	Percentage of the county that is rural	(Rural Population)/(Population)
Rank	County ranking by per capita income	Constant by County

Table 7.2: Explanation of Variables and Calculations. Sources for the fields within this data set include Census, CDC, Meeting Finder, Stacker, Vital Statistics, and Index Mundi.

We compiled a dataset based on the Table 7.2. The county poverty percentage distribution, the Gini index grouping calculations as well as the counties included in each Gini group can be seen in Tables 7.3, 7.4, and 7.5. The parameter values were calculated by county straight from the data set or using the transition rates from the simulation model. These can be seen in Table 7.6. Our τ values are based on a simulation model running using real-world data obtained from the phone interviews previously mentioned in the assumption section. This is shown in Table 7.7. From the simulation model, we were also able to see what a 30-day warm-up period would give us as initial values for each vaccination compartment. Table 7.8 gives these numbers. Thus, we were able to find the formulas for the initial values in general in our compartmental SIR model. These formulas are given in Table 7.9.

Poverty Group	**Poverty % Range**
Group 0	9.1 – 11.7
Group 1	11.7 – 14.3
Group 2	14.3 – 16.9
Group 3	16.9 – 19.4
Group 4	19.4 – 22.0
Group 5	22.0 – 24.6

Table 7.3: Poverty Percentage Bins

Gini Group	**Gini Range**	**Label**
0	0.4081 – 0.4309	Lowest inequality
1	0.4309 – 0.4414	Low-Mid
2	0.4414 – 0.4450	Mid
3	0.4450 – 0.4748	High
4	0.4748 – 0.5085	Highest inequality

Table 7.4: Gini Index Groupings and Interpretations

Index	Gini Inequality Group	List of Counties
0	Lowest Inequality	Washita, Grady, Hughes, McClain, Pontotoc, Rogers, Wagoner
1	Low-Mid Inequality	Bryan, Creek, Garfield, Jackson, Garvin, Ottawa
2	Moderate Inequality	Cleveland, Comanche, Craig, Lincoln, Mayes, Pottawatomie, Sequoyah
3	High Inequality	Delaware, LeFlore, McCurtain, Muskogee, Okmulgee, Seminole
4	Highest Inequality	Beckham, Cherokee, Custer, Oklahoma, Payne, Stephens, Tulsa

Table 7.5: Counties by Gini Inequality Group

Parameter	Value
$\beta_0(p)$	0.41
$\beta_1(p)$	0.05
$\alpha_1(p)$	0.10
$\alpha_2(p)$	0.40
$\alpha_3(p)$	0.60
$\psi(p)$	0 – Assumed Constant Population
Population	based on county
$\phi(p)$	based on county
$\mu(p)$	0 – Assumed Constant Population
$\delta(p)$	0 – Assumed Constant Population
f_v	0-1 based on county
Z	0.50
θ	0.54
χ	0.53

Table 7.6: Parameter values used in the model

Transition	Symbol	Distribution and Parameters
Vaccination 1 to 2 (if used)	τ_1	Assume Lognormal (Not explicitly provided)
Detox Length of Stay	τ_2	Beta Distribution: $\alpha = 1.37$, $\beta = 1.3$, $\times 0.5 + 0.5$
Treatment Length of Stay	τ_3	Triangular Distribution: (min = 1, mode = 28, max = 30)
Interarrival/Wait Impact	τ_4	Lognormal Distribution: (logmean = 0.234, logstd = 0.194, +1.15)

Table 7.7: Transition Times and Their Distributions

Category	**Count**
Total Detox Arrivals	157.000
Detox Patients (Follow-through)	109.900
Detox Completed	59.346
Detox Only (Complete, no treatment)	29.673
Detox to Treatment	29.673
Total Treatment Arrivals	151.200
Direct Treatment Patients	151.200
Total Treatment Patients (Direct + Detox)	180.873
Treatment Completed	95.863
Treatment Failed	85.010

Table 7.8: Detox and Treatment Outcomes after 30-Day Warm-up Period (Provided Data)

Class	Initial Values
$S(p,0)$	$\phi(p) - \beta_0(p)\phi(p)$
$I(p,0)$	$\beta_0(p)\phi(p) - \alpha_1(p)\beta_0(p)\phi(p)$
$R(p,0)$	$\alpha_1(p)\beta_0(p)\phi(p)$
$V_1(p,0)$	$\sum_{t=-30}^{0} f_v(p)I(p,t)$
$V_2(p,0)$	$\sum_{t=-30}^{0} \left(\frac{\theta}{\tau_1}V_1(p,t) + (1-\chi)V_2(p,t)\right)$

Table 7.9: Initial values for each class in the model

7.3.5 Results

Next, we plugged this data into our model and looked at 12 plots. The first six were based on poverty percentage. We looked at the three counties with the highest poverty percentage and the three counties with the lowest poverty percentage. We show the 5-year SIRV2 plot for each county. Those six plots can be seen in Figures 7.3, 7.4, 7.5, 7.6, 7.7, and 7.8

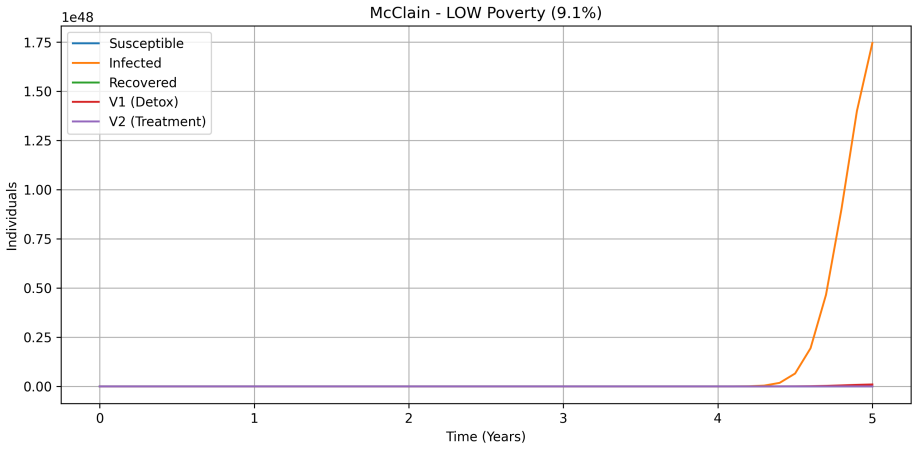


Figure 7.3: McClain Low

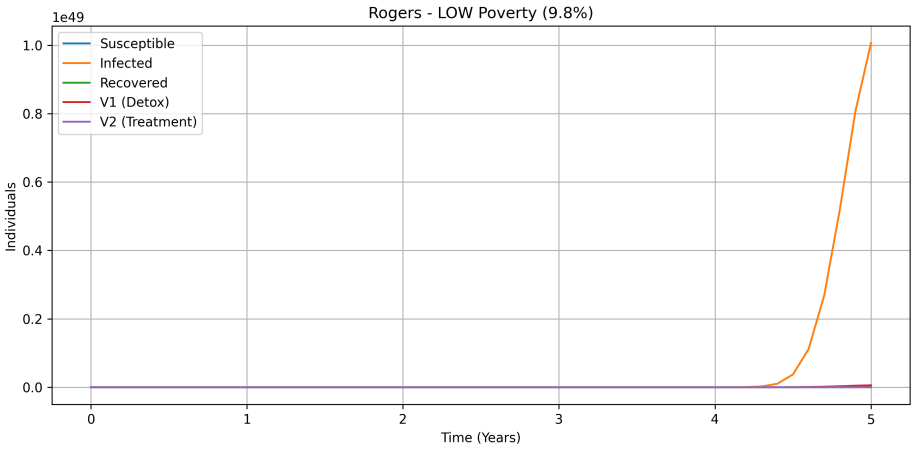


Figure 7.4: Rogers Low

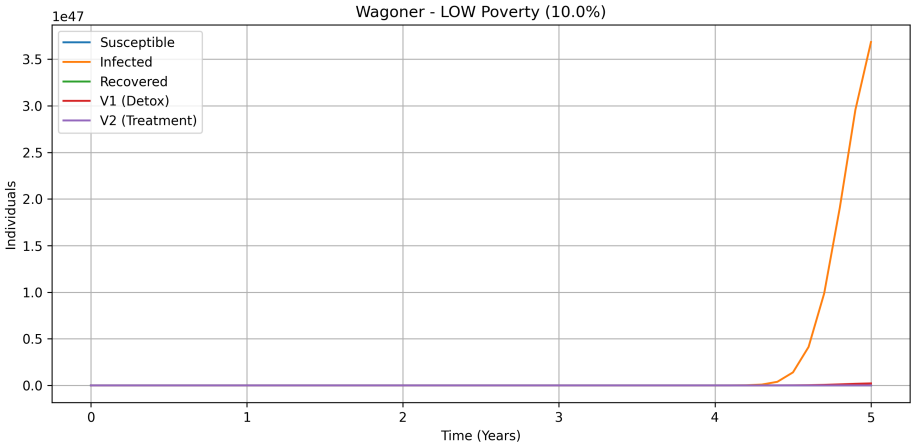


Figure 7.5: Wagoner Low

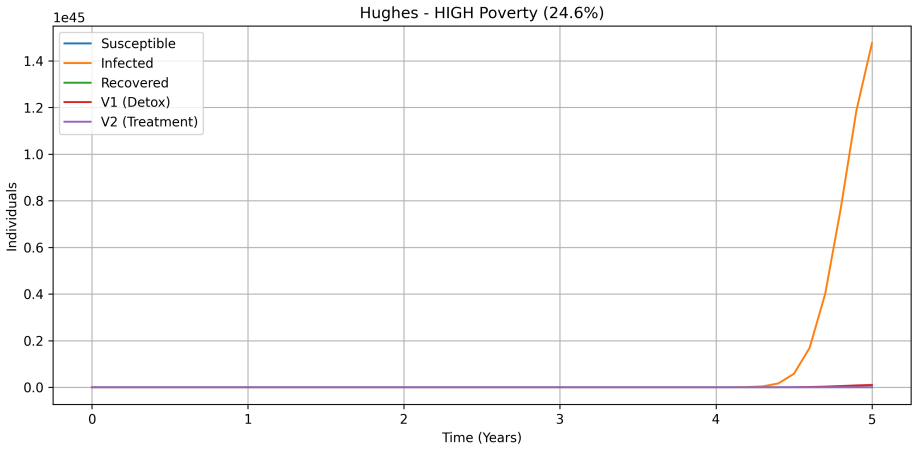


Figure 7.6: Hughes High

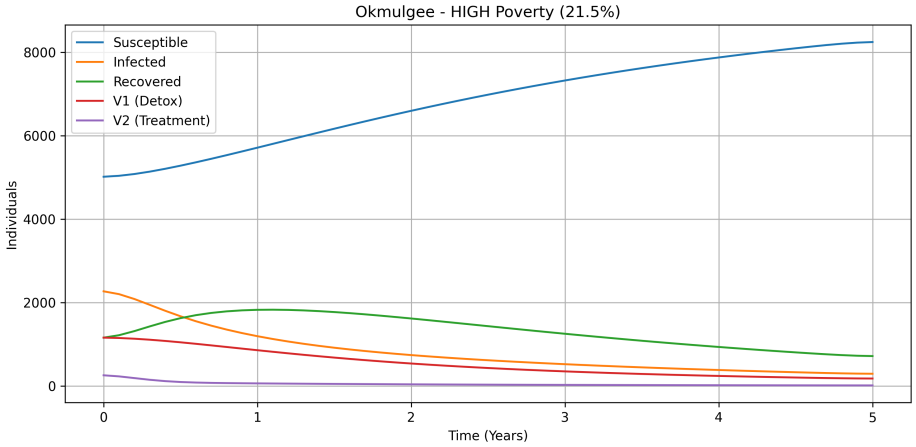


Figure 7.7: Okmulgee High

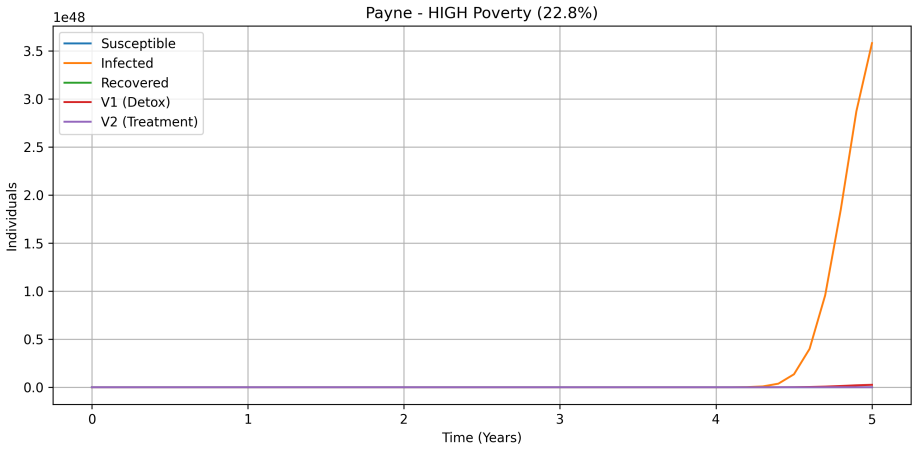


Figure 7.8: Payne High

From these plots we were able to verify that like expected, with the current initial conditions, the infected class exploded to infinity in almost all of the counties regardless of the poverty levels. For this reason, we look at what we can do to change this issue by tweaking the parameters. After changing the parameters, we can see the top three and bottom three charts look significantly different than the previous charts and now look like a typical SIR model result. These can be seen in Figures 7.9 through 7.14.

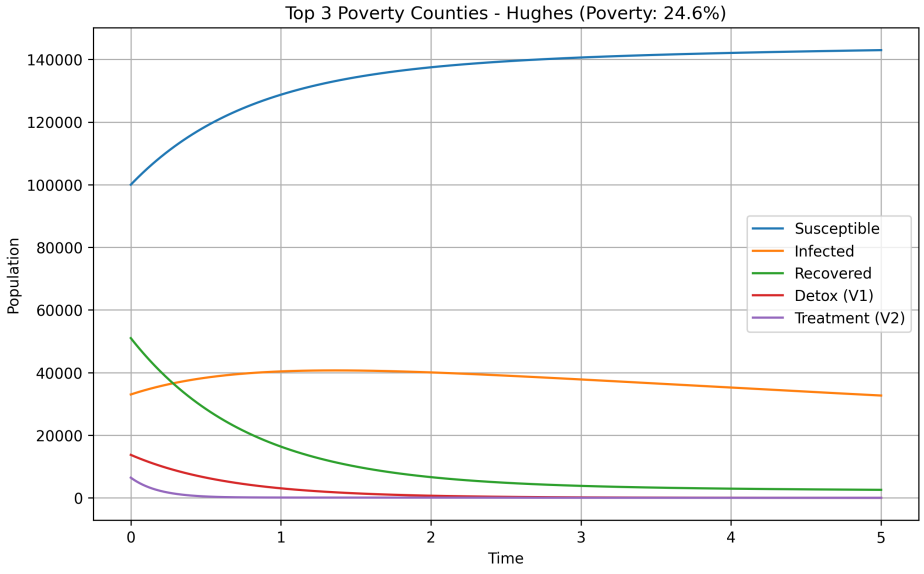


Figure 7.9: Hughes High

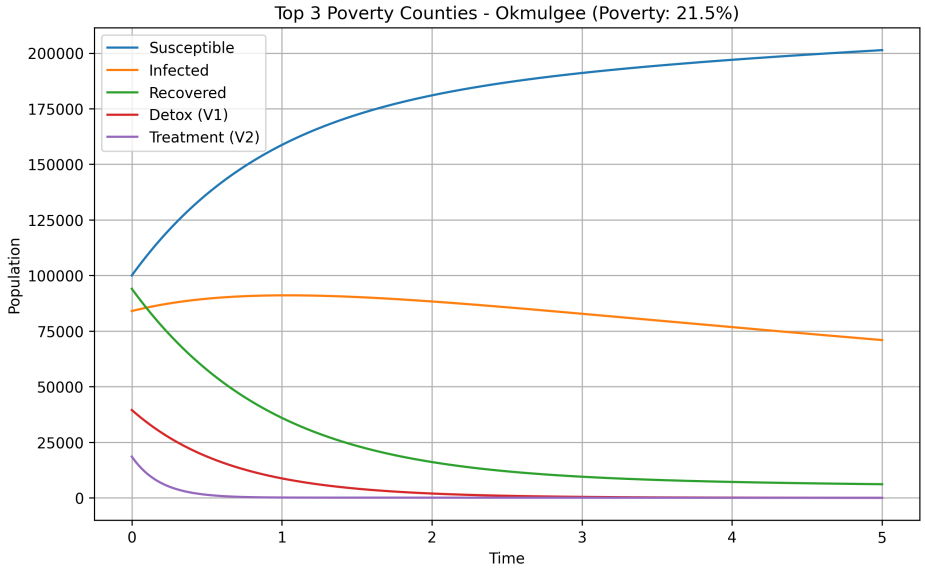


Figure 7.10: Okmulgee High

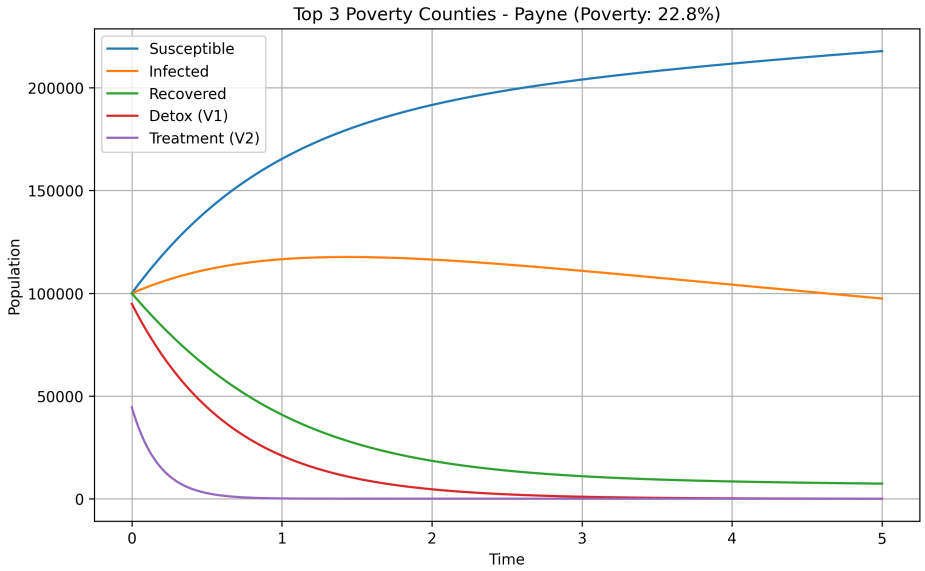


Figure 7.11: Payne High

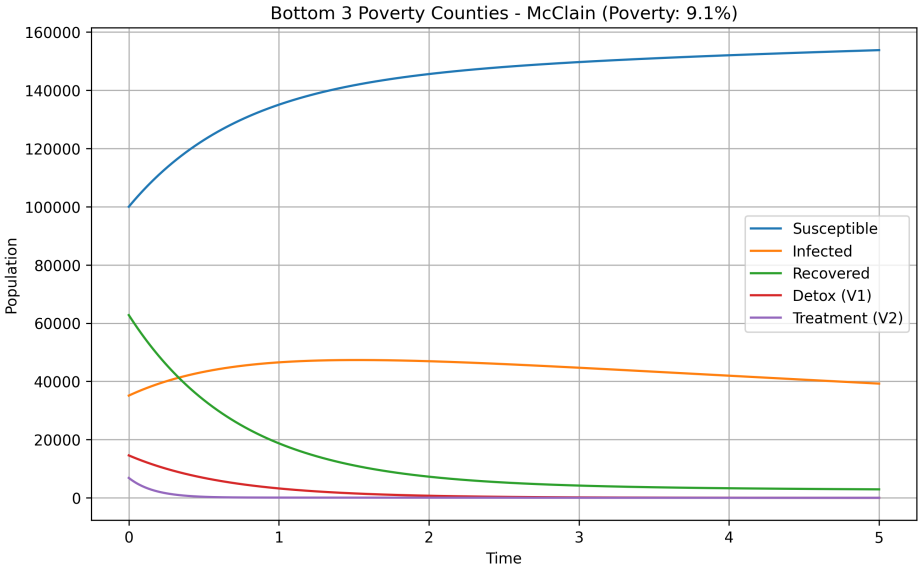


Figure 7.12: McClain Low

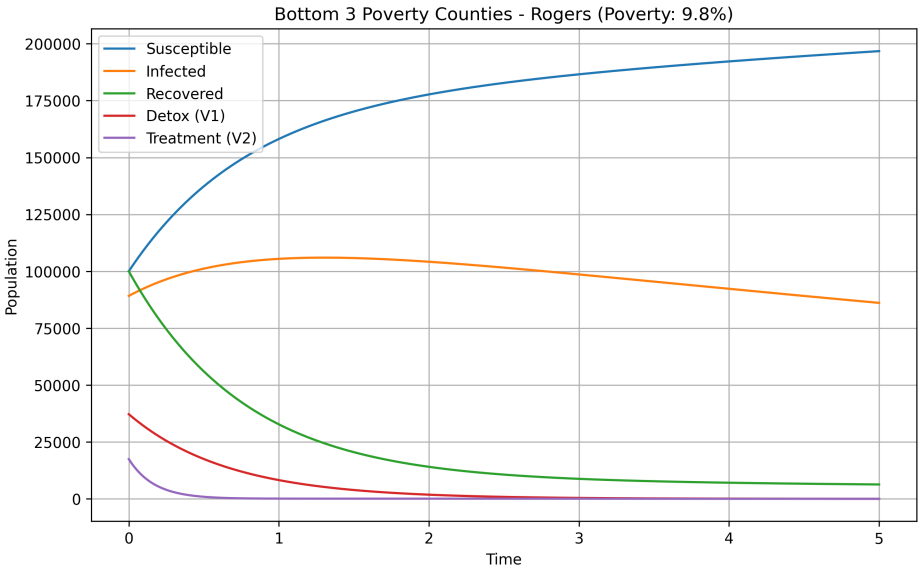


Figure 7.13: Rogers Low

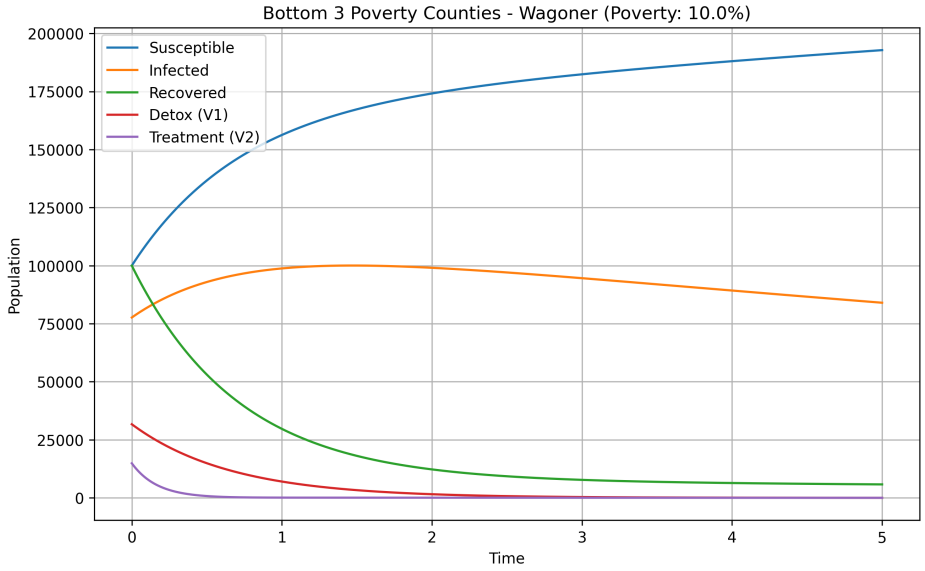


Figure 7.14: Wagoner Low

In reviewing the SIRV2 model results, differences are observed between counties with higher and lower poverty rates. In higher poverty counties such as Hughes, Okmulgee, and Payne counties have a susceptible population that decreases quickly at the start of the model, demonstrating a strong diffusion effect implying that methamphetamine use spreads rapidly through the population. Over time, the susceptible group stabilizes at a lower level as fewer individuals remain unexposed to the risk. The infected population in these counties rises sharply, peaks early, and then gradually declines, further illustrating the diffusion process followed by recovery or an exit from addiction. The recovered population grows slowly at first but increases later in the model, although it never fully compensates for the earlier losses in the susceptible group. The detox (V_1) and rehab (V_2) populations initially increase and then gradually decline, indicating a continued demand for treatment and a slower movement through recovery.

In contrast, the lower poverty counties such as McClain, Rogers, and Wagoner counties display a reverse diffusion behavior. The susceptible population actually increases over time, implying that fewer people develop addiction, and some individuals may transition back to a non-using state. The infected population increases slightly but rapidly plateaus and declines, showing the containment of the spread of addiction. The detox and rehab populations show sharp initial peaks followed by rapid declines, indicating that people who enter treatment tend to complete it quickly, with fewer requiring long-term care.

These results highlight a clear distinction between the diffusion of addiction

seen in counties of higher poverty and the reverse diffusion trends in counties of lower poverty, where susceptibility increases and the addiction rate declines more effectively over time. The parameters that were changed can be seen in Table 7.10.

Table 7.10: Parameter Changes from Original Model

Parameter	Original Value	Modified Value	Change Description
$\beta_0(p)$	0.41	0.25	Reduced addiction susceptibility.
$\beta_1(p)$	0.05	0.01	Lowered infection force.
$\alpha_1(p)$	Meth% / 100	Same	No change; based on meth use.
$\alpha_2(p)$	1− Completion Rate	Clipped at 0.8	Limited relapse rate.
$\alpha_3(p)$	Completion Rate	Same	No change.
$\mu(p)$	Deaths / Population	Same	No change.
$\delta(p)$	$0.2 \cdot \mu$	Same	No change.
$f_v(p)$	Gini Index	Same	No change
θ	0.54	same	No change.
χ	0.53	same	No change.
Z	0.50	same	No change.
τ_1	$1.15 + (1-$ Completion Rate)	Same	No change.
τ_2	1− Completion Rate	Min value of 0.1	Prevented low exit rates.
τ_3	Based on Income	Same	No change.
τ_4	1.15+ Gini Index	Same	No change.

7.3.6 The Statistics

In order to understand if this is a useful model, we decided to look at some of the statistics within our available data set. Now that we are looking at poverty percentages, we can look at data for each individual county for our model in the same way. For instance, we used CDC and Census data to find the natural death rates by county in Oklahoma for 2016. This can be done with the other county level data in the same way so that the model stays consistent. The next thing that was done, was to simplify all these poverty percentages by creating 5 poverty groups that were created based on the bins of the histogram. This was done in the table listed in the data section previously.

Using these new groups, we can look at some maps using the choropleth package in R (code can be found on GitHub (Roberts, 2023)). This helps us to see how poverty percentages are distributed around the state of Oklahoma. Similarly, we can plot maps using other variables such as: Number of AA meetings, meeting density per square mile, per capita income, percentage of county that is rural, number of people addicted to meth (based on a rate of 0.4 given in the National Institute of Drug Abuse journal article (National Institute on Drug Abuse, 2013), and drug overdose rate (Oklahoma

State Department of Health (OSDH), 2016), poverty percent, poverty group, number in drug court, drug court completion rate, and Gini index by county. These will be useful if we find that that there is not much of a correlation between poverty and drug addiction, thus allowing us to use the model with a different variable in the place of percent of county in poverty. Several maps will be shown in the following figures. By finding the data for each individual county, we can use poverty percent as a place holder for county or some other variable or variables in our table that is given for each of the counties. A map with the county names included in Figure 7.15 to be used for comparisons to those maps in which county name is not shown.



Figure 7.15: Map Counties (of County Commissioners of Oklahoma , ACCO)

Figures 7.16 and 7.17 present a set of visualizations describing poverty and income inequality across Oklahoma counties and how they are distributed and categorized geographically. The colors on these maps are broke into 5 groups to correlate with the 5 poverty groups. The first panel shows a histogram of county-level poverty percentages, grouped into five bins, illustrating that most counties fall within the 13% to 17.5% range. The second panel maps the percent of individuals in poverty by county, with counties such as Choctaw, Adair, and McCurtain exhibiting higher poverty levels. On the second set of maps, The first panel displays Gini Index values by county, representing income inequality, with Nowata, Kay, and Okfuskee showing higher levels of disparity. The second panel translates the raw poverty percentages into five discrete poverty groups, visually aligning with the color scale in the second panel on the first set of choropleth maps.

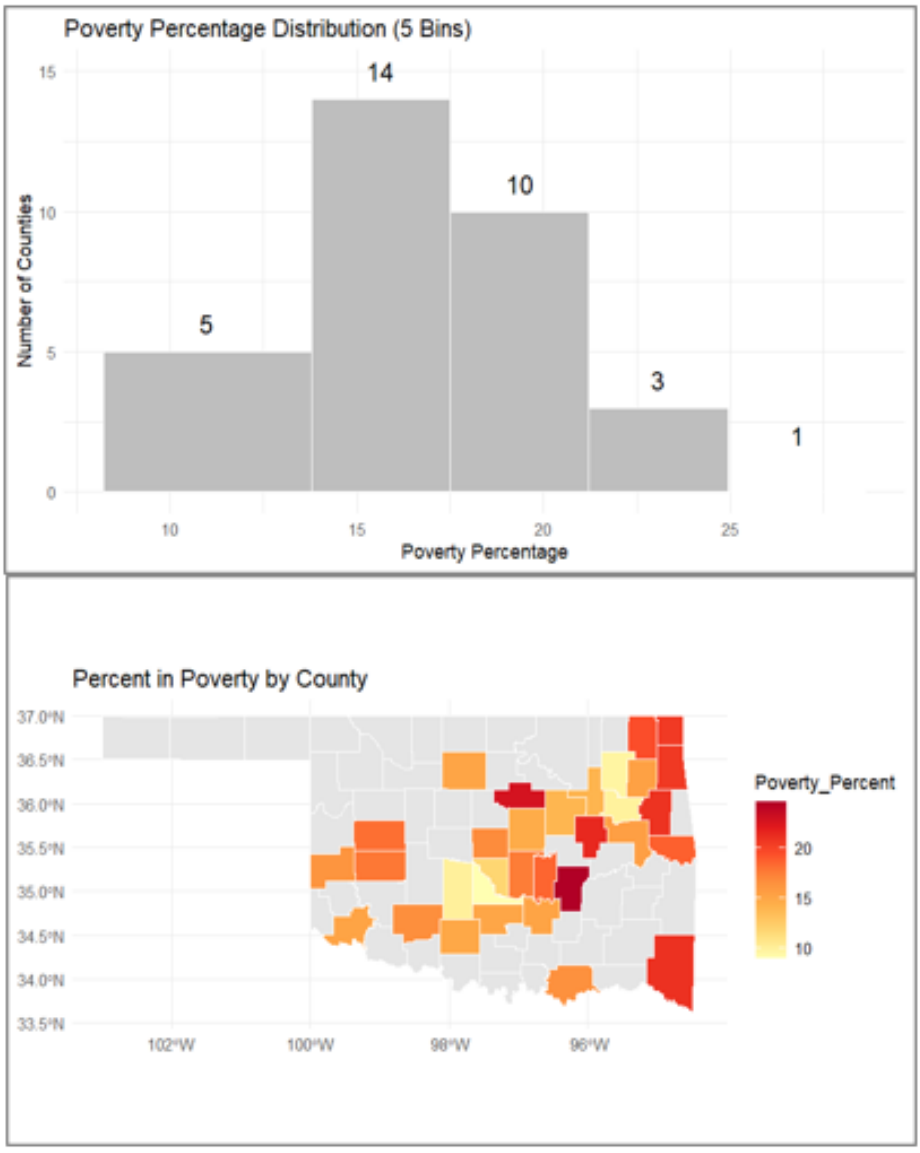


Figure 7.16: Poverty Percentage Histogram and Poverty Percentage by County

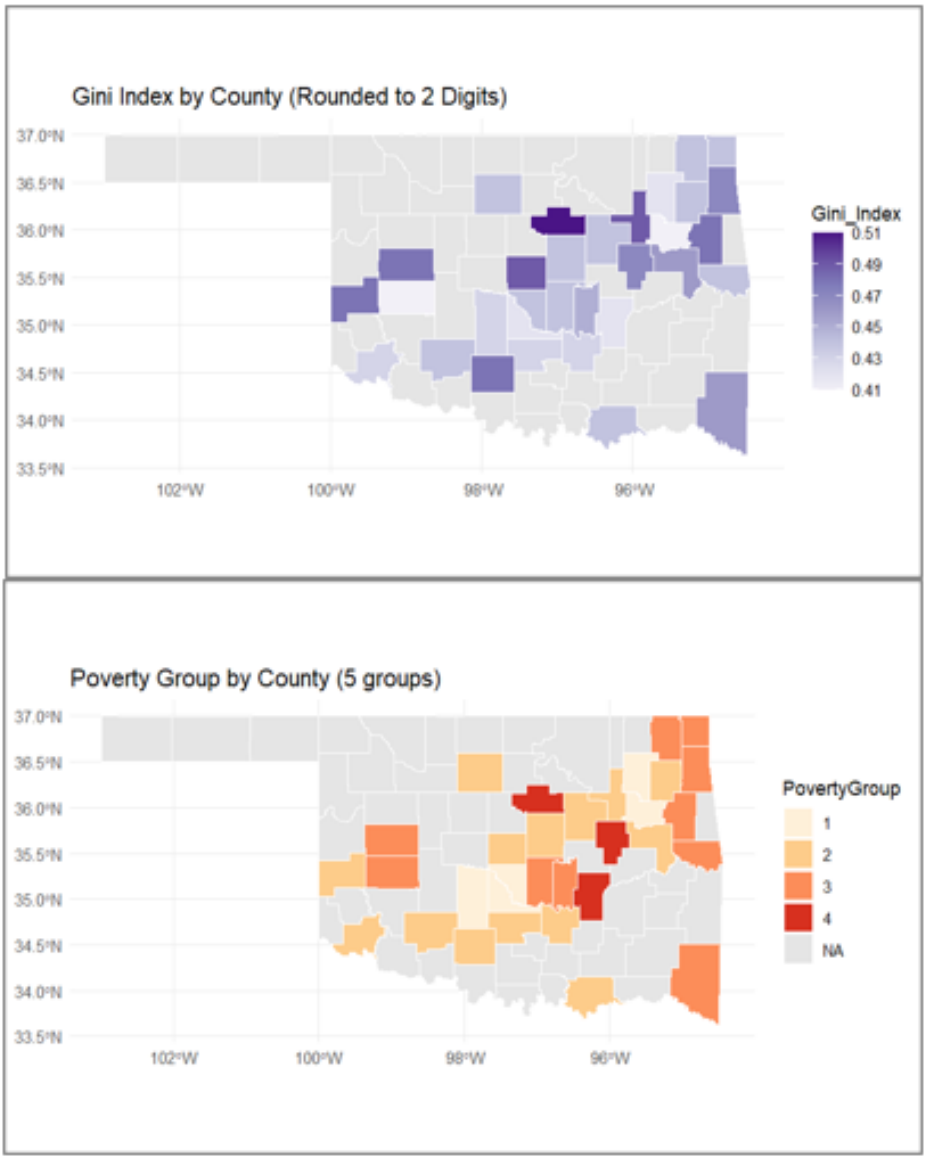


Figure 7.17: Gini Index and Poverty Group by County

Figures 7.18 and 7.19 present four choropleth maps visualizing percent rural, drug court participation, population size, and overdose mortality across Oklahoma counties. The first set of maps shows the percentage of land classified as rural by county, with Pushmataha, Harper, and Adair counties having the highest proportions, each exceeding 99%. The second map presents population by county, highlighting Oklahoma, Tulsa, and Cleveland counties as the areas with the largest populations. The first map on the second set of charts illustrates the number of individuals in drug court, where Tulsa, Payne, and Wagoner counties report the highest enrollment counts.The second map shows the number of drug overdose deaths per county, with Oklahoma and Muskogee counties having the highest totals. Taken together, these maps reveal contrasts between county characteristics, service participation, and health outcomes. Counties with larger populations tend to report higher drug court participation and overdose counts, while more rural counties often appear with fewer participants and lower reported totals.

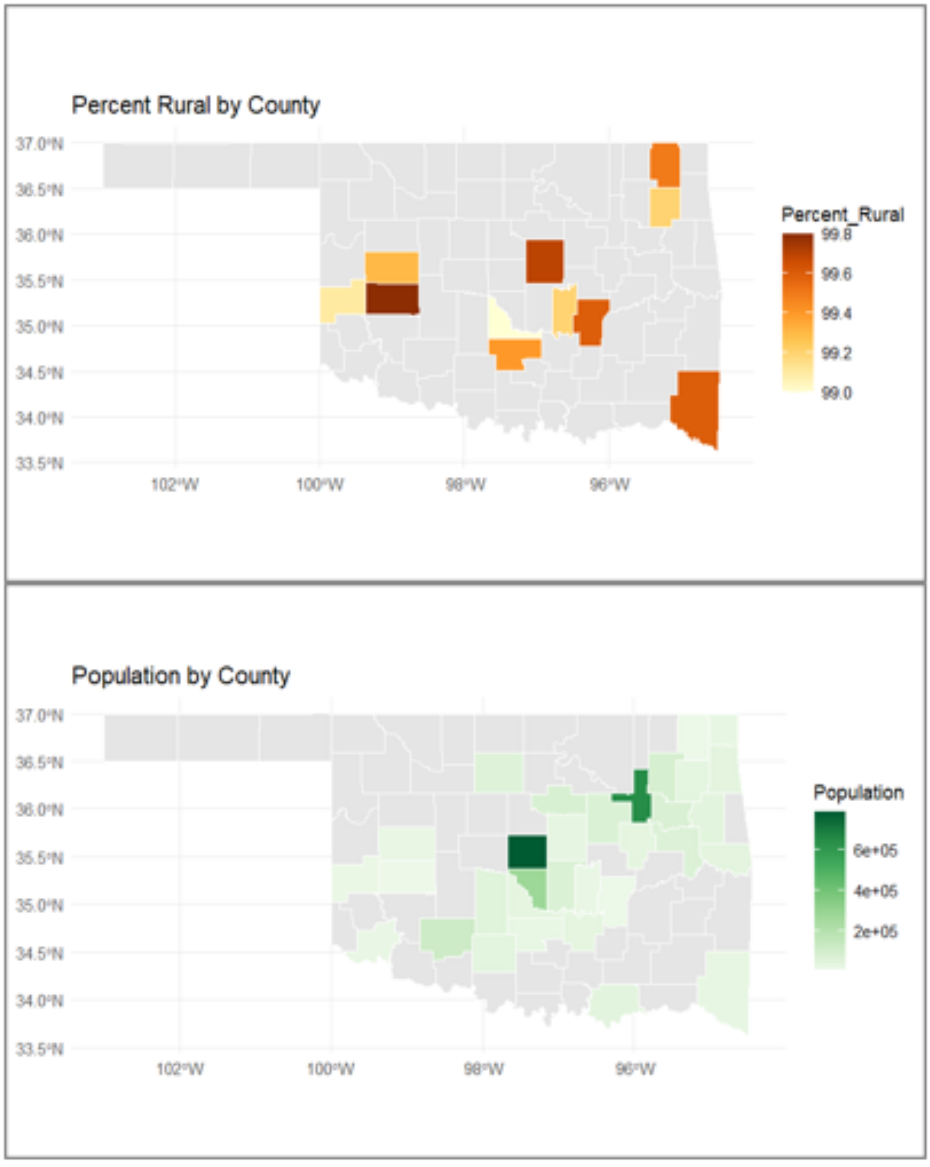


Figure 7.18: Percent Rural and Population by County

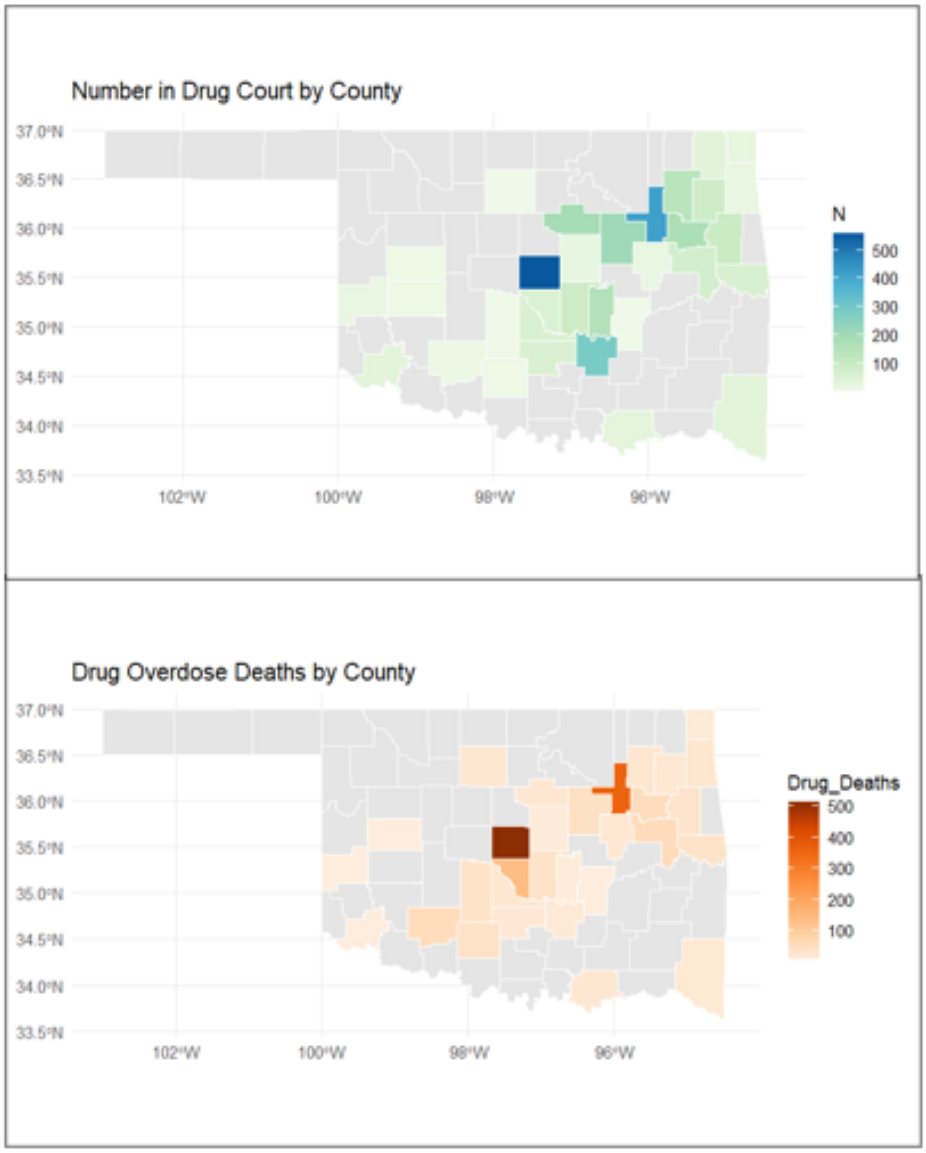


Figure 7.19: Number in Drug Court and Drug Deaths by County

Figures 7.20 and 7.21 present four choropleth maps depicting treatment-related metrics across Oklahoma counties. The first map displays drug court completion rates, with Craig, Ottawa, and Washington counties in the northeast exhibiting the highest values, while much of the south and west show little or no data. The bottom map represents the percentage of the population addicted to methamphetamine, highlighting Haskell, Pittsburg, and McCurtain counties in the southeast as having the highest methamphetamine exposure rates. The top map on the next set of charts illustrates the number of AA meetings, where Oklahoma County and Tulsa County stand out with the greatest counts. The map in the bottom of this figure shows the meeting density per square mile, with Oklahoma County showing the highest density, followed by Payne County, while most rural regions display extremely low values.

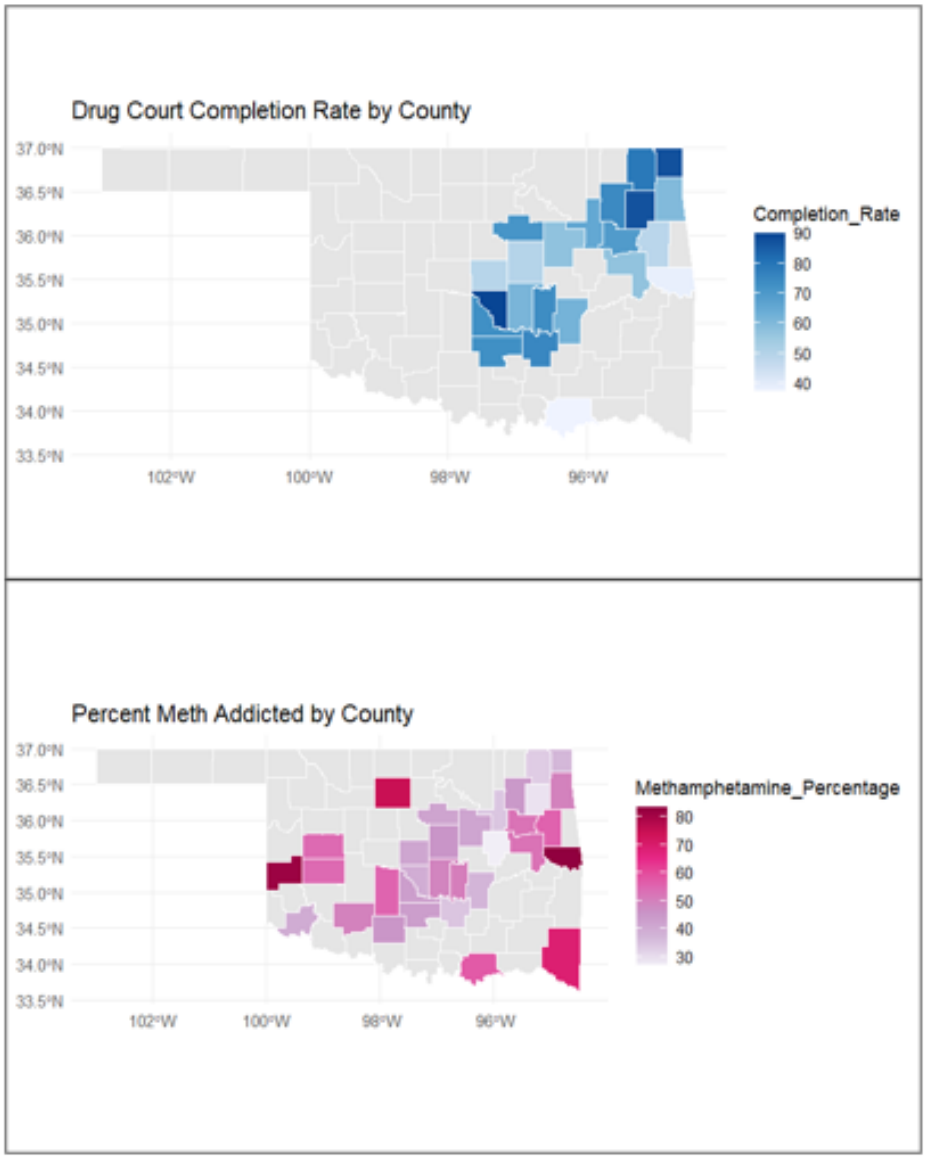


Figure 7.20: DC Completion Rate and Percent Meth Addicted by County

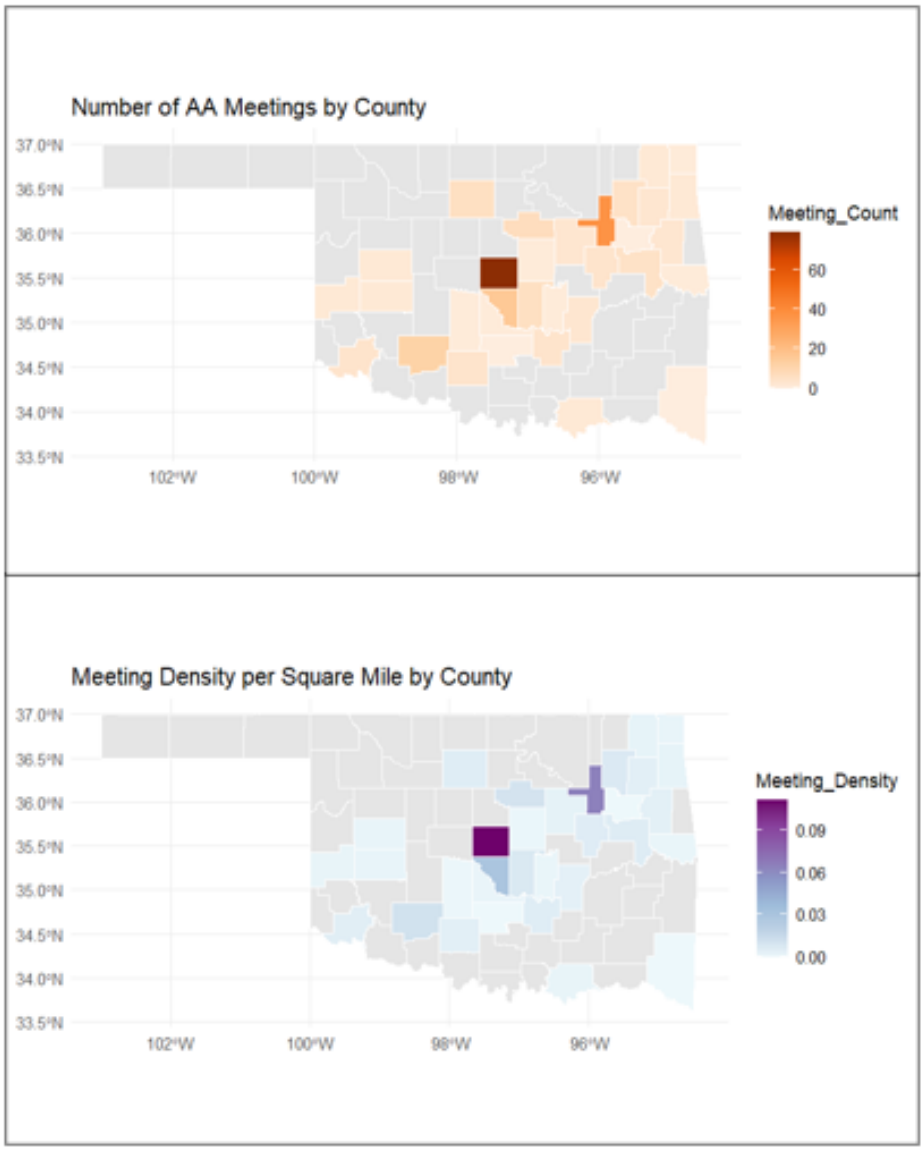


Figure 7.21: Number AA Meetings and Meeting Density by County

One of the things to mention on the map of AA meeting locations is that Oklahoma County and Tulsa County were outliers in the data set with 79 and 37 meetings respectively. The other counties that fell into the 6-79 category ranged between 6 and 16 meetings. We also found that it might be more beneficial to look at how many meetings there were per area in square mile. Finally, we wanted to see what the top 10 counties were with respect to poverty, overdose rates, per capita income, meeting density, percent of county that is rural, and number of people addicted. This was done using pivot tables in excel and those charts can be in Figures 7.22, 7.23, 7.24,7.25, 7.26, and 7.27.

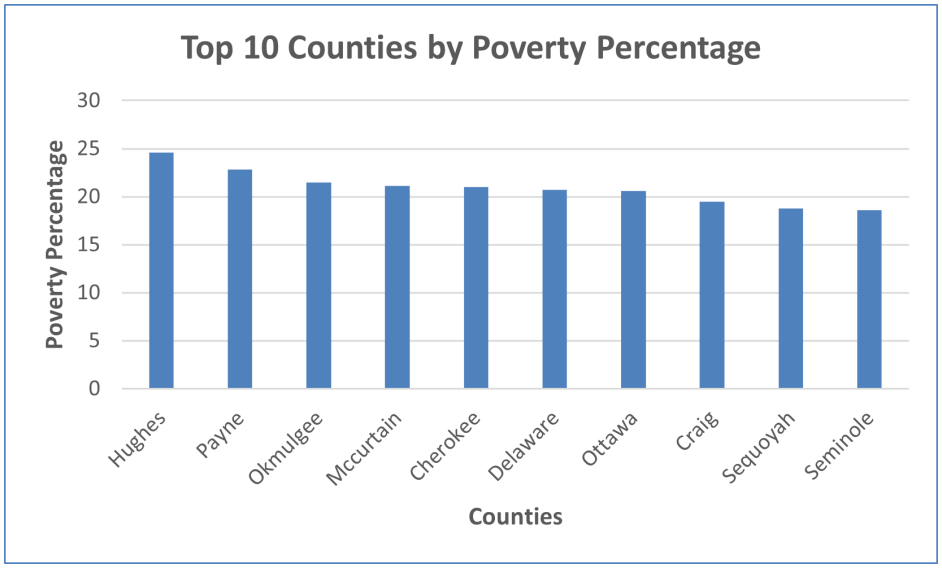


Figure 7.22: Top 10 Counties by Percentage County in Poverty

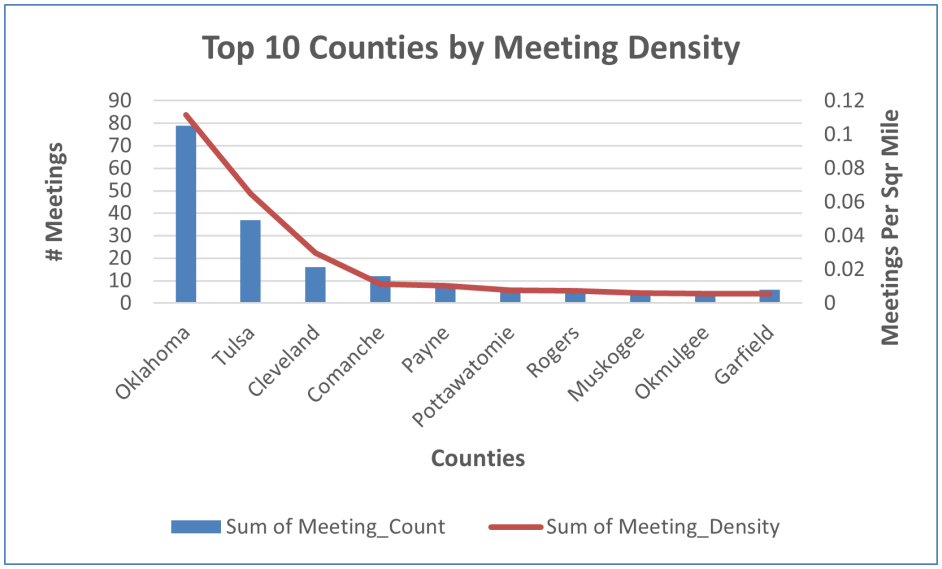


Figure 7.23: Top 10 Counties by Meeting Density

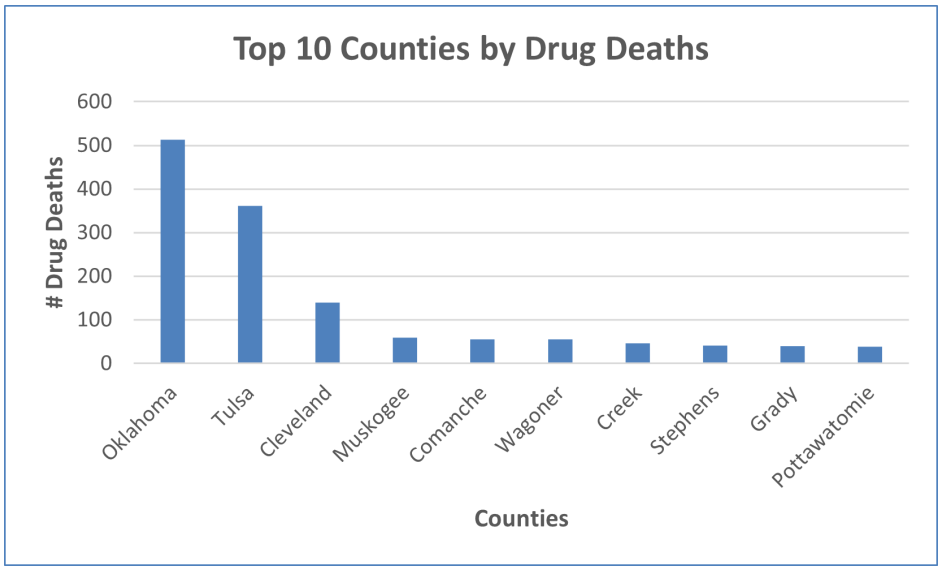


Figure 7.24: Top 10 Counties by Drug Deaths

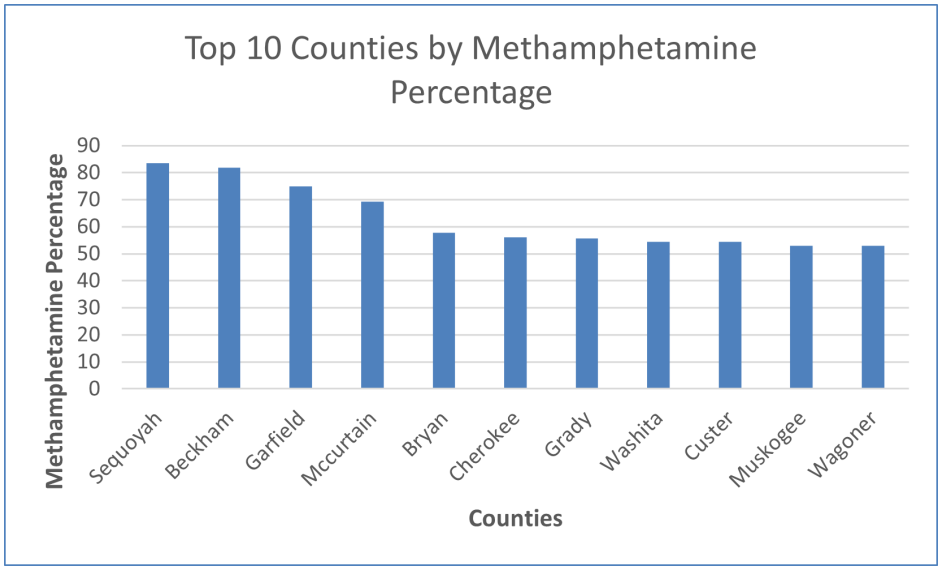


Figure 7.25: Top 10 Counties by Number of People Addicted to Methamphetamines Initially

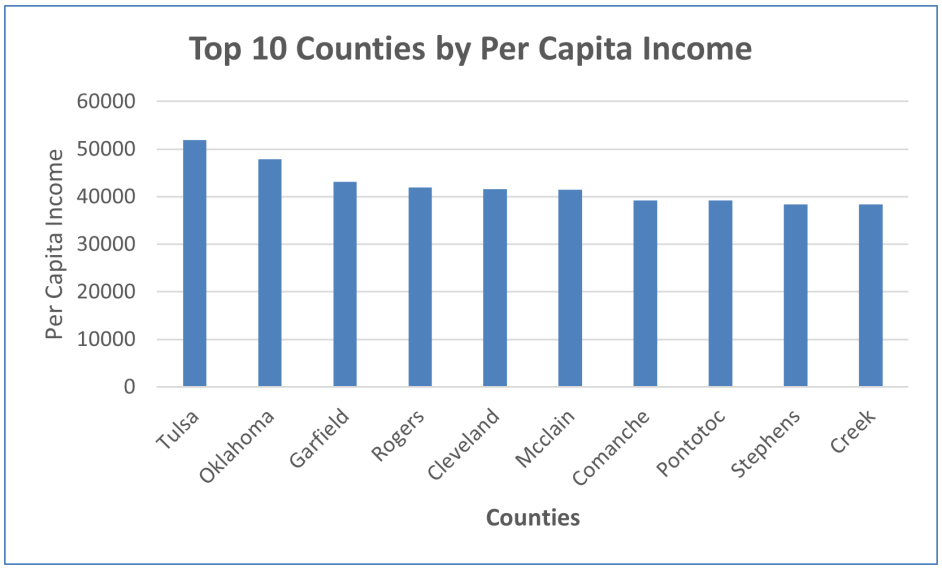


Figure 7.26: Top 10 Counties by Per Capita Income

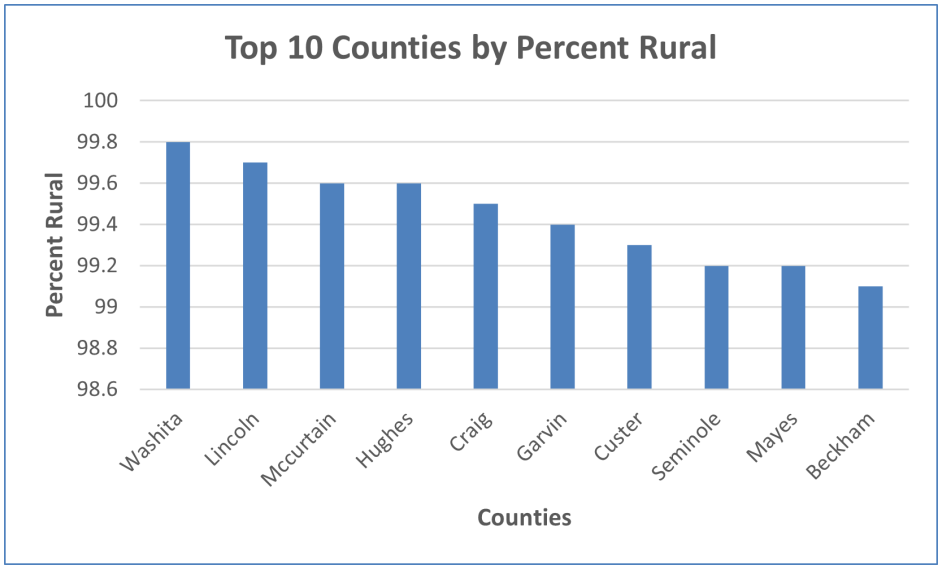


Figure 7.27: Top 10 Counties by Percent Rural

Next, we looked at the correlations between some of these variables that looked to form a pattern in the choropleth maps. Both the number of drug deaths correlation plot in Figure 7.28 and the per capita income correlation plot in Figure 7.29 showed high correlations with several of the other variables. With the rural counties not having access to the drug court program, it skews several of the numbers so we will not use percent rural or percent in poverty which is highly correlated with that. When looking at income and poverty it appears that per capita income shows much higher correlations with people addicted and number in drug court than looking at percentage of county in poverty did. For this reason, we concluded that the p variable in our PDE model should now represent per capita income.

Although the poverty percentage provided a straightforward stratification variable for model calibration, it captures only relative disadvantage within counties. Per capita income, on the other hand, allows for a more granular and continuous measure of economic well-being. Therefore, in the conclusion analysis, we pivot to income as the explanatory axis to better connect the simulation results with distributive fairness and policy implications.

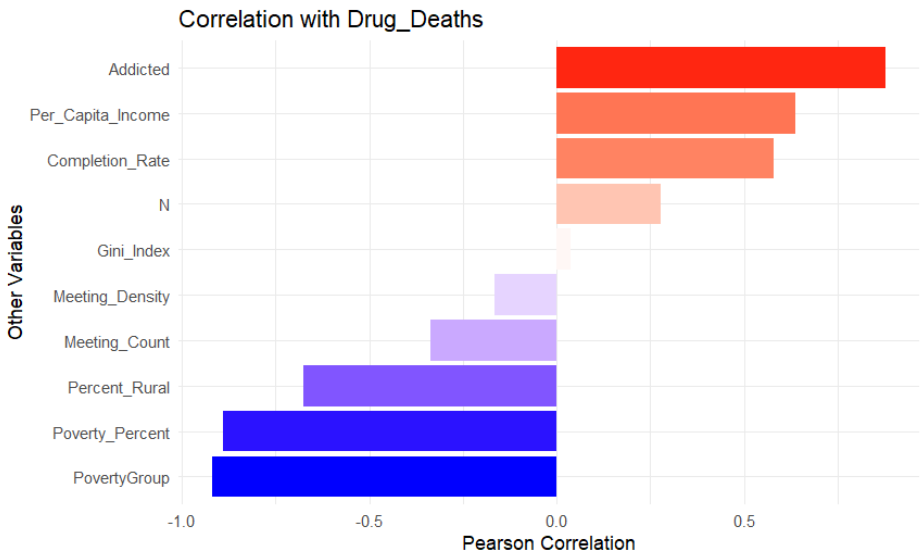


Figure 7.28: Drug Deaths Correlations

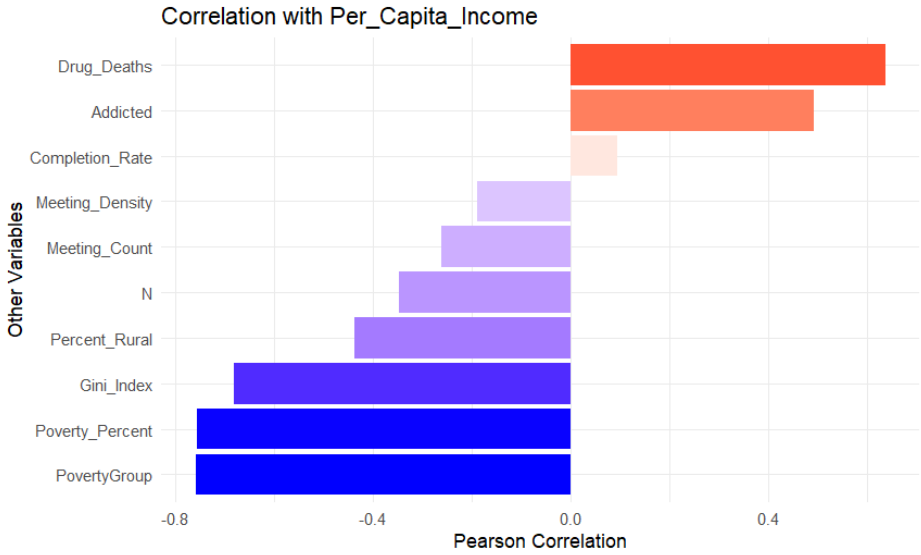


Figure 7.29: Per Capita Income Correlations

Finally, when looking at the choropleth maps, there seems to be a pattern or higher concentration of people that are addicted or in drug court in four counties: Oklahoma county, Cleveland county, Tulsa county, and Comanche county. The counties surrounding those counties seem to be the next highest in density for both addiction and number of people in drug court (if that county has a drug court program, i.e., is not whited out). From this we concluded that perhaps the use of a hotspot model would be beneficial to

use a hotspot spatial distribution model in order to see how the density of each of these classes looks over each of the counties. This would be useful in order to tie the percent rural and/or location of AA/NA meetings in with the poverty percentages.

7.4 Conclusions

The process of trying to build a model to show how methamphetamine addiction is spread as an epidemic over time within differing percentages of a population being impoverished is more difficult than simply using a poverty structured SIRV2 epidemic model, statistical analysis, or even hotspot spatio-temporal analysis. Without knowing an exact interaction coefficient, guesses must be made based on data or trends in the data over time. Comparisons could be made between models using various interaction coefficients to do a sort of sensitivity analysis on the behavior of the model. One fact that seemed pretty consistent was that there is a higher density of people on methamphetamines in the larger counties that are surrounded by counties with lower per capita income or higher percentage rural. Also, there seems to be a higher meeting density in the larger counties. While there appears to be a connection between meeting location and number of people addicted, there is no way of knowing if the meetings are there due to population size or because of a higher density of addicted people. Some of these things could be made clearer with more available data or by using more than one-year worth of drug court and poverty data to make better comparisons over time. Longitudinal drug court records could essentially refine relapse rates.

Chapter 8

Machine Learning Techniques with Fairness for Prediction of Completion of Drug and Alcohol Rehabilitation

8.1 Introduction

Substance abuse is one of the leading causes for mental illness and these issues are dealt with in the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD) (dsm, 2021). 36.2% of adults age 18-25 had a mental illness, 29.4% for age 26-49 and 13.9% 50 or older with 11.6% being serious mental illness for age 18-25, 7.6% for 26-49 years old and 3.0% for 50 years or older (National Survey on Drug Use and Health, 2022). The COMPAS assessment has been used in criminal sentencing to predict whether an offender will re-offend, however, it has proven to be racially biased. COMPAS is the Correctional Offender Management Profiling for Alternative Sanctions. It was developed in 1998, uses a recidivism risk scale, and has been used since 2000. It predicts a defendant's risk of committing a misdemeanor or felony within 2 years of assessment for 137 features about an individual and the individual's past criminal record (Dressel and Farid, 2018). The COMPAS assessment incorrectly predicted that whites would reoffend at a rate of 47.7% which was twice the rate of blacks at 28.0%. It favored white defendants over blacks. Its accuracy for white defendants was 67% and 63.8% for black defendants (Dressel and Farid, 2018). For this reason, considering fairness in these types of assessments is important. Also, taking our approach as to looking at whether an offender would complete rehab instead of going to prison is important as an alternative to sending them to prison due to overcrowding in prisons especially in Oklahoma. Oklahoma itself

ranks third if one looks at the world's incarceration rate considering every state as a country (Wildra and Herring, 2021a). For the scope of this Chapter, treatment episode data from the Substance Abuse and Mental Health Services website (Substance Abuse and Mental Health Services Administration, 2021) is used to analyze whether a person completed a drug and alcohol rehabilitation program, the number of previous treatments a person has been to when entering treatment, and then the concatenation of both variables.(Chawla et al., 2002) thoroughly reviewed
SMOTE: Synthetic Minority Over-Sampling Technique for both the continuous and Nominal case. The nominal case is what we will use in this work. The algorithms and use of the value distance metric were described for SMOTE-N and SMOTE-NC and how these relate to the ROC curve. Dual and three-way interactions of both additive and multiplicative types are discussed in (Veenstra, 2011) and the intersection of bias terms is shown in this work. Demographic Parity, Disparate Impact, Statistical Parity Difference, Equal Opportunity, Equalized Odds, and Min-Max Fairness were all discussed in length in (Irfan et al., 2023). They all compared models such as Support Vector Classifiers, Gaussian Process Classifiers, Gaussian Naïve Bayesian, and Linear Discriminant Analysis. The Worst-case fairness metrics such as the demographic parity ratio, the disparate impact ratio, the conditional statistical parity ratio, the equal opportunity ratio and the equal odds ratio, as well as a multiclass example were shown in (Ghosh et al., 2021). These fairness measures can be used when there are multiple classes or when the reweighting or intersectionality has been done and there are more than two ratios to evaluate between.

The fairness methods in ML from the literature review of several papers were used in this chapter. Disparate impact, statistical parity difference, equalized opportunity, equalized odds, and the idea of intersectional fairness was expanded upon. This Chapter is organized as follows: in section 2 we discuss the methodology. In sections 3, 4 and 5 we discuss issues of data cleaning, encoding, balancing and data sensitivities and distributions. Section 6 discusses fairness measures applied to our data. In sections 7 and 8 we explore several machine learning models and interpretation of the results. Sections 9 and 10 discusses reweighting and new fairness calculations. Finally sections 11 and 12 discuss the conclusions and future work.

8.2 Research Questions

- **What are the optimal machine learning models for predicting drug and alcohol rehabilitation outcomes in Oklahoma?**

  - Identifies the most accurate models for predicting rehabilitation outcomes, balancing prediction accuracy and efficiency.

- **How do fairness measures affect the accuracy and equity of these machine learning models?**

  - Examines the impact of fairness measures on model performance, focusing on fairness between demographic groups.

- **What role do intersubsectional bias variables play in mitigating biases in rehabilitation predictions?**

  - Analyzes how intersubsectional bias variables reduce prediction bias, ensuring equitable performance across subgroups.

8.3 Methodology

Various machine learning algorithms were used on this data with a focus on kernel methods for support vector machines. With these data being highly categorical in nature, encoding techniques were used to transform the data to make it more manageable. The data were balanced to make the predictions more accurate and the missing data was imputed. Fairness measures were compared before and after weighing the variables using two- and three-way interactions with chi squared significance and internationalizing the nine bias variables. The nine bias variables explored were gender, veteran status, marital status, education, age, employment, pregnancy status, race, and ethnicity. Four kernels were compared; linear, polynomial, radial basis function, and sigmoid. These were tried on a range of degrees, c values, gammas, and r values. Decision trees, random forests, and neural networks were compared. The models were then compared with each other to see which models outperformed each other.

8.4 Data Clean Up Process

The data set was filtered down to Oklahoma and down to 30-day rehabilitation centers only. A calculated column was created called COMPLETED in which a value of 'COMPLETE' was taken if the REASON variable took on a value of 1 and 'INCOMPLETE' if REASON took on a value of anything else. The reason code translations can be seen in the following figure.

Value	Label
1	Treatment completed
2	Dropped out of treatment
3	Terminated by facility
4	Transferred to another treatment program or facility
5	Incarcerated
6	Death
7	Other

Figure 8.1: Reason Codes

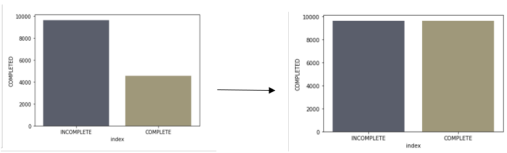


Figure 8.2: Completed SMOTE Applied

Some user defined functions were created to clean up the data. One removes unneeded columns that just have a constant value in them or had an identification number for the CASEID, admit year (ADMYR), REGION and STFIPS could be removed since the data was filtered for Oklahoma. STFIPS was the FIPS code for each state. This was filtered for 40 for Oklahoma as mentioned previously.

8.5 Encoding and Balancing Data

Then, three data sets were created. One was for the completed predictions (whether someone completed treatment or not), one was for the reasons predictions (based on the reason code for why they left treatment), and one was for the no-priors (number of times in treatment) predictions. The COMPLETED_NOPRIOR data set was created from a combination of these data sets. The final data sets used in the analysis are the COMPLETED, NOPRIOR, and COMPLETED_NOPRIOR data sets. Each data set went through an encoding and data balancing processes. One hot encoding was used to encode the categorical variables into new binary variables. Each class within that variable is created into a new column which takes on the value of 1 if it appears for that record and 0 otherwise.

Next, we look at how the data is imbalanced in all three data sets. We use the SMOTEN (Synthetic Minority Over-Sampling Technique Nominal) package in Python which uses K-Nearest neighbors (Gajawada, 2021) algorithm to balance the data. See the following figures. This is done for the three data frames used in prediction.

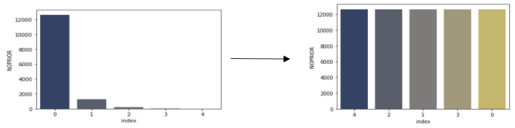


Figure 8.3: NOPRIOR SMOTE Applied

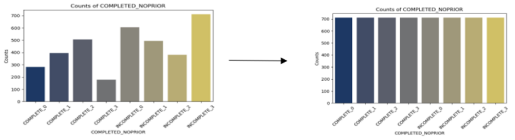


Figure 8.4: Completed_NORPIOR SMOTE Applied

Typically SMOTE is used which helps to improve the prediction accuracy. It distributes the instances of the majority class and the minority class equally. SMOTE technique increases the predictive accuracy over the minority class by creating synthetic instances of that minority class, (Jishan et al., 2015). SMOTEN extends SMOTE for nominal features by getting the nearest neighbors by using the modified version of the Value Difference Metric which looks at the overlap of feature values over all feature vectors. A matrix of features for all feature vectors is created and the distance between those features is defined in the following equation:

$$\delta\left(V_1, V_2\right) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k \tag{8.1}$$

Where V_1 and V_2 are the corresponding feature values and C_1 is the total number of times V_1 occurs and C_{1i} is the total occurrences of feature V_1 for class i. The same holds for V_2, C_2, and C_{2i}. k is some constant that is typically set to 1. This gives the matrix of value differences for each nominal value in the set of feature vectors(Chawla et al., 2002). An example given in the following figures.

Let F1 = Red , Orange, Yellow, Green, Blue

Let F2 = Red, Purple, Yellow, Teal, Brown

Let F3 = Black, Orange, Yellow, Green, Brown

SMOTEN would create the following:

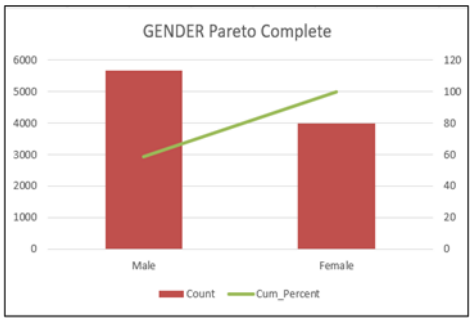
FS= Red, Orange, Yellow, Green, Brown

Figure 8.5: SMOTEN Example

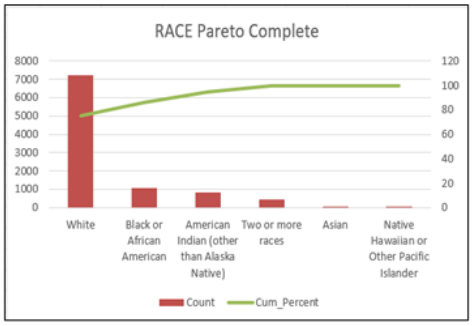
8.6 Data Sensitivities/Distributions

8.6.1 Pareto Charts

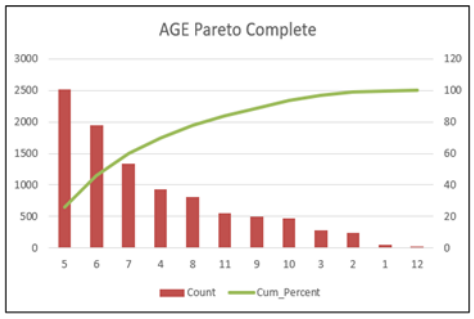
Nine variables were chosen and distributed by Pareto charts to see where bias occurred. These variables were gender, race, age, ethnic, veteran status, education status, marital status, employment status, and pregnancy status. These are shown for completion rehab outcomes in the Figures 8.6a- 8.8c.



(a) Gender Pareto

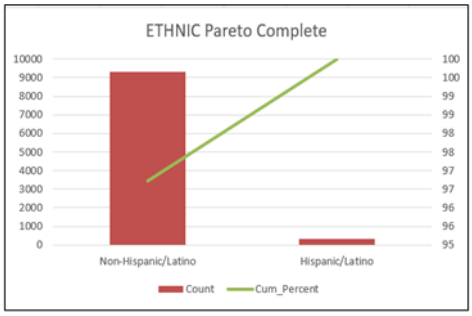


(b) Race Pareto

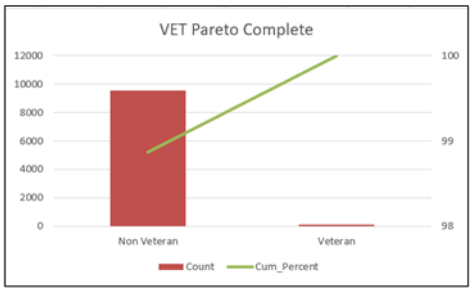


(c) Age Pareto

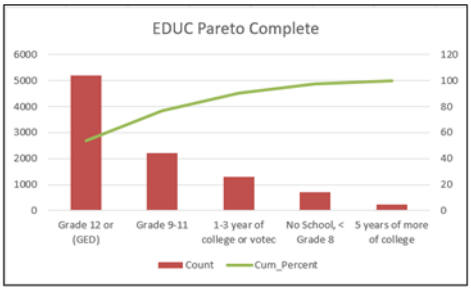
Figure 8.6: Pareto Charts for Gender, Race, and Age



(a) ETHNIC Pareto

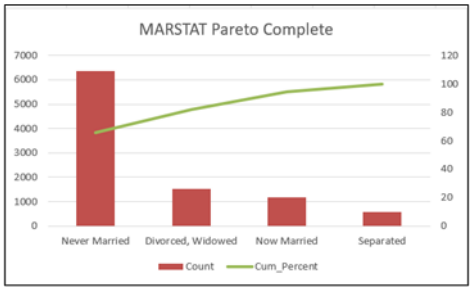


(b) VET Pareto

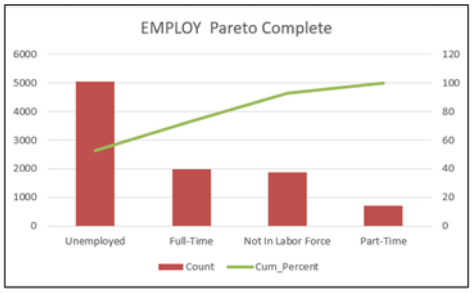


(c) EDUC Pareto

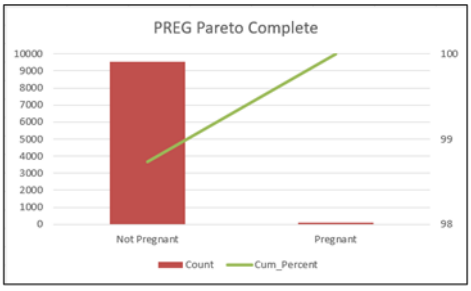
Figure 8.7: Pareto Charts for Ethnicity, Veteran Status, and Education Status



(a) MARSTAT Pareto



(b) EMPLOY Pareto



(c) PREG Pareto

Figure 8.8: Pareto Charts for Marital Status, Employment Status, and Pregnancy Status

8.6.2 Bucketized Categories

Based on the Pareto charts previously shown, each variable was bucketized into dichotomous variables. Variables such as gender, pregnancy, ethnicity, and veteran status were already dichotomous so they did not change. Race became either white or non-white, employment status became employed or not employed, age became under 40 or 40 plus, marital status became never married or married/previously married, and education status became college or no college.

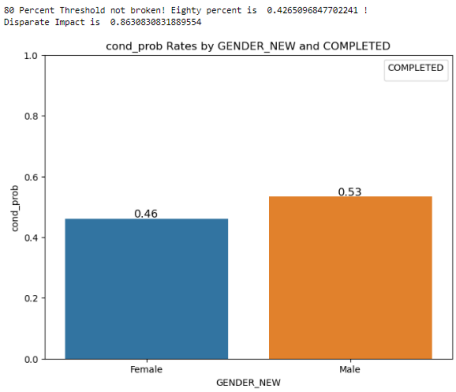
8.7 Fairness Measures

8.7.1 Disparate Impact

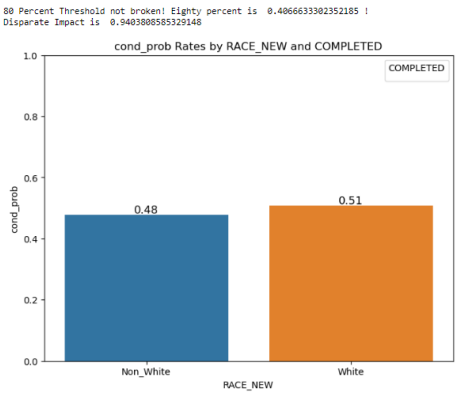
To find where the discrimination occurs, we first look at disparate impact in the dichotomous variables.

$$DI = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \tag{8.2}$$

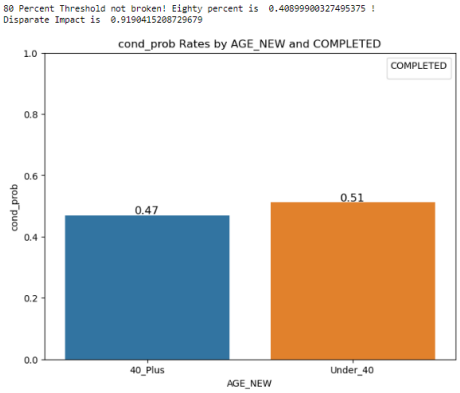
This compares the proportion of individuals receiving a favorable outcome for a privileged and underprivileged group. The closer to 1 it is, the more fair it is. $\hat{Y}$ is the model predictions and A is the protected attribute, with 0 being the underprivileged class and 1 being the privileged class (Irfan et al., 2023). The resulting disparate impact charts for the completed model can be seen in the following figures. Typically the 80% rule is followed for the threshold to be considered discriminatory (Ghosh et al., 2021).



(a) DI Dichotomous Gender

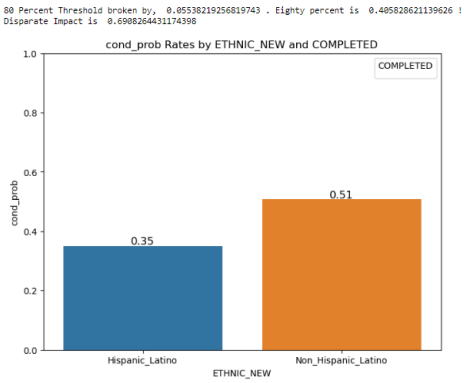


(b) DI Dichotomous Race

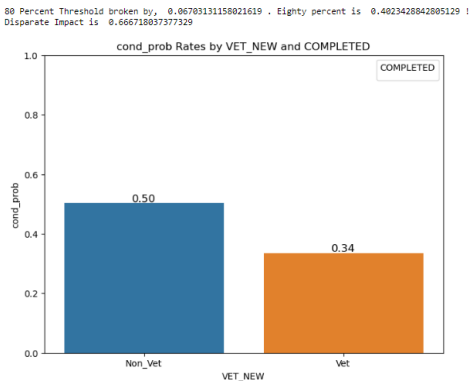


(c) DI Dichotomous Age

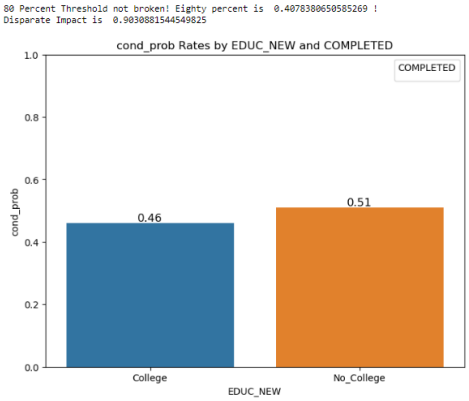
Figure 8.9: DI Dichotomous Charts for Gender, Race, and Age



(a) DI Dichotomous ETHNIC

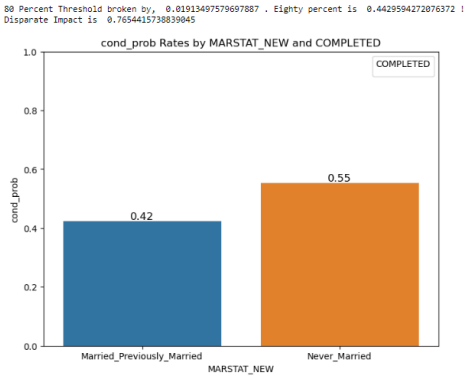


(b) DI Dichotomous VET

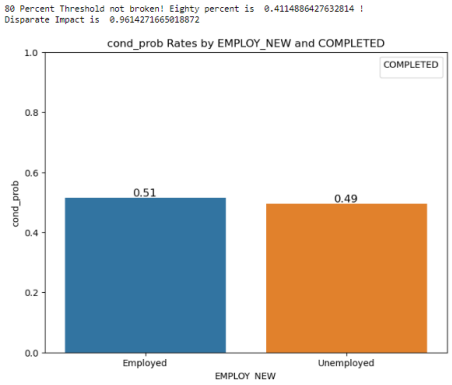


(c) DI Dichotomous EDUC

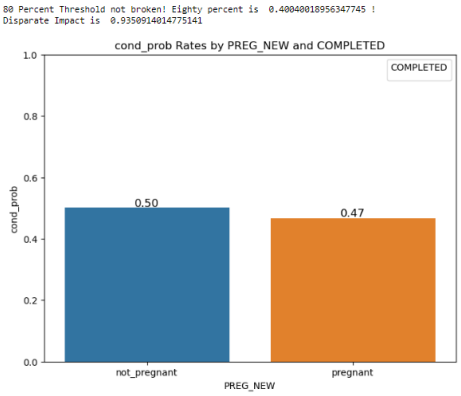
Figure 8.10: DI Dichotomous Charts for Ethnicity, Veteran Status, and Education Status



(a) DI Dichotomous MARSTAT Pareto



(b) DI Dichotomous EMPLOY



(c) DI Dichotomous PREG

Figure 8.11: DI Dichotomous Charts for Marital Status, Employment Status, and Pregnancy Status

8.7.2 Disparate Impact – Multiclass

Next we look at the multiclass case in which we find the disparate impact for each class and divide the minimum conditional probability when the target matches the outcome by the maximum conditional probability when the target matches the outcome. In the following formula c represents the class of interest and a is the different values within the protected attribute such as gender, age, race, etc.

$$\text{DIMulticlass}_c = \frac{\min\limits_{a \in A} P(\hat{Y} = c \,|\, A = a)}{\max\limits_{a \in A} P(\hat{Y} = c \,|\, A = a)} \tag{8.3}$$

The outputs for noprior model for each of the variables can be seen in the following table.

Table 8.1: Disparate Impact Analysis Multiclass

Var	Class	DI	80% Broke
GENDER	0	0.43	UNFAIR
	3	0.78	UNFAIR
	2	0.02	UNFAIR
	1	0.81	FAIR
AGE	0	0.56	UNFAIR
	3	0.01	UNFAIR
	2	0.65	UNFAIR
	1	0.97	FAIR
VET	0	0.28	UNFAIR
	3	1.00	FAIR
	2	0.17	UNFAIR
	1	0.33	UNFAIR
EDUC	0	0.80	UNFAIR
	3	1.00	FAIR
	2	0.38	UNFAIR
	1	0.73	UNFAIR

Continued Next Page

Table 8.1 Continued from previous page

Var	Class	DI	80% Broke
MARSTAT	0	0.58	UNFAIR
	3	0.42	UNFAIR
	2	0.67	UNFAIR
	1	0.77	UNFAIR
EMPLOY	0	0.36	UNFAIR
	3	1.00	FAIR
	2	1.00	FAIR
	1	0.75	UNFAIR
RACE	0	0.92	FAIR
	3	0.00	UNFAIR
	2	0.34	UNFAIR
	1	0.81	FAIR
ETHNIC	0	0.64	UNFAIR
	3	1.00	FAIR
	2	0.43	UNFAIR
	1	0.20	UNFAIR
PREG	0	0.30	UNFAIR
	3	0.17	UNFAIR
	2	1.00	FAIR
	1	0.52	UNFAIR

8.7.3 Statistical Parity Difference

For the completed model we next look at Statistical Parity Difference (SPD). The closer the result is to 0, the more fair it is.

$$SPD = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1), \tag{8.4}$$

where $\hat{Y}$ is the models predictions, $A = 0$ is the protected attribute for the unprivileged class and $A = 1$ is the protected attribute for the privileged

class. (Irfan et al., 2023). The results can be seen for each variable in the following table.

Table 8.2: Disparate Impact Analysis

Var	**Class**	**DI**	**80% Broke**
Var	Class	SPD	SPD Diff
GENDER	Male	0.53	0.07
	Female	0.46	
AGE	Under 40	0.51	0.04
	40 Plus	0.47	
VET	Veteran	0.34	0.16
	Non-Veteran	0.50	
EDUC	No College	0.51	0.05
	College	0.46	
MARSTAT	Married/Previously Married	0.55	0.13
	Never Married	0.42	
EMPLOY	Unemployed	0.49	0.02
	Employed	0.51	
RACE	White	0.51	0.03
	Non-White	0.48	
ETHNIC	Hispanic/Latino	0.51	0.16
	NonHispanic/Latino	0.35	
PREG	Pregnant	0.47	0.03
	Not Pregnant	0.50	

8.7.4 Statistical Parity Difference – Multiclass

This was done again for each class of noprior and the results can be seen in the following table. The max threshold is the maximum threshold for which a fairness is achieved for the spd value for each class. If at least one of the classes has an SPD of zero, then SPD is satisfied for that variable, otherwise it is not.

Table 8.3: Statistical Parity Difference (SPD) Analysis

Variable	Class	SPD	Max Threshold	At least one SPD = 0
GENDER	0	0.27	0.26	Not Satisfied for All Classes
	3	0.07	0.06	
	2	0.29	0.26	
	1	0.05	0.01	
AGE	0	0.17	0.16	Not Satisfied for All Classes
	3	0.30	0.26	
	2	0.12	0.11	
	1	0.01	0.01	
VET	0	0.63	0.61	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.21	0.21	
	1	0.17	0.16	
EDUC	0	0.05	0.01	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.29	0.26	
	1	0.08	0.06	
MARSTAT	0	0.14	0.11	Not Satisfied for All Classes
	3	0.18	0.16	
	2	0.10	0.06	
	1	0.06	0.06	
EMPLOY	0	0.42	0.41	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.00	0.00	
	1	0.08	0.06	
RACE	0	0.02	0.01	Not Satisfied for All Classes
	3	0.35	0.31	
	2	0.32	0.31	
	1	0.05	0.01	
ETHNIC	0	0.14	0.11	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.32	0.31	
	1	0.20	0.16	
PREG	0	0.58	0.56	Satisfied for At Least One Class
	3	0.21	0.21	
	2	0.00	0.00	
	1	0.12	0.11	

8.7.5 Equal Opportunity

When the true positive rate (TPR) is the same for both the privileged and unprivileged groups, it is considered fair (Irfan et al., 2023).

$$P\left(\hat{Y} = 1 \mid Y = 1, A = 0\right) = P\left(\hat{Y} = 1 \mid Y = 1, A = 1\right) \tag{8.5}$$

$$\text{TPR}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{8.6}$$

$$\text{EqOpp}_{\text{diff}} = \max_i(\text{TPR}_i) - \min_i(\text{TPR}_i) \tag{8.7}$$

This fairness measure was ran at four different C values with the default values for gamma and r, where $gamma = scale = 1/n$ and $r = 0$. C was tried at .1, 1, 10, and 100. The optimal results are shown in the following table. Notice again that this is tried at different thresholds. The difference is found and then the maximum threshold value for which the $EqOpp_{diff} < threshhold$ is recorded in the table for each model.

Next we do the same thing but for each class and take the max difference and the minimum difference to calculate the fairness in each class where c is in each class. The results can be seen in the following table. The TPR difference is shown in the table and this is the maximum value that the threshold can be taken at. Again this is shown for the optimal C value with the default values of gamma, r, and degree.

$$TPR_{sc} = \frac{TP_{sc}}{TP_{sc} + FN_{sc}} \tag{8.8}$$

$$TPR_{\text{diff}} = \max(TPR_{sc}) - \min(TPR_{sc}) \tag{8.9}$$

$$TPR_{sc} = \frac{TP_{sc}}{TP_{sc} + FN_{sc}} \tag{8.10}$$

$$FPR_{sc} = \frac{FP_{sc}}{FP_{sc} + TN_{sc}} \tag{8.11}$$

$$\text{Max_TPR}_{\text{diff}} = \max_c\left(\max_s(TPR_{sc})\right) \tag{8.12}$$

$$\text{Min_TPR}_{\text{diff}} = \min_c\left(\min_s(TPR_{sc})\right) \tag{8.13}$$

Table 8.4: Equal Opportunity - COMPLETED

Variable	Optimal C Value	Model	Max TPR	Min TPR	EqOpp TPR Diff	Fairness
GENDER	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
AGE	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0.23	0.77	UNFAIR
VET	0.1	Linear	0.99	0	0.99	UNFAIR
	10	Poly	0.99	0	0.99	UNFAIR
	10	RBF	0.99	0	0.99	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
EDUC	0.1	Linear	0.99	0	0.99	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
MARSTAT	0.1	Linear	1	0.7	0.3	UNFAIR
	10	Poly	1	0.7	0.3	UNFAIR
	10	RBF	1	0.7	0.3	UNFAIR
	1	Sigmoid	1	0.36	0.64	UNFAIR
EMPLOY	0.1	Linear	0.99	0	0.99	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
RACE	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0.47	0.53	FAIR
ETHNIC	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
PREG	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR

$$\text{EqOpp}_{\text{diff}} = |\text{Max}_{\text{TPR}_{\text{diff}}} - \text{Min}_{\text{TPR}_{\text{diff}}}| \tag{8.14}$$

Table 8.5: Equalized Opportunity - NOPRIOR

Variable	Optimal C Value	Model	TPR Difference	FPR Difference	Equalized Odds Diff	Fair/Unfair
Variable	C=Optimal	Model	Max TPR	Min TPR	Equalized Opp Diff	Fair/Unfair
GENDER	10	Linear	0.69	0.37	0.32	UNFAIR
	1	Poly	0.70	0.56	0.14	FAIR
	10	RBF	0.69	0.52	0.17	FAIR
	1	Sigmoid	0.65	0.40	0.25	UNFAIR
AGE	10	Linear	0.69	0.43	0.26	UNFAIR
	1	Poly	0.73	0.58	0.15	FAIR
	10	RBF	0.71	0.54	0.17	FAIR
	1	Sigmoid	0.67	0.44	0.23	UNFAIR
VET	10	Linear	0.50	0.00	0.50	UNFAIR
	1	Poly	0.63	0.00	0.63	UNFAIR
	10	RBF	0.50	0.00	0.50	UNFAIR
	1	Sigmoid	0.51	0.00	0.51	UNFAIR
EDUC	10	Linear	0.55	0.23	0.32	UNFAIR
	1	Poly	0.66	0.41	0.25	UNFAIR
	10	RBF	0.63	0.38	0.25	UNFAIR
	1	Sigmoid	0.54	0.32	0.22	UNFAIR
MARSTAT	10	Linear	0.50	0.49	0.01	FAIR
	1	Poly	0.62	0.62	0.00	FAIR
	10	RBF	0.60	0.56	0.04	FAIR
	1	Sigmoid	0.54	0.45	0.09	FAIR
EMPLOY	10	Linear	0.51	0.12	0.39	UNFAIR
	1	Poly	0.62	0.50	0.12	FAIR
	10	RBF	0.60	0.38	0.22	UNFAIR
	1	Sigmoid	0.51	0.25	0.26	UNFAIR
RACE	10	Linear	0.50	0.48	0.02	FAIR
	1	Poly	0.60	0.62	0.02	FAIR
	10	RBF	0.60	0.54	0.06	FAIR
	1	Sigmoid	0.54	0.49	0.05	FAIR
ETHNIC	10	Linear	0.50	0.00	0.50	UNFAIR
	1	Poly	0.63	0.00	0.63	UNFAIR
	10	RBF	59.00	0.00	59.00	UNFAIR
	1	Sigmoid	0.50	0.50	0.00	FAIR
PREG	10	Linear	0.50	0.00	0.50	UNFAIR
	1	Poly	0.62	0.00	0.62	UNFAIR
	10	RBF	0.59	0.00	0.59	UNFAIR
	1	Sigmoid	0.51	0.00	0.51	UNFAIR

8.7.6 Equalized Odds

Let $A = 1$ and $A = 0$ be the privileged and unprivileged groups respectively. When $Y = 1$, the equations shows the TPR and when $Y = 0$ it shows the FPR. This shows the ROC where TPR and FPR are equal for the privileged and unprivileged demographics.
(Irfan et al., 2023).

$$P\left[\hat{Y} = 1 \middle| Y = y, A = 0\right] = P[\hat{Y} = 1 | Y = y, A = 1] \tag{8.15}$$

where $y \in \{0, 1\}$.

The following table shows the TPR and FPR difference as well as the Equalized odds difference and whether it is fair or unfair. The differences are calculated by subtracting the two values from each class from each other for

the TPF and FPR respectively for each variable. Fairness is determined based on a .2 threshold. Any threshold could be chosen. Since equality determines fairness, a threshold closer to zero is wanted.

Table 8.6: Equal Odds - COMPLETED

Variable	Optimal C Value	Model	TPR Difference	FPR Difference	Equalized Odds Diff	Fair/Unfair
GENDER	0.1	Linear	0.32	0.03	0.29	UNFAIR
	10	Poly	0.14	0.09	0.05	FAIR
	10	RBF	0.17	0.06	0.11	FAIR
	1	Sigmoid	0.25	0.06	0.19	FAIR
AGE	0.1	Linear	0.26	0.03	0.23	UNFAIR
	10	Poly	0.15	0.05	0.10	FAIR
	10	RBF	0.16	0.08	0.08	FAIR
	1	Sigmoid	0.22	0.01	0.21	UNFAIR
VET	0.1	Linear	0.51	0.22	0.29	UNFAIR
	10	Poly	0.63	0.18	0.45	UNFAIR
	10	RBF	0.60	0.16	0.44	UNFAIR
	1	Sigmoid	0.51	0.22	0.29	UNFAIR
EDUC	0.1	Linear	0.32	0.29	0.03	FAIR
	10	Poly	0.25	0.48	0.23	UNFAIR
	10	RBF	0.25	0.36	0.11	FAIR
	1	Sigmoid	0.22	0.00	0.22	UNFAIR
MARSTAT	0.1	Linear	0.02	0.03	0.01	FAIR
	10	Poly	0.01	0.02	0.01	FAIR
	10	RBF	0.04	0.01	0.03	FAIR
	1	Sigmoid	0.09	0.02	0.07	FAIR
EMPLOY	0.1	Linear	0.39	0.03	0.36	UNFAIR
	10	Poly	0.12	0.04	0.08	FAIR
	10	RBF	0.22	0.04	0.18	FAIR
	1	Sigmoid	0.26	0.01	0.25	UNFAIR
RACE	0.1	Linear	0.02	0.00	0.02	FAIR
	10	Poly	0.00	0.02	0.02	FAIR
	10	RBF	0.06	0.03	0.03	FAIR
	1	Sigmoid	0.05	0.04	0.01	FAIR
ETHNIC	0.1	Linear	0.50	0.17	0.33	UNFAIR
	10	Poly	0.63	0.13	0.50	UNFAIR
	10	RBF	0.59	0.16	0.43	UNFAIR
	1	Sigmoid	0.00	0.25	0.25	UNFAIR
PREG	0.1	Linear	0.50	0.25	0.25	UNFAIR
	10	Poly	0.62	0.19	0.43	UNFAIR
	10	RBF	0.59	0.21	0.38	UNFAIR
	1	Sigmoid	0.51	0.25	0.26	UNFAIR

8.7.7 Equalized Odds - Multiclass

We do the same thing for the multiclass except now we take the maximum and minimum of each class to get the TPR difference and FPR difference for each class where $s \in sense\ variable$, $c \in class\ label$.

$$TPR_{sc} = \frac{TP_{sc}}{TP_{sc} + FN_{sc}} \tag{8.16}$$

$$FPR_{sc} = \frac{FP_{sc}}{FP_{sc} + TN_{sc}} \tag{8.17}$$

$$TPR_{diff} = \max(TPR_{sc}) - \min(TPR_{sc}) \tag{8.18}$$

$$FPR_{diff} = \max(FPR_{sc}) - \min(FPR_{sc}) \tag{8.19}$$

$$\begin{aligned} EqOdds_{diff} = &|\max(TPR_{sc}) - \min(TPR_{sc})| \\ &- |\max(FPR_{sc}) - \min(FPR_{sc})| \end{aligned} \tag{8.20}$$

The TPR difference, FPR difference, Equalized Odds difference and whether it is fair or unfair for each variable can be seen in the following table. Again, this was done at a .2 threshold and the default values for gamma, degree, and r with optimal C values for the multiclass NOPRIOR model. Note that closer to zero for an $EqOdds_{diff}$ is more fair.

Table 8.7: Equalized Odds NOPRIOR

ETHNIC							
C-Value	**Model**	**Max TPR**	**Min TPR**	**Max FPR**	**Min FPR**	**EOD Diff**	**Fairness**
GENDER							
10	Linear	1.00	0.00	0.26	0.00	0.74	UNFAIR
1	Poly	1.00	0.00	0.14	0.00	0.86	UNFAIR
10	RBF	1.00	0.00	0.12	0.00	0.88	UNFAIR
1	Sigmoid	1.00	0.00	0.29	0.00	0.71	UNFAIR
AGE							
10	Linear	0.99	0.00	0.20	0.00	0.79	UNFAIR
1	Poly	1.00	0.00	0.07	0.00	0.93	UNFAIR
10	RBF	0.99	0.00	0.07	0.00	0.92	UNFAIR
1	Sigmoid	1.00	0.27	0.28	0.00	0.45	UNFAIR
VET							
10	Linear	0.99	0.00	0.17	0.00	0.82	UNFAIR
1	Poly	0.99	0.00	0.17	0.00	0.82	UNFAIR
10	RBF	0.99	0.00	0.17	0.00	0.82	UNFAIR
1	Sigmoid	1.00	0.00	0.22	0.00	0.78	UNFAIR
EDUC							
10	Linear	0.99	0.00	0.09	0.00	0.90	UNFAIR
1	Poly	1.00	0.00	0.06	0.00	0.94	UNFAIR
10	RBF	1.00	0.00	0.07	0.00	0.93	UNFAIR
1	Sigmoid	1.00	0.00	0.31	0.00	0.69	UNFAIR
MARSTAT							
10	Linear	1.00	0.69	0.11	0.00	0.20	UNFAIR
1	Poly	1.00	0.70	0.11	0.00	0.19	FAIR
10	RBF	1.00	0.71	0.12	0.00	0.17	FAIR
1	Sigmoid	1.00	0.29	0.34	0.00	0.37	UNFAIR
EMPLOY							
10	Linear	0.99	0.00	0.25	0.00	0.74	UNFAIR
1	Poly	1.00	0.00	0.06	0.00	0.94	UNFAIR
10	RBF	1.00	0.00	0.06	0.00	0.94	UNFAIR
1	Sigmoid	1.00	0.00	0.89	0.00	0.11	FAIR
ETHNIC							
10	Linear	1.00	0.00	0.09	0.00	0.91	UNFAIR
1	Poly	1.00	0.00	0.06	0.00	0.94	UNFAIR
10	RBF	1.00	0.00	0.06	0.00	0.94	UNFAIR
1	Sigmoid	1.00	0.34	0.28	0.00	0.38	UNFAIR
ETHNIC							
10	Linear	1.00	0.00	0.08	0.00	0.92	UNFAIR
1	Poly	1.00	0.00	0.10	0.00	0.90	UNFAIR
10	RBF	1.00	0.00	0.07	0.00	0.93	UNFAIR
1	Sigmoid	1.00	0.00	1.00	0.00	0.00	FAIR
PREG							
10	Linear	1.00	0.00	1.00	0.00	0.00	FAIR
1	Poly	1.00	0.00	1.00	0.00	0.00	FAIR
10	RBF	1.00	0.00	1.00	0.00	0.00	FAIR
1	Sigmoid	1.00	0.00	1.00	0.00	0.00	FAIR

8.8 The Model

8.8.1 Support Vector Machines

The data sets are split 70/30 training and test sets with shuffle set at true which splits the data randomly. Support Vector Machines were run with four kernels. Linear, Polynomial, Radial Basis Function, and Sigmoid. The formulas for these kernels can be seen in the following equations.

$$\text{Linear Kernel :}$$
$$K(X, Y) = X^T Y \tag{8.21}$$
$$\text{Polynomial Kernel :}$$
$$K(X, Y) = (X^T Y + C)^d \tag{8.22}$$
$$\text{Radial Basis Function Kernel :}$$
$$K(X, Y) = e^{-\gamma \|X-Y\|^2}, \tag{8.23}$$
$$\gamma > 0$$
$$\text{Sigmoid Kernel :}$$
$$K(X, Y) = \tanh(\gamma X^T Y + C) \tag{8.24}$$

$$\text{where } \gamma = \text{auto} \quad = \frac{1}{n} \tag{8.25}$$
$$\text{where } \gamma = \text{scale} \quad = \frac{1}{n \times \text{Var}(X)} \tag{8.26}$$
$$\text{where Var}(x) \quad = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{8.27}$$

After further research, it was determined that with algorithms such as SVM, feature importance is not that useful. It is more useful to select an appropriate C value instead. The C value is a parameter that sets how much you want to penalize your model for each misclassified point. This means that the larger the C value, the smaller the margin, the smaller the C value, the larger the margin. The goal is to maximize the margin while minimizing the classification error. Next, we looked at different C values. In the models shown, the C value was set to four different C-values, .1, 1, 10, and 100 and the optimal values were chosen for each model. Note in Figure 8.12 (Felipe, 2019), the difference between large and small C values.

Large values of C	Small Values of C
Large effect of noisy points. A plane with very few misclassifications will be given precedence.	Low effect of noisy points. Planes that separate the points well will be found, even if there are some misclassifications

Figure 8.12: The Influence of C Parameter (Felipe, 2019)

The results of using the different C-Values with a grid search of *gamma*, *r*, and *degree* can be seen on the COMPLETED and NOPRIOR datasets for all four kernels shown. The best performing model was the rbf-wgt model with a C value of 10 having a prediction 10 fold CV accuracy of 93.78% for the COMPLETED data set. The confusion matrix can be seen in Figure **??**. The polynomial kernel model was a close second with an accuracy of 92.45% with a $C = 100$. The lower the c value the better it performed in all the models.

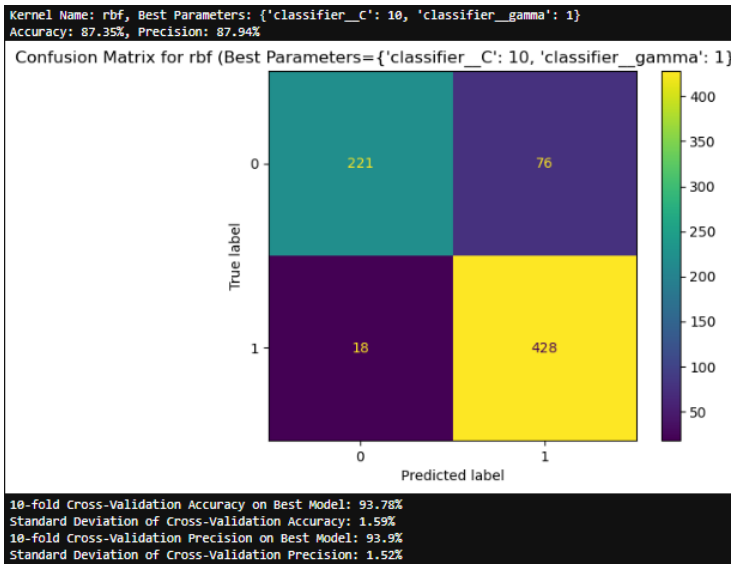


Figure 8.13: Optimal Confusion COMPLETED

For the multiclass model, the accuracy was more consistent with poly coming in first with an accuracy of 91.02% with a C-value of 1 and polynomial coming in second with an accuracy of 90.3% and a C-value of 10. The linear model was also higher at 79.35% with a C-value of 10 as opposed to the 66.24% in the COMPLETED data. The confusion matrix can be seen in the following figures.

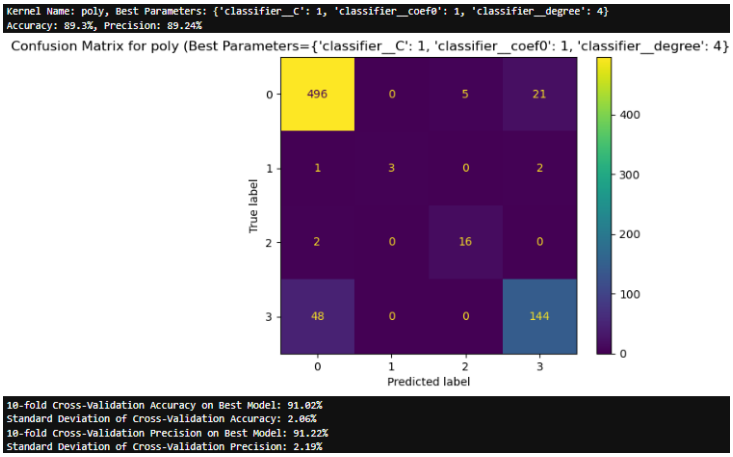


Figure 8.14: Optimal Confusion NOPRIOR

8.8.2 Decision Trees

A grid search approach was used for decision trees with the following parameters:

- **Max Depth**: [None, 2, 4, 6, 8, 10]
- **Min Samples Split**: [2, 5, 10]
- **Min Samples Leaf**: [1, 2, 4]

The optimal results for each data set can be seen in the following figures:

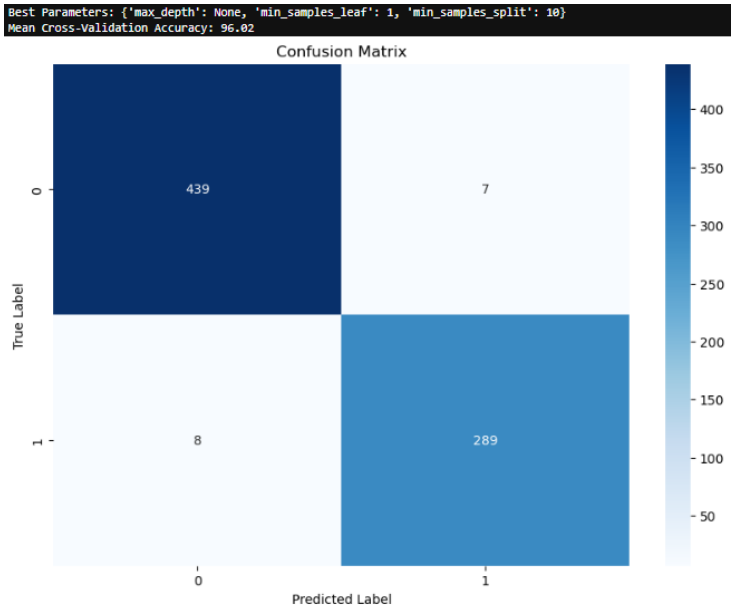


Figure 8.15: Optimal Confusion COMPLETED DT

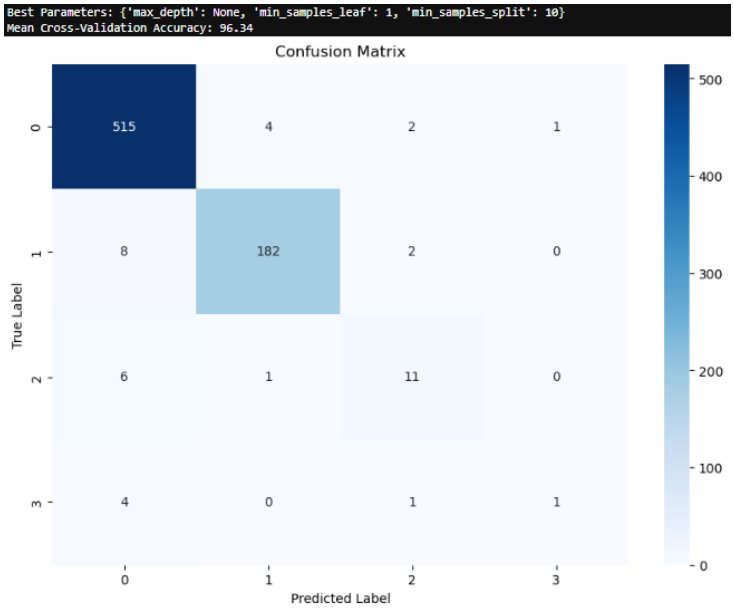


Figure 8.16: Optimal Confusion NOPRIOR DT

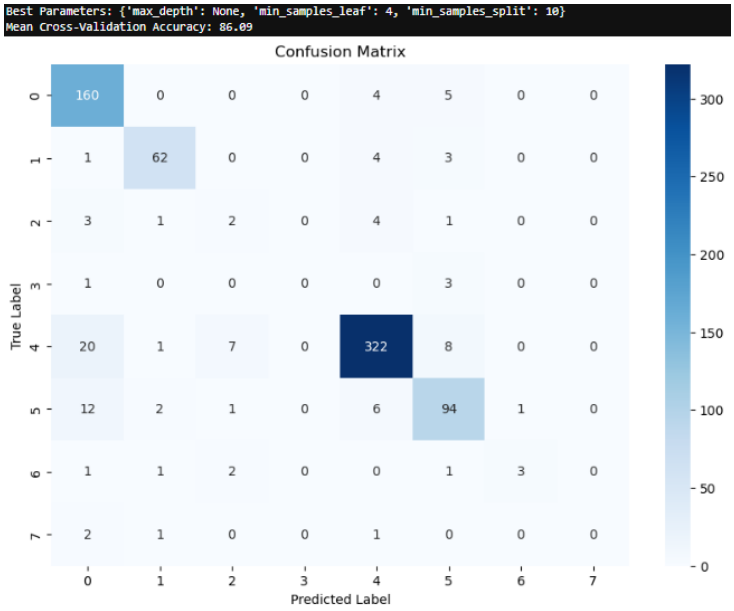


Figure 8.17: Optimal Confusion COMPLETED_NOPRIOR DT

8.8.3 Random Forests

A parameter grid approach was used for random forests with the following parameters:

- **Number of Estimators**: [10, 50, 100, 200]
- **Max Features**: ['auto', 'sqrt', 'log2']
- **Max Depth**: [None, 5, 10, 20]
- **Min Samples Split**: [2, 5, 10]
- **Min Samples Leaf**: [1, 2, 4]

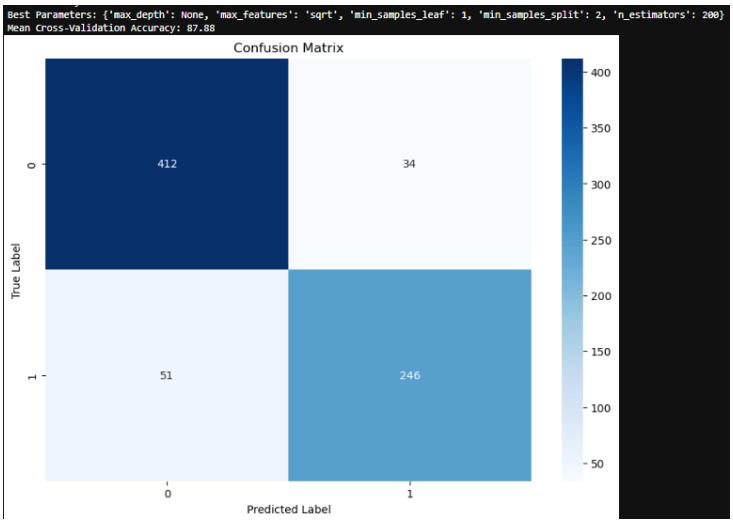


Figure 8.18: Optimal Confusion COMPLETED RF

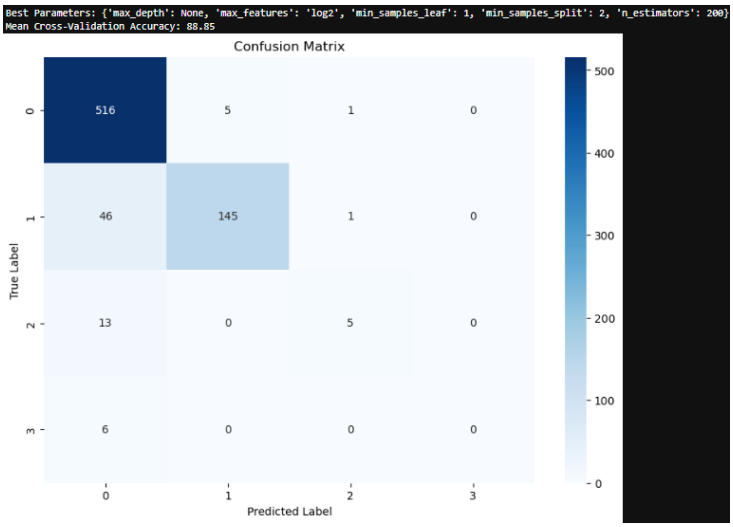


Figure 8.19: Optimal Confusion NOPRIOR RF

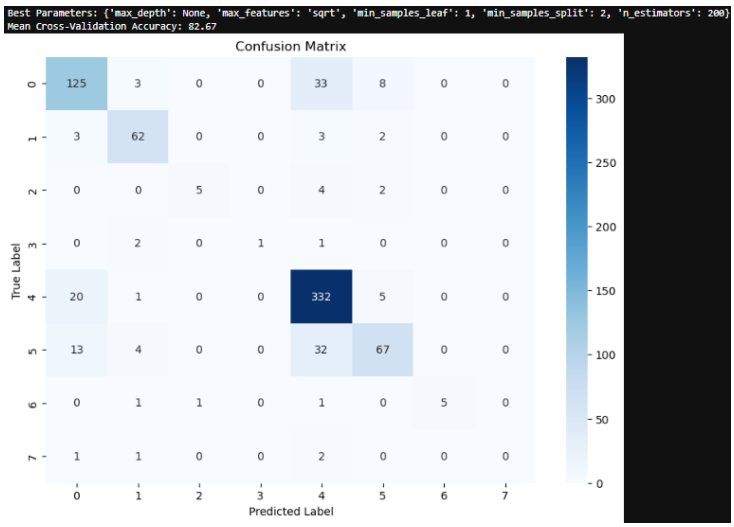


Figure 8.20: Optimal Confusion COMPLETED_NOPRIOR RF

8.8.4 Neural Networks

The parameter grid for the neural network model is as follows:

- **Units in Layer 1**: [8, 10, 20, 30]
- **Units in Layer 2**: [8, 10, 20, 30]
- **Activation Function in Input Layers**: ['relu', 'tanh', 'sigmoid']
- **Optimizer**: ['adam', 'sgd']
- **Loss Function**: ['categorical_crossentropy', 'mean_squared_error']

The resulting optimal parameters along with their confusion matrices for each data set can be seen in the following figures:

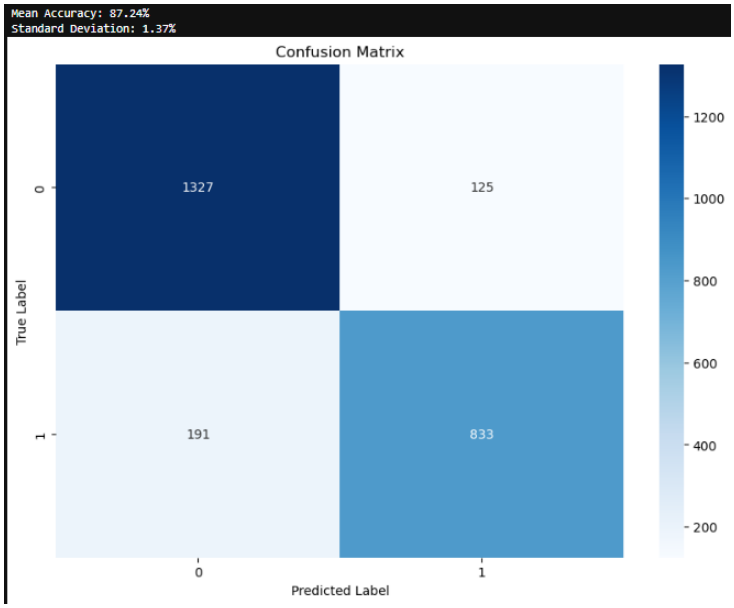


Figure 8.21: Optimal Confusion COMPLETED NN

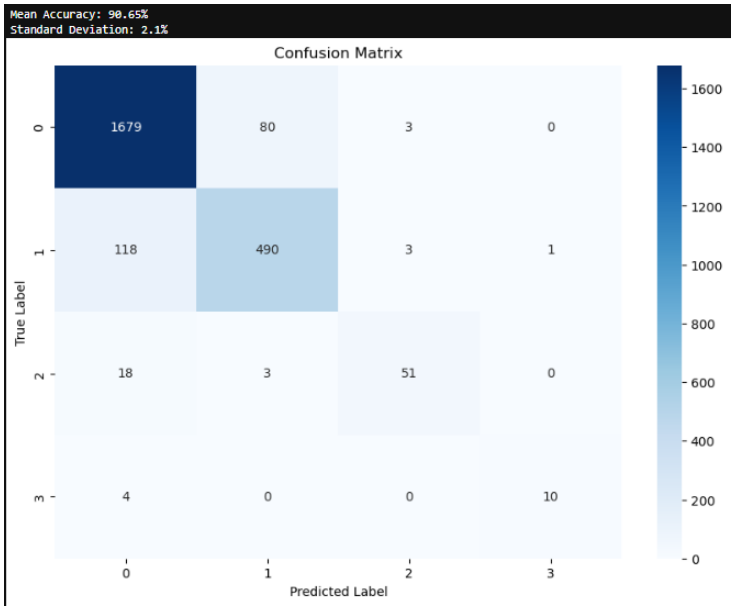


Figure 8.22: Optimal Confusion NOPRIOR NN

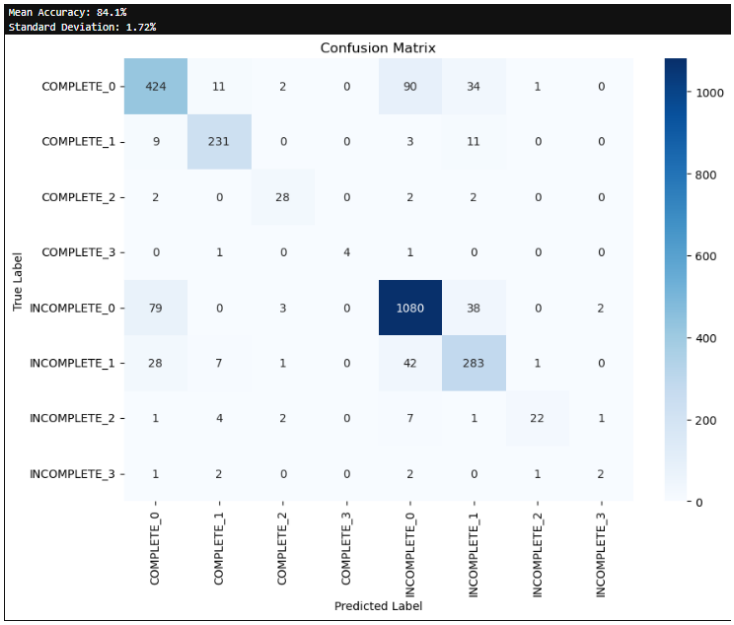


Figure 8.23: Optimal Confusion COMPLETED_NOPRIOR NN

Value	Label
0	No prior treatment episodes
1	One prior treatment episode
2	Two prior treatment episodes
3	Three prior treatment episodes
4	Four prior treatment episodes
5	Five or more prior treatment episodes

Figure 8.24: NOPRIOR Translation

8.9 Interpretation of Results

The NOPRIOR variable translation can be seen in the following figure. This is the translation to what is shown in the x-axis of the previous confusion matrix.

The encoded variables used in the data sets and their translated meanings can be seen in the following tables.

Table 8.8: Variable Descriptions

Variable	**Meaning**
SERVICES_4	Rehab/residential, short term (30 days or fewer)
SERVICES_5	Rehab/residential, long term (more than 30 days)
SERVICES_7	Ambulatory, non-intensive outpatient
PSOURCE_2	Alcohol/drug use care provider
PSOURCE_3	Other health care provider
PSOURCE_4	School (educational)
PSOURCE_5	Employer/EAP
PSOURCE_6	Other community referral
PSOURCE_7	Court/criminal justice referral/DUI/DWI
METHUSE_2	No medication-assisted opioid therapy
LIVARAG_2	Dependent living
LIVARAG_3	Independent living
ALCFLG_1	Alcohol reported at admissions
COKEFLG_1	Cocaine/crack reported at admissions
MARFLG_1	Marijuana/hashish reported at admissions

Variable	Meaning
HERFLG_1	Heroin reported at admissions
METHFLG_1	Non-rx methadone reported at admissions
OPSYNFLG_1	Other opiates/synthetics reported at admission
PCPFLG_1	PCP reported at admission
HALLFLG_1	Hallucinogens reported at admission
MTHAMFLG_1	Methamphetamine/speed reported at admission
AMPHFLG_1	Other amphetamines reported at admission
STIMFLG_1	Other stimulants reported at admission
BENZFLG_1	Benzodiazepines reported at admission
TRNQFLG_1	Other tranquilizers reported at admission
BARBFLG_1	Barbiturates reported at admission
SEDHPFLG_1	Other sedatives/hypnotics reported at admission
INHFLG_1	Inhalants reported at admission
OTCFLG_1	Over-the-counter medication reported at admission
OTHERFLG_1	Other drug reported at admission
ALCDRUG_1	Alcohol only
ALCDRUG_2	Other drugs only
ALCDRUG_3	Alcohol and other drugs
DETNLF_3	Retired, disabled not in labor force
DETNLF_4	Resident of institution not in labor force
DETNLF_5	Other not in labor force
DETCRIM_2	Formal adjudication process
DETCRIM_3	Probation/parole
DETCRIM_6	Prison
NOPRIOR_1	One prior treatment
NOPRIOR_2	Two prior treatments
NOPRIOR_3	Three prior treatments
PSOURCE_2	Alcohol/drug use care provider
PSOURCE_3	Other health care provider

Variable	Meaning
PSOURCE_6	Other community referral
PSOURCE_7	Court/criminal justice referral/DUI/DWI
ARRESTS_1	Once
ARRESTS_2	Two or more times
SUB3_10	Methamphetamine tertiary
SUB3_11	Other amphetamines tertiary
SUB3_12	Other stimulants tertiary
SUB3_13	Benzodiazepines tertiary
SUB3_16	Other sedatives or hypnotics tertiary
SUB3_19	Other drugs tertiary
SUB3_2	Alcohol tertiary
SUB3_3	Cocaine/crack tertiary
SUB3_4	Marijuana/hashish tertiary
SUB3_5	Heroin tertiary
SUB3_6	Non-prescription methadone tertiary
SUB3_7	Other opiates and synthetics tertiary
SUB3_8	PCP tertiary
SUB3_9	Hallucinogens tertiary

variable	meaning
METHUSE_2	no medication-assisted opioid therapy
PSYPROB_2	no coocurring mental
	and substance abuse disorders
DSMCRIT_10	cannabis use
DSMCRIT_11	other substance use
DSMCRIT_12	opioid abuse
DSMCRIT_13	cocaine abuse
DSMCRIT_2	substance-induced disorder
DSMCRIT_3	alcohol intoxication
DSMCRIT_4	alcohol dependency
DSMCRIT_5	opioid dependency
DSMCRIT_6	cocaine dependency
DSMCRIT_7	cannabis dependency
DSMCRIT_8	other substance dependency
DSMCRIT_9	alcohol abuse
SUB1_10	Methamphetamine/speed primary
SUB1_11	other amphetamines primary
SUB1_12	other stimulants primary
SUB1_13	benzodiazepines primary
SUB1_19	other drugs primary
SUB1_2	alcohol primary
SUB1_3	cocaine/crack primary
SUB1_4	marijuana/hashish primary
SUB1_5	heroin primary

Table 8.9: Variable Descriptions Cont'd

8.10 Reweighting

8.10.1 Interactions

The Chi Squared test was used to test significant interactions between bias variables.

$$\chi^2 = \frac{(E-O)^2}{E}, \tag{8.28}$$

where E is the expected value and O is referring to the observed frequencies in the contingency table respectively. We looked at both dual and three-way interactions and then combined the datasets and calculated the probability that at least one of these interactions occurred as our new reweighted weight. If there is a significant interaction, then the interaction counts for each x1, x2 for the dual interaction or x1, x2, and x3 for the three-way interaction are then divided by the total value counts for x1, x2, with the target or x1, x2, and x3 with the target. This gives the probability. These would be multiplicative models (Veenstra, 2011).

8.10.1.1 Dual Interaction

$$\begin{aligned} S_1 &= \{x_1, x_2\} \\ S_2 &= \{x_1, x_2, \text{target}\} \\ \text{DualInteraction}_{\text{Probability}} &= \frac{|S_1|}{|S_2|} \end{aligned} \tag{8.29}$$

Example:

$$\frac{P(\text{White} \cap \text{Male})}{P(\text{Complete} \cap \text{White} \cap \text{Male})} \tag{8.30}$$

8.10.1.2 Three Way Interaction

$$S_3 = \{x_1, x_2, x_3\} \tag{8.31}$$

$$S_4 = \{x_1, x_2, x_3, \text{target}\} \tag{8.32}$$

$$\text{3wayInteraction}_{\text{Probability}} = \frac{|S_3|}{|S_4|} \tag{8.33}$$

Example:

$$\frac{P(\text{White} \cap \text{Female} \cap \text{Pregnant})}{P(\text{Complete} \cap \text{White} \cap \text{Female} \cap \text{Pregnant})} \tag{8.34}$$

8.10.2 Intersectionalization

By taking the case for which at least one interaction occurs, we are reweighting the fairness. Then we combined the two-way and three-way interaction datasets together to get our new df dataset. This allows for the case when a person can be in two groups at once. An example would be veteran and male or veteran and male under 40.

$$\text{FinalProbability} = 1 - \left(1 - \prod \text{probability}_{\text{interactions}}\right) \tag{8.35}$$

$$\begin{aligned} \text{probability}_{\text{interactions}} &= (\text{DualInteraction}_{\text{Probability}}) \\ &\cup \left(\text{3wayInteraction}_{\text{Probability}}\right) \end{aligned} \tag{8.36}$$

8.11 New Fairness Calculations

8.11.1 Worst Case Demographic Parity

Demographic parity compares the pass rate of (positive outcome) of two groups. It is satisfied if:

$$P\left(\hat{Y} \middle| A \in \text{sg}_i\right) = P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i, j \in \mathbb{N},\ i \neq j \tag{8.37}$$

Where N is the total number of subgroups. When using the worst-case ratio formula, it is also known as demographic parity ratio, (Ghosh et al., 2021):

$$\text{DPR} = \frac{\min\left\{P\left(\hat{Y}\middle|A \in \text{sg}_i\right)\forall i \in \mathbb{N}\right\}}{\max\left\{P\left(\hat{Y}\middle|A \in \text{sg}_i\right)\forall i \in \mathbb{N}\right\}} \tag{8.38}$$

Minimum Demographic Parity Ratio: 0.9524058315188466

	COMPLETED	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2	0.999484	0.999981	0.999503
1	0	3	0.951284	0.998822	0.952406
2	1	2	0.999836	1.000000	0.999837
3	1	3	0.974790	0.999861	0.974925

Figure 8.25: Merged - COMPLETED

Minimum Demographic Parity Ratio: 0.7218107423676939

	NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2.0	0.981420	1.000000	0.981420
1	0	3.0	0.903449	0.999998	0.903451
2	1	2.0	0.985489	0.994492	0.990947
3	1	3.0	0.920917	0.985644	0.934330
4	2	2.0	0.964407	0.999969	0.964437
5	2	3.0	0.721809	0.999998	0.721811
6	3	2.0	0.841075	0.999551	0.841453
7	3	3.0	0.831436	0.999546	0.831813

Figure 8.26: Merged - NOPRIOR

Minimum Demographic Parity Ratio: 0.02458828773206945

	COMPLETED_NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2.0	0.645950	0.962295	0.671259
1	0	3.0	0.487470	0.901409	0.540787
2	1	2.0	0.524531	0.895904	0.585477
3	1	3.0	0.581854	0.823778	0.706324
4	2	2.0	0.859057	0.967239	0.888154
5	2	3.0	0.571817	0.937434	0.609981
6	3	2.0	0.630716	0.998639	0.631576
7	3	3.0	0.819125	0.998596	0.820277
8	4	2.0	0.762554	0.941340	0.810073
9	4	3.0	0.417672	0.981201	0.425674
10	5	2.0	0.024203	0.984348	0.024588
11	5	3.0	0.636053	0.959108	0.663171
12	6	2.0	0.562646	0.899390	0.625586
13	6	3.0	0.355316	0.499445	0.711422
14	7	2.0	0.918633	0.981632	0.935823
15	7	3.0	0.838083	0.906130	0.924903

Figure 8.27: Merged - COMPLETED_NOPRIOR

8.11.2 Worst Case Demographic Parity – Multiclass

$$\text{DPR}_{\text{multiclass}} = \min \left\{ \frac{\min \left\{ P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i \in N \right\}}{\max \left\{ P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i \in N \right\}} \right\} \tag{8.39}$$

The same is done for the multiclass DPR except the overall minimum is taken to get the minimum over all classes. [Gosh et al., 2022](Ghosh et al., 2021). The results can be seen in the following figures for COMPLETED, NORPIOR, and COMPLETED_NORPIOR merged on two-way and three-way interactions.

Minimum Demographic Parity Ratio: 0.02458828773206945

	COMPLETED_NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2.0	0.645950	0.962295	0.671259
1	1	2.0	0.524531	0.895904	0.585477
2	2	2.0	0.859057	0.967239	0.888154
3	3	2.0	0.630716	0.998639	0.631576
4	4	2.0	0.762554	0.941340	0.810073
5	5	2.0	0.024203	0.984348	0.024588
6	6	2.0	0.562646	0.899390	0.625586
7	7	2.0	0.918633	0.981632	0.935823

Figure 8.28: 2-Way Interaction NOPRIOR

8.11.3 Worst Case Disparate Impact

$$\text{DI} = \min \left\{ \frac{P(\hat{Y}|A \in \text{sg}_i)}{P(\hat{Y}|A \in \text{sg}_i)}; \forall i, j \in N, i \neq j \right\}, \tag{8.40}$$

Minimum Demographic Parity Ratio: 0.4256741722674935

	COMPLETED_NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	3.0	0.487470	0.901409	0.540787
1	1	3.0	0.581854	0.823778	0.706324
2	2	3.0	0.571817	0.937434	0.609981
3	3	3.0	0.819125	0.998596	0.820277
4	4	3.0	0.417672	0.981201	0.425674
5	5	3.0	0.636053	0.959108	0.663171
6	6	3.0	0.355316	0.499445	0.711422
7	7	3.0	0.838083	0.906130	0.924903

Figure 8.29: 3-Way Interaction NOPRIOR

where in this case i and j are unique pairs in the series. The data is grouped by number of terms (2- or 3-way interactions), The probabilities are calculated between the pairs of ratios for each group of target class and pair of probabilities. The minimum ratios are all aggregated across all pairs, and then minimum among these aggregated ratios is the worst-case scenario for disparate impact. The 80% rule must be passed for fairness to be achieved. This is also known as the four fifth rule. The pass rate of group 1 to group 2 must be greater than 80%, with group 1 and 2 interchangeable (Ghosh et al., 2021). Again, the results are shown in the following figures.

Minimum Min_Ratio_Pairs: 0.9524058315188466

	COMPLETED	Number_of_Terms	Min_Ratio_Pairs
0	0	2	0.999503
1	0	3	0.952406
2	1	2	0.999837
3	1	3	0.974925

Figure 8.30: Merged COMPLETED DI

Minimum Min_Ratio_Pairs: 0.7218107423676939

	NOPRIOR	Number_of_Terms	Min_Ratio_Pairs
0	0	2.0	0.981420
1	0	3.0	0.903451
2	1	2.0	0.990947
3	1	3.0	0.934330
4	2	2.0	0.964437
5	2	3.0	0.721811
6	3	2.0	0.841453
7	3	3.0	0.831813

Figure 8.31: Merged NOPRIOR DI

Minimum Min_Ratio_Pairs: 0.02458828773206945

	COMPLETED_NOPRIOR	Number_of_Terms	Min_Ratio_Pairs
0	0	2.0	0.671259
1	0	3.0	0.540787
2	1	2.0	0.585477
3	1	3.0	0.706324
4	2	2.0	0.888154
5	2	3.0	0.609981
6	3	2.0	0.631576
7	3	3.0	0.820277
8	4	2.0	0.810073
9	4	3.0	0.425674
10	5	2.0	0.024588
11	5	3.0	0.663171
12	6	2.0	0.625586
13	6	3.0	0.711422
14	7	2.0	0.935823
15	7	3.0	0.924903

Figure 8.32: Merged COMPLETED_NOPRIOR DI

Note that, COMPLETED is the fairest, NOPRIOR is second most fair, and COMPLETED_NOPRIOR is the least fair. It is fair individually, however in a worst-case scenario it has the lowest score.

8.11.4 Conditional Statistical Parity Ratio

This uses the worst-case min-max ratio, with predictor $\hat{Y}$, member A, and legitimate attribute L (Corbett-Davies et al., 2017).

$$P\left(\hat{Y} \mid L = 1, A \in \text{sg}_i\right) = P\left(\hat{Y} \mid L = 1, A \in \text{sg}_j\right) \quad \forall i, j \in N, i \neq j \tag{8.41}$$

$$CSPR = \frac{\min\left\{P(\hat{Y} \mid L = 1, A \in \text{sg}_i)\ \forall i \in N\right\}}{\max\left\{P(\hat{Y} \mid L = 1, A \in \text{sg}_i)\ \forall i \in N\right\}} \tag{8.42}$$

$$CSPR_{\text{Multiclass}} = \min\left\{\frac{\min\left\{P(\hat{Y} \mid L = 1, A \in \text{sg}_i)\ \forall i \in N\right\}}{\max\left\{P(\hat{Y} \mid L = 1, A \in \text{sg}_i)\ \forall i \in N\right\}}\right\} \tag{8.43}$$

Minimum CSPR value: 0.9759400182323508

	Number_of_Terms	CSPR
0	2	0.999856
1	3	0.975940

Figure 8.33: CSPR COMPLETED

This uses the worst-case min-max ratio, with predictor $\hat{Y}$, member A, and legitimate attribute L (Corbett-Davies et al., 2017).

Minimum CSPR value: 0.7218110561977474

Figure 8.34: CSPR NOPRIOR

Minimum CSPR value: 0.4968096639431072

Figure 8.35: CSPR COMPLETED_NOPRIOR

8.11.5 Equal Opportunity Ratio - Multiclass

The TPR is compared for the protected and unprotected group. The minimum of these overall is taken for the multiclass case (Ghosh et al., 2021).

$$\begin{aligned} &P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1) = \\ &\quad P(\hat{Y} = y_k | A \in \text{sg}_j, Y = 1) \\ &\quad \forall i, j \in N, i \neq j \end{aligned} \tag{8.44}$$

$$\text{EOppR} = \frac{\min\left\{P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1) \forall i \in N \forall k \in K\right\}}{\max\left\{P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1) \forall i \in N \forall k \in K\right\}} \tag{8.45}$$

$$\text{EOppR}_{\text{Multi}} = \min\left\{\frac{\min\left\{P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1), \forall i \in N, \forall k \in K\right\}}{\max\left\{P(\hat{Y} = y_k | A \in \text{sg}_i, Y = 1), \forall i \in N, \forall k \in K\right\}}\right\} \tag{8.46}$$

where K is the set of all possible output classes.

```
Minimum Equal Opportunity Ratio: 0.95
{(2, 0): 1.0, (2, 1): 1.0, (3, 0): 0.95, (3, 1): 0.97}
```

Figure 8.36: Merged COMPLETED Equal Opporunity

```
Minimum Equal Opportunity Ratio: 0.72
{(2.0, 0): 0.98,
 (2.0, 1): 0.99,
 (2.0, 2): 0.96,
 (2.0, 3): 0.84,
 (3.0, 0): 0.9,
 (3.0, 1): 0.93,
 (3.0, 2): 0.72,
 (3.0, 3): 0.83}
```

Figure 8.37: Merged NOPRIOR Equal Opportunity

```
Minimum Equal Opportunity Ratio: 0.02
{(2.0, 0): 0.67,
 (2.0, 1): 0.59,
 (2.0, 2): 0.89,
 (2.0, 4): 0.81,
 (2.0, 3): 0.63,
 (2.0, 5): 0.02,
 (2.0, 6): 0.63,
 (2.0, 7): 0.94,
 (3.0, 0): 0.54,
 (3.0, 1): 0.71,
 (3.0, 2): 0.61,
 (3.0, 3): 0.82,
 (3.0, 4): 0.43,
 (3.0, 5): 0.66,
 (3.0, 7): 0.92,
 (3.0, 6): 0.71}
```

Figure 8.38: Merged COMPLETED_NOPRIOR Equal Opportunity

8.11.6 Equal Odds Ratio

K is the set of all possible output classes. The closer the value is to 1, the lower the disparity is. For this case we calculate the worst-case odds for each output class y and then take the minimum of all those values (Ghosh et al., 2021). This is the multiclass case for equalized odds ratio.

$$P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right) = P\left(\hat{Y} = y_k \middle| A \in \text{sg}_j\right) \quad \forall i, j \in N, i \neq j, \forall k \in K \tag{8.47}$$

$$\text{EOddR}_{\text{Multiclass}} = \min\left\{\frac{\min\left\{P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right)\right\}}{\max\left\{P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right)\right\}}\right\} \forall i \in N, \forall k \in K \tag{8.48}$$

where K is the set of all possible output classes.

```
Equalized Odds Ratio for class '0' in term group '2': 1.0
Equalized Odds Ratio for class '1' in term group '2': 1.0
Equalized Odds Ratio for class '0' in term group '3': 0.98
Equalized Odds Ratio for class '1' in term group '3': 1.02
Minimum Equalized Odds Ratio: 0.98
```

Figure 8.39: Merged COMPLETED Equalized Odds

```
Equalized Odds Ratio for class '0' in term group '2.0': 0.99
Equalized Odds Ratio for class '1' in term group '2.0': 1.01
Equalized Odds Ratio for class '2' in term group '2.0': 0.98
Equalized Odds Ratio for class '3' in term group '2.0': 0.86
Equalized Odds Ratio for class '0' in term group '3.0': 0.97
Equalized Odds Ratio for class '1' in term group '3.0': 1.03
Equalized Odds Ratio for class '2' in term group '3.0': 0.8
Equalized Odds Ratio for class '3' in term group '3.0': 0.92
Minimum Equalized Odds Ratio: 0.8
```

Figure 8.40: Merged NOPRIOR Equalized Odds

```
Equalized Odds Ratio for class '0' in term group '2.0': 1.15
Equalized Odds Ratio for class '1' in term group '2.0': 0.87
Equalized Odds Ratio for class '2' in term group '2.0': 1.32
Equalized Odds Ratio for class '4' in term group '2.0': 1.21
Equalized Odds Ratio for class '3' in term group '2.0': 0.94
Equalized Odds Ratio for class '5' in term group '2.0': 0.04
Equalized Odds Ratio for class '6' in term group '2.0': 0.93
Equalized Odds Ratio for class '7' in term group '2.0': 1.39
Equalized Odds Ratio for class '0' in term group '3.0': 0.77
Equalized Odds Ratio for class '1' in term group '3.0': 1.31
Equalized Odds Ratio for class '2' in term group '3.0': 1.13
Equalized Odds Ratio for class '3' in term group '3.0': 1.52
Equalized Odds Ratio for class '4' in term group '3.0': 0.79
Equalized Odds Ratio for class '5' in term group '3.0': 1.23
Equalized Odds Ratio for class '7' in term group '3.0': 1.71
Equalized Odds Ratio for class '6' in term group '3.0': 1.32
Minimum Equalized Odds Ratio: 0.04
```

Figure 8.41: Merged COMPLED_NOPRIOR Equalized Odds

COMPLETED									Reweighted Fairness				
Model	C-Value	Gamma	r	degree	10 Fold CV Accuracy	CV Accuracy Std Dev	CV Precision	CV Precision Std Dev	DIR	DPR	CSPR	EqOppR	EqOddsR
Linear-wgt	1000				66.23	2.8	66.75	3.41	1	0.95	0.98	0.95	0.98
Poly-wgt	100		-0.5	5	92.45	1.78	92.5	1.79					
RBF-wgt	10	1			93.78	1.59	93.9	1.52					
Sigmoid-wgt	1000	0.01	-1		81.46	1.95	81.56	1.97					

Figure 8.42: Completed Results

NOPRIOR									Reweighted Fairness				
Model	C-Value	Gamma	r	degree	10 Fold CV Accuracy	CV Accuracy Std Dev	CV Precision	CV Precision Std Dev	DIR	DPR	CSPR	EqOppR	EqOddsR
Linear-wgt	10				79.35	2.33	81.55	1.96	0.72	0.72	0.72	0.72	0.8
Poly-wgt	1		1	4	91.02	2.06	91.22	2.19					
RBF-wgt	10	scale			90.3	1.76	90.37	1.94					
Sigmoid-wgt	10	0.01	-1		76.22	2.54	78.28	2.18					
* Note: only went to c=10 on sigmoid due to runtime complexity													

Figure 8.43: NOPRIOR Results

COMPLETED-NOPRIOR									Reweighted Fairness				
Model	C-Value	Gamma	r	degree	10 Fold CV Accuracy	CV Accuracy Std Dev	CV Precision	CV Precision Std Dev	DIR	DPR	CSPR	EqOppR	EqOddsR
Linear-wgt	10				71.98	3.28	71.44	3.2	0.02	0.02	0.5	0.02	0.04
Poly-wgt	10		0	4	90.23	1.79	90.67	1.74					
RBF-wgt	10	0.1			89.25	1.32	89.7	1.22					
Sigmoid-wgt	1000	0.01	-1		81.91	2.93	82.01	2.99					

Figure 8.44: Completed_NOPRIOR Results

Next we did hypothesis testing to compare the polynomial and rbf kernels to see if they were significantly different. All were significantly different from each other in all three data sets. The results can be seen in the following table.

Data Set	T-statistic	P-value	Significance
COMPLETED	-2.4441	0.0371	Statistically significant
NOPRIOR	3.5839	0.0059	Statistically significant
COMPLETED_NOPRIOR	4.2794	0.0021	Statistically significant

Table 8.10: SVM T-test Results

Class	Precision	Recall	F1-Score	Support
Incomplete	0.98	0.99	0.99	446
Complete	0.98	0.97	0.98	297
CV Accuracy: 96.48				

Table 8.11: Classification Report Completed DT

Class	Precision	Recall	F1-Score	Support
0	0.97	0.99	0.98	522
1	0.97	0.95	0.96	192
2	0.69	0.61	0.65	18
3	0.50	0.17	0.25	6
CV Accuracy: 96.11				

Table 8.12: Classification Report NOPRIOR DT

Class	Precision	Recall	F1-Score	Support
COMPLETE_0	0.81	0.93	0.87	169
COMPLETE_1	0.91	0.91	0.91	70
COMPLETE_2	0.57	0.36	0.44	11
COMPLETE_3	0.00	0.00	0.00	4
INCOMPLETE_0	0.94	0.91	0.93	358
INCOMPLETE_1	0.82	0.84	0.83	116
INCOMPLETE_2	1.00	0.25	0.40	8
INCOMPLETE_3	0.00	0.00	0.00	4
CV Accuracy: 86.15				

Table 8.13: Classification Report Completed_NOPRIOR DT

Class	Precision	Recall	F1-Score	Support
Incomplete	0.89	0.92	0.91	446
Complete	0.88	0.83	0.85	297
CV Accuracy: 87.88				

Table 8.14: Classification Report Completed RF

Class	Precision	Recall	F1-Score	Support
0	0.89	0.99	0.94	522
1	0.97	0.76	0.85	192
2	0.71	0.28	0.40	18
3	0.00	0.00	0.00	6
CV Accuracy: 88.85				

Table 8.15: Classification Report NOPRIOR RF

Class	Precision	Recall	F1-Score	Support
COMPLETE_0	0.77	0.74	0.76	169
COMPLETE_1	0.84	0.89	0.86	70
COMPLETE_2	0.83	0.45	0.59	11
COMPLETE_3	1.00	0.25	0.40	4
INCOMPLETE_0	0.81	0.93	0.87	358
INCOMPLETE_1	0.80	0.58	0.67	116
INCOMPLETE_2	1.00	0.62	0.77	8
INCOMPLETE_3	0.00	0.00	0.00	4
CV Accuracy: 82.67				

Table 8.16: Classification Report Completed_NOPRIOR RF

Model	Completed	NOPRIOR	Completed_NOPRIOR
Number of Layers	2	2	2
Number of Neurons Layer 1	220	220	200
Number of Neurons Layer 2	10	200	220
Activation Function	ReLU (Rectified Linear Unit)	ReLU (Rectified Linear Unit)	ReLU (Rectified Linear Unit)
Network Optimizer	adam	adam	adam
Loss Function	mean squared error	mean squared error	categorical crossentropy
Epoch	20	20	20
Accuracy	77.52	85.09	73.65
Recall Incomplete	0.86		
Recall Complete	0.65		
Precision Incomplete	0.79		
Precision Complete	0.75		
Recall 0		0.92	
Recall 1		0.71	
Recall 2		0.61	
Recall 3		0.00	
Precision 0		0.89	
Precision 1		0.76	
Precision 2		0.73	
Precision 3		0.00	
Recall 0			0.54
Recall 1			0.83
Recall 2			0.36
Recall 3			0.25
Recall 4			0.87
Recall 5			0.61
Recall 6			0.75
Recall 7			0.50
Precision 0			0.70
Precision 1			0.78
Precision 2			0.57
Precision 3			1.00
Precision 4			0.77
Precision 5			0.64
Precision 6			0.60
Precision 7			1.00

Table 8.17: Model performance metrics

The Rectified Linear Unit activation function is 0 when x is negative and x when x is positive $f(x) = max(x, 0)$, (Ramachandran et al., 2018). This was the optimal activation function in all three models.

8.12 Conclusions

Initially this research set out to predict whether a person completed treatment or not and it proved difficult due to the imbalance of the data with roughly 32% of people completing treatment. When this proved to have poor accuracy with SVMs, other algorithms such as decision trees and random forest approaches were utilized, A different approach was taken to look at predicting how many prior times a person had been in treatment as well as how the combination of completed and NOPRIOR. Fairness was explored with 9 variables and after reweighting the variables, a higher fairness was achieved in all the models with only a small decrease in accuracy. The highest model for the completed data set was the rbf-wgt at 93.78%. Note that the reweighted COMPLETED model is fair when it comes to all fairness measures with a fairness of 95% for demographic parity, 98% for CSPR, equal opportunity of 95% and equal odds of 98%. For the NOPRIOR data set, poly-wgt had the best accuracy with the best fairness. It had an accuracy of 91.22% and a fairness of 72% for all fairness except equal odds which was 80%. COMPLETED_NOPRIOR was fair looking at individual fairness but not at worse case fairness, due to some classes having poor fairness. The accuracy of this model was 90.23% for poly-wgt. Based on fairness and accuracy, the COMPLETED model predicts the best and NOPRIOR is the next best. An RBF or poly kernel would be the best to use when it comes to SVM, c value of 10, r=1, gamma=1, degree=4. This provides the best results. In comparison, decision trees performed much better overall having an accuracy of 96.48% for COMPLETED, 96.11% for NOPRIOR and 86.15% for COMPLETED_NOPRIOR. Neural Networks performed the third best with an accuracy of 89.0% for COMPLETED, 90.65% for NORPIOR, and 84.1% for COMPLETED_NOPRIOR. Random forests performed in the worst with an accuracy of 87.88% for COMPLETED, 88.85% for NOPRIOR, and 82.67% for COMPLETED_NOPRIOR. Overall decision trees performed the best with the COMPLETED data set being the best predicting data set, NOPRIOR being next and COMPLETED_NOPRIOR coming in last in accuracy. For this reason, we would choose to use the decision trees model to predict all three data sets.

Chapter 9

Conclusion

This dissertation has examined fairness across machine learning, optimization, and system modeling, with particular application to substance abuse treatment and criminal justice in Oklahoma. The work began with the question of how fairness can be defined and assessed in machine learning, especially when algorithms are applied to sensitive characteristics such as gender, race, or age. The analysis of fairness measures and mitigation techniques showed that attention to fairness is not simply a technical exercise. Rather, it requires balancing predictive performance with ethical considerations so that models do not reinforce existing inequities.

The discussion then turned to fairness in optimization. Here the focus was on how equity can be embedded directly into decision-making algorithms that allocate resources. By applying measures such as the Gini coefficient and the Hoover index, the study demonstrated how optimization can go beyond efficiency alone to include concerns about distribution and fairness. This perspective broadens the scope of optimization, re-framing it as a tool not just for minimizing costs, but also for designing systems that treat populations more equitably.

Based on this foundation, facility location modeling was applied to the placement of rehabilitation centers in Oklahoma. Different distance metrics were compared and fairness restrictions were introduced to evaluate how treatment access could be distributed more evenly across the state. The results highlight that without fairness considerations, rural and under-served communities risk being left behind. When fairness is included, solutions emerge that provide more balanced access to treatment facilities.

The dissertation also considered the prediction of treatment outcomes. Using SAMHSA data, machine learning models were developed to estimate the likelihood of treatment completion. The study found that decision trees and random forests provided the most consistent balance between accuracy and fairness, while methods such as oversampling and intersectional analysis

further reduced disparities across subgroups. These results underscore the importance of fairness-aware predictive modeling in practical decision making.

The system models were then constructed to compare treatment-first and punishment-first approaches to drug offenses. The treatment-focused model showed significant reductions in incarceration rates and public costs compared to punishment-first strategies. Sensitivity analysis identified the system transitions most responsible for cost variation, and the findings were consistent across counties of different sizes. This robustness strengthens the case for treatment-oriented policies. Sobol analysis helped nail down the values at which unknown parameter values could be in order to gauge the incarceration flow. This is better at prediction than simply using a 50/50 coin flip guess and is more accurate in prediction, providing clearer targets for policy makers.

Finally, a poverty-structured epidemic model was developed to explore how socioeconomic disparities affect the dynamics of addiction. By modeling poverty as a structural driver of relapse and limited recovery, the analysis revealed that inequities in income play a central role in addiction outcomes. The results suggest that incorporating fairness into treatment allocation can help reduce these disparities and improve recovery prospects for disadvantaged groups.

Together, these studies connect fairness in algorithms, optimization, and system modeling with practical policy design. They demonstrate that fairness is not only a technical criterion but also a social imperative. By linking predictive methods, access to treatment, and poverty-sensitive modeling, this dissertation provides evidence-based insights to improve rehabilitation outcomes, reduce incarceration, and support more equitable recovery systems. Since parts of the analysis use 2002–2005 data, treat the results as examples rather than as up-to-date estimates. Our goal here is to lay out the methods and model design. As newer data are released, we will rerun the analysis, recheck fairness and facility placement decisions, and report updated, policy-ready numbers.

Chapter 10

Future Work

In Chapter 4, looking at lexicographic optimization methods is something that could be explored. Looking at a time component could be added by setting a constraint to the max number of days a person can stay in rehab thus allowing for rotations of people in and out over time. Having alternate rehab paths available if the closest one has no open beds would be something to consider. A better source of data could be found, perhaps using incarceration data instead of drug court data would get a more accurate description of the true demand needed since many counties in Oklahoma do not have a drug court program in place. Setting an objective to find the optimal fairness value could be interesting as it changes significantly based on what this is set at.

In Chapter 5, the Free-to-Treatment rate needs to be thought about as a different way to obtain this statistic. The way it is currently being looked at is that we are looking at the percentage of people in drug court in for a drug crime. This skews the data if the whole population is in for drugs and not alcohol. One thing to think about would be to set the rate for free to treatment then same as free to incarcerated since they have the opportunity to go to either compartment based on the model. It would also be interesting to take the number of people incarcerated and in treatment over time into a network model as supply and demand components of a network of counties that are connected based on their distances. A sensitivity analysis could also be performed to look at how changing the rates in either direction affects the outcome of the model. Additionally, adding a fairness aspect to the model as to who gets treatment by county. This could be based off of population size or demographics. In Chapter 6, We think it would be interesting to look at the spatial distribution of each of the classes about the state. The paper, "Nonlinear Patterns in Urban Crime – Hotspots, Bifurcations, and Suppression" (Short et al., 2010) would be a good start for how to accomplish this. By using their methods in parallel with the model we have already created, a clearer and concise view of what is happening would be able to be seen. This paper also considers how to go about solving a nonlinear partial

differential equation with the use of weakly nonlinear analysis. It would also be interesting to utilize counties that have drug court data and use those counties to train the model in order to use it to predict what might happen in counties that have similar poverty, income, and percent rural details. Finally, we would like to compare the spatio-temporal clusters of addiction with respect to poverty as well as how the density of people within each of the SIRV2 model classes changes over time with the spatial distribution of the AA/NA meetings.

For Chapter 7, it would be beneficial to look at linear discriminant analysis for dimensionality reduction. Also, MCA or Multiple Correspondence Analysis would be good to look at for reducing dimensionality of categorical data. Researching kernel-based MCA would be interesting to see how that affects the accuracy of the model. Other methods of balancing the data that are more sophisticated than SMOTE using K-nearest neighbors or deep learning methods would be interesting to investigate. Looking deeper into the data to see what other information could be extracted. Perhaps a clustering analysis of what drugs/substances occur in treatment together in a patient and how those are related to age and demographics.

For future work overall, we can expand this research outside of the state of Oklahoma to the entire United States or regions of the United States and compare it with the counties of Oklahoma. We could also investigate things such as uncertainty and how it could fit into the various models. We can apply other machine learning models or distance metrics, respectively, to the chapters presented. In the compartmental model that did not include fairness, we could find a way to add fairness to that model instead of just creating a partial differential equation model like we did and changing the research direction. This way we could still look at how tax dollars are spent and have a focus on recidivism. Things such as intersectional fairness would be something to focus on, as that is an interesting topic that does not seem to have a lot of research applied to it. Overall, there are several ways in which we can improve upon our research, as there are always new methods to add as fairness is being studied.

Bibliography

Recidivism rates by state 2022. https://worldpopulationreview.com/state-rankings/recidivism-rates-by-state. Accessed: 2022-04-07.

Performance and outcome report on drug courts for fy'02-fy'05. Technical report, Oklahoma Department of Mental Health and Substance Abuse Services, 2005. URL https://oklahoma.gov/content/dam/ok/en/odmhsas/documents/a0002/fiscal-year-2002-2005.pdf. Accessed: 2022-04-07.

2005 oklahoma vital statistics births and deaths. Technical report, Oklahoma State Department of Health, 2008. URL https://oklahoma.gov/content/dam/ok/en/health/health2/documents/hci-vs2005report.pdf. Accessed: 2022-04-07.

The price of prisons - prison spending in 2015. Technical report, Vera Institute of Justice, 2015. Accessed: 2022-04-07.

Black's Law Dictionary. Thomson Reuters, 11th edition, 2019.

Diagnostic and Statistical Manual of Mental Disorders. 5th edition, 2021. URL https://doi.org/10.1176/appi.books.9780890425596. Accessed 21 Dec 2023.

Advanced Recovery Systems. What is the meth capital of the us | also meth capital of the world?, 2018. URL https://www.therecoveryvillage.com/meth-addiction/meth-capital/#gref. Accessed: 2018-04-01.

AIDSVu. Income inequality in canadian county, oklahoma, 2025. Retrieved February 12, 2025, from https://map.aidsvu.org/incomeinequality/county/ratio/none/none/canadian-county-ok-oklahoma?geoContext=national.

Anup Roop Akkihal. *Inventory Pre-positioning for Humanitarian Operations*. PhD thesis, Massachusetts Institute of Technology, 2006. URL https://dspace.mit.edu/handle/1721.1/36318.

Wilson E. Alvarez, Itelhomme Fene, Kimberly Gustein, Brenda Martinez, Diego Chowell, Anuj Mubayi, and Luis Melara. Prison reform programs and their impact on recidivism. Technical report, Arizona State University, 2012. Accessed: 2022-04-07.

ABA Standards for Criminal Justice: Sentencing. American Bar Association, 2017.

Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6): 54–61, June 2018. doi: 10.1145/3209581.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*, chapter 3: Classification. MIT Press, 2023a.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*, chapter 6: Understanding United States anti-discrimination law. MIT Press, 2023b.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Chapter 8: Fairness and Machine Learning in Practice. In *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023c. Available at https://fairmlbook.org/.

Rajan Batta, Anjan Ghose, and Udatta S. Palekar. Locating facilities on the manhattan metric with arbitrarily shaped barriers and convex forbidden regions. *Operations Research*, 38(3):427–436, 1990.

Nicholas A. Battista. A comparison of heroin epidemic models, May 2009. Unpublished paper.

Nicholas A. Battista. A comparison of heroin epidemic models, 2015. URL https://arxiv.org/abs/1510.04581.

Ruben Becker, Gianlorenzo D'Angelo, and Sajjad Ghobadi. On the cost of demographic parity in influence maximization, 2023.

Steven Belenko, Matthew Hiller, and Leah Hamilton. Treating substance use disorders in the criminal justice system. *Current Psychiatry Reports*, 15 (11):414, 2013a. doi: 10.1007/s11920-013-0414-z.

Steven Belenko, Matthew Hiller, and Leah Hamilton. Treating substance use disorders in the criminal justice system. *Current psychiatry reports*, 15(11): 414, 2013b. doi: 10.1007/s11920-013-0414-z.

Nathan Berg. Non-response bias. *ENCYCLOPEDIA OF SOCIAL MEASUREMENT*, Vol. 2:865–873, 2005. doi: 10.2139/ssrn.1694533. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1694533. 34 Pages Posted: 15 Oct 2010.

Richard Berk. How you can tell if the simulations in computational criminology are any good. *Journal of Experimental Criminology*, 4:289–308, 2008. doi: 10.1007/s11292-008-9053-5. URL https://doi-org.ezproxy.lib.ou.edu/10.1007/s11292-008-9053-5.

P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975. doi: 10.1126/science.187.4175.398.

Laura Cantlebary and Linwei Li. Facility location problem, 2020. URL https://optimization.cbe.cornell.edu/index.php?title=Facility_location_problem.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, pages 321–357, 2002.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. 2018.

Xinying V. Chen and John N. Hooker. A guide to formulating fairness in an optimization model. *Annals of Operations Research*, 2023. doi: 10.1007/s10479-023-05264-y.

Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we? In *Proceedings of the 46th International Conference on Software Engineering (ICSE 2024)*, page 13, New York, NY, USA, 2023. ACM. doi: 10.1145/nnnnnn.nnnnnnn.

Alex Chohlas-Wood, Joe Nudell, Keniel Yao, Zhiyuan (Jerry) Lin, Julian Nyarko, and Sharad Goel. Blind justice: Algorithmically masking race in charging decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 35–45, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462524. URL https://doi.org/10.1145/3461702.3462524.

Personal Communication. Phone interview with the referral center. Personal communication, nov 2022. Statistical data collection.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Jaclyn Cosgrove. Meth use continues to plague oklahoma, August 2017. URL http://newsok.com/article/5558008. Accessed: 2018-04-01.

Frank A. Cowell. *Measuring Inequality*. Oxford University Press, 2011.

CSG Justice Center. Oklahoma criminal justice data snapshot. Technical report, The Council of State Governments Justice Center, January 2025. URL https://justicereinvestmentinitiative.org/wp-content/uploads/2025/01/Oklahoma-Criminal-Justice-Data-Snapshot.pdf. Version 2.1, updated 01.06.2025.

Fernando G De Maio. Income inequality measures. *Journal of Epidemiology & Community Health*, 61(10):849–852, 2007. doi: 10.1136/jech.2006.052969.

Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. Springer, 2009.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011. URL http://arxiv.org/abs/1104.3913.

Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness and discrimination in information access systems. *CoRR*, abs/2105.05779, 2021. URL https://arxiv.org/abs/2105.05779.

Reza Zanjirani Farahani, Maryam SteadieSeifi, and Nasrin Asgari. Multiple criteria facility location problems: A survey. *Applied Mathematical Modelling*, 34(7):1689–1709, 2010. ISSN 0307-904X. doi: https://doi.org/10.1016/j.apm.2009.10.005. URL https://www.sciencedirect.com/science/article/pii/S0307904X09003242.

Felipe. The c parameter, 2019. URL https://queirozf.com/entries/choosing-c-hyperparameter-for-svm-classifiers-examples-with-scikit-learn.

James R. Foulds and Shimei Pan. An intersectional definition of fairness. *CoRR*, abs/1807.08362, 2018. URL http://arxiv.org/abs/1807.08362.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4): 136–143, 2021. doi: 10.1145/3433949.

K. Sampath Gajawada. Chi-square test for feature selection in machine learning, 2021. URL https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223.

Sandro Galea, Roland M. Merchant, and Nicole Lurie. The mental health consequences of covid-19 and physical distancing: The need for prevention and early intervention. *JAMA*, 324(6):603–604, 2020. doi: 10.1001/jama.2020.8942. URL https://jamanetwork.com/journals/jama/fullarticle/2766672. Accessed: 2025-06-19.

A. Ghosh et al. Characterizing intersectional group fairness with worst-case comparisons. In *AIDBEI*, 2021.

Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons, 2022.

Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023. International Joint Conferences on Artificial Intelligence Organization, August 2023. doi: 10.24963/ijcai.2023/742. URL http://dx.doi.org/10.24963/ijcai.2023/742.

LLC Gurobi Optimization. Gurobi optimizer reference manual, 2024. URL https://www.gurobi.com. Accessed November 10, 2025.

Gurobi Optimization, LLC. Specifying multiple objectives. https://www.gurobi.com/documentation/current/refman/specifying_multiple_object.html, 2023. Accessed: 2024-06-25.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, Stockholm, Sweden, 2018. PMLR.

Z. Irfan, F. McCaffery, and R. Loughran. *Evaluating Fairness Metrics*, volume 1840 of *Communications in Computer and Information Science*. Springer, Cham, 2023. URL https://doi.org/10.1007/978-3-031-37249-0_3.

Patricia Janis. Interview with women's firstep. Personal communication, nov 2022. Statistical data collection.

S.T. Jishan, R.I. Rashu, N. Haque, et al. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decis. Anal.*, 2(1), 2015. URL https://doi.org/10.1186/s40165-014-0010-2.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Journal Name*, Volume Number(Issue Number):Page Range, Month 2011. Received: 23 November 2010 / Revised: 23 August 2011 / Accepted: 16 November 2011 / Published online: 3 December 2011.

Bruce P Kennedy, Ichiro Kawachi, and Deborah Prothrow-Stith. Income distribution and mortality: cross-sectional ecological study of the robin hood index in the united states. *BMJ*, 312:1004, 1996. doi: 10.1136/bmj.312.7037.1004.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, pages 853–862. ACM, April 2018. doi: 10.1145/3178876.3186133. URL https://doi.org/10.1145/3178876.3186133.

Sara-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

Elle Lett and William G La Cava. Translating intersectionality to fair machine learning in health sciences. *Nature machine intelligence*, 5(5):476–479, 2023. ISSN 2522-5839.

Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness, 2018.

Alejandra Lopez, Naomi Moreira, Ariana Rivera, Bechir Amdouni, Baltazar Espinoza, and Christopher M. Kribs. Economics of prison: Modeling the dynamics of recidivism. Technical report, Arizona State University, 2018. Accessed: 2022-04-07.

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. Fair without leveling down: A new intersectional fairness definition. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9018–9032, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.558. URL https://aclanthology.org/2023.emnlp-main.558.

Trisha Mahoney, Kush R. Varshney, and Michael Hind. *AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning*, chapter Introduction. O'Reilly, 2020a.

Trisha Mahoney, Kush R. Varshney, and Michael Hind. *AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning*, chapter 1: Which Metric Should You Use? O'Reilly, 2020b.

Trisha Mahoney, Kush R. Varshney, and Michael Hind. *AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning*, chapter 2: Algorithms for Bias Mitigation. O'Reilly, 2020c.

Evi Maria, Edy Budiman, H. Haviluddin, and Medi Taruk. Measure distance locating nearest public facilities using haversine and euclidean methods. *Journal of Physics: Conference Series*, 1450(1):012080, 2020. doi: 10.1088/1742-6596/1450/1/012080.

David McMillon, Carl P. Simon, and Jeffrey Morenoff. Modeling the underlying dynamics of the spread of crime. *PLOS ONE*, 9(4):e88923, 2014. doi: 10.1371/journal.pone.0088923.

Héctor Mora, José M Mora-Pascual, Alberto García-García, and Paulino Martínez-González. Computational analysis of distance operators for the iterative closest point algorithm. *PloS one*, 11(10):e0164694, 2016. doi: 10.1371/journal.pone.0164694.

Animesh Mukherjee. *AI and Ethics: A Computational Perspective*, chapter 3: Algorithmic Decision Making. IOP Publishing, 2023a.

Animesh Mukherjee. *AI and Ethics: A Computational Perspective*, chapter 2: Descrimination Bias, and Fairness. IOP Publishing, 2023b.

Animesh Mukherjee. *AI and Ethics: A Computational Perspective*, chapter 6: Newly Emerging Paradigms. IOP Publishing, 2023c.

National Institute on Drug Abuse. Methamphetamine, September 2013. URL https://www.drugabuse.gov/publications/research-reports/methamphetamine. Accessed: 2018-03-25.

National Survey on Drug Use and Health. Key substance use and mental health indicators in the united states: Results from the 2022 national survey on drug use and health. Substance Abuse and Mental Health Services Administration, 2022. Accessed 21 Dec 2023.

Aileen Nielsen. *Practical Fairness*. O'Reilly, 2020. First Edition, ISBN: 9781492075738.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/abs/10.1126/science.aax2342.

Association of County Commissioners of Oklahoma (ACCO). Oklahoma map. URL https://www.okacco.com/oklahoma-map.html.

Bureau of Justice Statistics. Recidivism of prisoners released in 30 states, 2018.

Wlodzimierz Ogryczak, Hanan Luss, Michał Pióro, Dritan Nace, and Artur Tomaszewski. Fair Optimization and Networks: A Survey. *Journal of Applied Mathematics*, 2014(SI08):1 – 25, 2014. doi: 10.1155/2014/612018. URL https://doi.org/10.1155/2014/612018.

Oklahoma Department of Mental Health and Substance Abuse Services. Performance and outcome report 2002-2005. https://oklahoma.gov/content/dam/ok/en/odmhsas/documents/a0002/fiscal-year-2002-2005.pdf, May 2006. Retrieved December 30, 2023.

Oklahoma State Department of Health (OSDH). Vital statistics 2016 on oklahoma statistics on health available for everyone (ok2share), 2016. URL http://www.health.ok.gov/ok2share. Accessed: 2018-02-17.

Elizabeth Paluck, Porat Levy, Chelsey S Clark, and Donald P. Green. Prejudice reduction: What works? a review and assessment of research and practice. *Annual Review of Psychology*, 60:533–560, 2020.

U. Peters. What is the function of confirmation bias? *Erkenntnis*, 87:1351–1376, 2022. doi: 10.1007/s10670-020-00252-1. URL https://doi.org/10.1007/s10670-020-00252-1.

David J Rader. Deterministic operations research: Models and methods in linear optimization, 2010.

Prajit Ramachandran, Barret Zoph, and Quoc Le. Searching for activation functions. 2018. URL https://arxiv.org/pdf/1710.05941.pdf.

K. Roberts. Modeling meth addiction: R script with gini coefficient. https://github.com/KRoberts25/ModelingMethAddiction/blob/main/ModelingMethAddiction_W_Gini.R, 2023. Accessed: 2023-09-13.

Neal J. Roese and Kathleen D. Vohs. Hindsight bias. *Perspectives on psychological science*, 7(5):411–426, 2012. ISSN 1745-6916.

A Rose. Are face-detection cameras racist?, 2010. URL https://content.time.com/time/business/article/0,8599,1954643,00.html. January 24, 2024.

M. B. Short, A. L. Bertozzi, and P. J. Brantingham. Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression. *SIAM Journal on Applied Dynamical Systems*, 9(2):462–483, 2010. doi: 10.1137/090759069. URL https://doi.org/10.1137/090759069.

Substance Abuse and Mental Health Services Administration. Treatment episode data set (teds): 2019, 2021.

Systems Complex. System dynamics – a key to understanding complex systems, 2023. URL https://systemscomplex.com/system-dynamics/. Accessed: 2025-06-19.

Donald M. Taylor and Janet R. Doria. Self-serving and group-serving bias in attribution. *The Journal of social psychology*, 113(2):201–211, 1981. ISSN 0022-4545.

Ines Valentim, Nuno Lourenço, and Nuno Antunes. The impact of data preparation on the fairness of software systems. 2022. CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal.

G. Veenstra. Race, gender, class, and sexual orientation: intersecting axes of inequality and self-rated health in canada. *Int J Equity Health*, 10(3), 2011. URL https://doi.org/10.1186/1475-9276-10-3.

Ventana Systems, Inc. Vensim – system dynamics simulation software, 2024. URL https://vensim.com/. Accessed: 2025-06-19.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*, pages 1–7. ACM, May 2018. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

Sahil Verma, Michael Ernst, and Rene Just. Removing biased data to improve fairness and accuracy, 2021.

Dan Wagener. Financial assistance for substance abuse and drug rehab treatment. American Addiction Centers, 2022. URL https://americanaddictioncenters.org/rehab-guide/public-assistance. Accessed: 2022-04-07.

Emily Wildra and Tiana Herring. States of incarceration: The global context 2021. Prison Policy Initiative, 2021a. URL https://www.prisonpolicy.org/global/2021.html. Accessed April 21, 2022.

Emily Wildra and Tiana Herring. States of incarceration: The global context 2021. Prison Policy Initiative, 2021b. URL https://www.prisonpolicy.org/global/2021.html?gclid=Cj0KCQjwgYSTBhDKARIsAB8KukuXiNQycVGSq8hL-uA3fhB1IzqPVZjXpJPzBvs0SVilZXeqUPQTgyoaAo_sEALw_wcBlicy.org%2Fglobal%2F2021.html%3Fgclid. Accessed: 2022-04-07.

Emily Wildra and Tiana Herring. States of incarceration: The global context 2021, 2022. URL https://www.prisonpolicy.org/global/2021.html.

World Population Review. Meth use by state 2024, 2024. Retrieved February 12, 2025, from https://worldpopulationreview.com/state-rankings/meth-use-by-state.

World Population Review. Recidivism rates by state 2025. Online, 2025. URL https://worldpopulationreview.com/state-rankings/recidivism-rates-by-state. Data retrieved from WorldPopulationReview.com.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing, 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. 2017.

Ángel Pavón Pérez, Miriam Fernandez, Hasan Al-Madfai, Grégoire Burel, and Harith Alani. Tracking machine learning bias creep in traditional and online lending systems with covariance analysis. In *Proceedings of the 15th ACM Web Science Conference (WebSci '23)*, pages 184–195. ACM, April 2023. doi: 10.1145/3578503.3583605. URL https://doi.org/10.1145/3578503.3583605. Published: 30 April 2023.