

ABSTRACT

Title of Dissertation: Bi-modal flood Evacuation Planning using
Synthetic Population from A Generative
Adversarial Network Model

Mohammad Motalleb Nejad

Dissertation directed by: Professor Ali Haghani, Department of Civil and
Environmental Engineering

Flooding is a significant disaster that threatens the lives of millions of people within the United States. The 100-year coastal floodplain in the United States accommodates over 60,000 miles of major roadways. Moreover, United States Census Data reveals that more than 50% of the U.S. population lives close enough to the shoreline to be influenced by potential flooding. One major problem in this regard is identifying people who strongly depend on public transportation.

A Conditional Wasserstein Generative Adversarial Network model and a couple of statistical approaches are introduced to capture the interdependency of people's characteristics in hazard zones. This method can provide a robust tool to estimate the demand for evacuation. An available statistical tool, namely the copula model, and its limitations are discussed, and the superiority of GANs model in addressing these limitations is mentioned. Also, a dynamic mixed integer linear problem has been developed to use the estimated demand and minimize the total evacuation time, the total departure time, and the number of utilized shelters. The model is solved for a simple network. A heuristic genetic algorithm is designed to solve the problems of actual life scales.

BI-MODAL FLOOD EVACUATION PLANNING USING SYNTHETIC
POPULATION FROM A GENERATIVE ADVERSARIAL NETWORK MODEL

by

Mohammad Motaleb Nejad

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

2022

Advisory Committee:

Professor Ali Haghani, Chair

Professor Gerrit Knaap, Dean's Representative

Professor Paul Schonfeld

Professor Cinzia Cirillo

Dr. Sevgi Erdogan

© Copyright by
Mohammad Motalleb Nejad
2022

Acknowledgments

Thank God for all your blessings and your kind hands.

I want to express my gratitude to my advisor, Dr. Ali Haghani, who has helped me patiently throughout this study. I would also like to thank my committee members for letting my defense be an enjoyable moment and your thoughtful comments and suggestions.

I also sincerely appreciate my wife and my lovely daughter, Nora, whose patience and kindness helped me throughout the Ph.D. journey.

Table of Content

ACKNOWLEDGMENTS	ii
LIST OF FIGURES.....	v
LIST OF TABLES	viii
CHAPTER 1. INTRODUCTION	1
1.1. MOTIVATION	1
1.2. SCOPE OF STUDY	6
1.3. DISSERTATION OUTLINE	8
CHAPTER 2. TECHNICAL PROBLEM STATEMENT	10
2.1. SMALL AREA POPULATION ESTIMATION	10
2.2. GENERATIVE MODELS AND TRANSPORTATION DATA.....	13
2.3. OPERATIONAL EVACUATION PROGRAMMING MODELS.....	13
CHAPTER 3. LITERATURE REVIEW.....	15
3.1. DATA NEEDS FOR EVACUATION OF SMALL AREAS.....	15
3.2. SYNTHETIC POPULATION.....	16
3.3. GENERATIVE MODELS	19
3.4. BIMODAL EVACUATION MODELS.....	22
CHAPTER 4. GENERATIVE COPULA MODEL AND ITS LIMITATIONS	25
4.1. COPULA MODEL	25
4.1.1. <i>Introduction to the generative copula models.....</i>	25
4.1.2. <i>Generating synthetic population using Copula (methodology)</i>	25
4.1.3. <i>Anne Arundel County Application</i>	31
4.1.4. <i>Identification of Low-income People who do not own cars.....</i>	36
4.1.5. <i>Limitations of the copula model.....</i>	40
CHAPTER 5. GANS AND SYNTHETIC POPULATION GENERATION	43
5.1. INTRODUCTION TO GAN MODEL FOR SOCIODEMOGRAPHIC DATA GENERATION	43
5.1.1. <i>Basic GANs model for synthetic population generation.....</i>	43
5.1.2. <i>Challenges to be addressed</i>	48
5.1.3. <i>Generating fake data feed for discriminator pre-training</i>	49
5.1.4. <i>Applicability of GANs to one-dimensional data.....</i>	50
5.2. DATA PREPARATION	53
5.3. WASSERSTEIN GANS	55
5.4. THE STRUCTURE OF WGANs MODELS EXAMINED IN THIS STUDY	57
5.5. SAMPLE GENERATION FOR EACH MICRO-AREA.....	62
5.6. RESULTS OF THE GANS MODELS ON SOCIODEMOGRAPHIC DATA	66
CHAPTER 6. BIMODAL EVACUATION PLANNING MODEL	75
6.1. UNDERSTANDING THE REQUIREMENTS FOR FLOOD EVACUATION PLANNING.....	75
6.2. EVACUATION OPTIMIZATION PROBLEMS	77
6.2.1. <i>Flood evacuation problem specifications.....</i>	77
6.2.2. <i>Problems associated with flood evacuation operations.....</i>	80
6.2.3. <i>Highly non-linear nature of the evacuation problems.....</i>	81

6.3.	PROPOSED EVACUATION MODEL.....	84
6.3.1.	<i>Formulating the problem</i>	84
6.3.2.	<i>Application of TET MILP on a simple network</i>	89
CHAPTER 7.	A HEURISTIC SOLUTION TO THE TET MILP EVACUATION MODEL.....	94
7.1.	TET MILP LIMITATIONS AND THE PROPOSED HEURISTIC SOLUTION'S STRENGTHS	94
7.2.	ALGORITHM DESIGN AND ASSUMPTIONS	95
7.3.	RESULTS OF THE HEURISTIC ALGORITHM	103
CHAPTER 8.	CONCLUSION AND FUTURE STUDY PATHWAYS	111
8.1.	CONCLUSIONS AND HIGHLIGHTS	111
8.2.	FUTURE RESEARCH SUGGESTIONS	114
REFERENCES		118

List of Figures

FIGURE 1. RESCUE EFFORTS IN NEW ORLEANS IN THE DAYS AFTER HURRICANE KATRINA (SOURCE: HTTPS://WWW.INDIEWIRE.COM/2020/03/FLOODLINES-PODCAST-HOST-VANN-NEWKIRK-HURRICANE-KATRINA-1202220318/)	4
FIGURE 2. THOUSANDS OF HOUSTONIAN EVACUEES IN MASSIVE, CONGESTED TRAFFIC AS HURRICANE RITA APPROACHES (SOURCE: PUBLIC DOMAIN)	6
FIGURE 3. TWO LAYERS OF GEOGRAPHICAL AREA, PUMA LEVEL (P) AND CENSUS TRACT LEVEL (C)	12
FIGURE 4. BOOTSTRAP FLOWCHART TO CALCULATE THE P-VALUE FOR A SELECTED COPULA.	29
FIGURE 5. FLOWCHART OF THE POPULATION-SYNTHESIZING ALGORITHM.	31
FIGURE 6. THE MEDIAN INCOME FOR DIFFERENT CENSUS TRACTS WITHIN ANNE ARUNDEL COUNTY DATA (NOTE THAT THE DATA WAS OBTAINED FROM INTERACTIVE MAPS AVAILABLE AT THE UNITED STATES CENSUS BUREAU WEBSITE IN 2018, CURRENTLY AVAILABLE IN A DIFFERENT FORMAT AT HTTPS://WWW.CENSUS.GOV/PROGRAMS- SURVEYS/GEOGRAPHY/DATA/INTERACTIVE-MAPS.HTML).	32
FIGURE 7. ILLUSTRATION OF PUMAS AND CENSUS TRACTS IN ANNE ARUNDEL COUNTY.	33
FIGURE 8. A SAMPLE VIEW OF IRS DATA USED TO GENERATE CDF FUNCTIONS FOR INCOME.	35
FIGURE 9. A SAMPLE VIEW OF CROSSWALKS BETWEEN ZIP CODE AND CENSUS TRACTS.	36
FIGURE 10. CAR OWNERSHIP AND INCOME LEVEL FOR CENSUS TRACTS WITHIN ANNE ARUNDEL COUNTY.	40
FIGURE 11. ILLUSTRATIVE REPRESENTATION OF THE SECOND ISSUE WITH SMALL AREA ESTIMATIONS USING <i>COPULA</i>	42
FIGURE 12. DISCRIMINATOR PRETRAINING STEP USING THE SAMPLES FROM REAL AND FALSE DATA	44
FIGURE 13. THE STRUCTURE OF AN EXAMPLE GENERATOR	45
FIGURE 14. THE STRUCTURE OF A COMPILED GANS MODEL	45
FIGURE 15. THE TRAINING PROCESS FOR A COMPILED GANS MODEL	47
FIGURE 16. A SIMPLE SPIRAL FUNCTION USED TO EXPLORE THE APPLICABILITY OF GANS TO GENERATE THE SAMPLES FROM MARGINAL DISTRIBUTIONS OF THE VARIABLES	50
FIGURE 17. SYNTHETIC DATA GENERATED BY TRAINED GANS AFTER 200, 5000, AND 10000 ITERATIONS	51
FIGURE 18. ACCURACY OF THE DISCRIMINATOR IN IDENTIFYING THE FAKE AND REAL SAMPLES	52
FIGURE 19. ACCURACY OF THE DISCRIMINATOR IN IDENTIFYING THE FAKE AND REAL SAMPLES	52
FIGURE 20. (A) GENERATING CDFS AND INVERSE CDFS FOR 10,000 RANDOM SAMPLES FROM SPIRAL FUNCTION (B) THE GENERATOR SAMPLES AFTER 50, 200, AND 10,000 TRAINING ITERATIONS	53
FIGURE 21. TWO SAMPLES REPRESENTED AS (A) A ONE-HOT VECTOR, AND (B) AN IMAGE	55
FIGURE 22. CRITIC STRUCTURE USED FOR MOST OF THE MODELS IN THIS STUDY	57
FIGURE 23. GENERATOR'S STRUCTURE FOR GEN-1 MODEL	58
FIGURE 24. GENERATOR'S STRUCTURE FOR GEN-2 MODEL	59
FIGURE 25. THE STRUCTURE OF THE DISCRIMINATOR AND GENERATOR FOR GEN-CONV2D	61
FIGURE 26. AN EXAMPLE OF CONDITIONAL WGANs USING THE DATA FROM TWO DIFFERENT SPIRAL FUNCTIONS, (A) SAMPLES GENERATED FOR THE FIRST SPIRAL FUNCTION, (B) SAMPLES GENERATED FOR THE SECOND SPIRAL FUNCTION, (C) LOSSES FOR THE CRITIC AND THE GENERATOR FOR THE FIRST 6,000 ITERATIONS (EACH DATA POINT IS THE MEAN FOR THE LAST 20 ITERATIONS)	64

FIGURE 27. INPUTS TO CONDITIONAL WGANs (A) GENERATOR INPUT (B) CRITIC INPUT IN THE MLP MODEL (C) CRITIC INPUT IN THE CONVOLUTIONAL MODEL	65
FIGURE 28. MODE COLLAPSE AFTER TRAINING CONVOLUTIONAL GANs WITH GEN-CON2D MODEL	67
FIGURE 29. WGANs RESULTS WITH GEN-CONV2D MODEL AFTER (A) 100 (B) 1200, AND (C) 3000 ITERATIONS	67
FIGURE 30. (A) THE PROBABILITY DISTRIBUTION FUNCTION FOR A SET OF 170,756 SAMPLES FROM TRAINED GEN-1 MODEL (B) PERCENTAGE DEVIATION FROM REAL DATA FOR 100 SETS OF 170,756 SAMPLES FROM TRAINED GEN-1 MODEL	68
FIGURE 31. PERCENTAGE DEVIATION FROM REAL SAMPLES FOR 100 SETS OF 170,756 SAMPLES FROM (A) WGANs TRAINED WITH GEN-2 GENERATOR (B) WGANs TRAINED WITH GEN-3 GENERATOR	69
FIGURE 32. THE LOSS FUNCTION FOR WGANs TRAINED WITH GEN-1, GEN-2, AND GEN-4	70
FIGURE 33. PROBABILITY DISTRIBUTION FUNCTION FOR SYNTHETIC DATA PRODUCED BY CONDITIONAL WGANs MODEL FOR THREE SUBAREAS OF PUMA	70
FIGURE 34. PERCENTAGE DEVIATION FROM THE REAL SAMPLE DISTRIBUTION FOR SYNTHETIC DATA (100 SETS OF SAMPLES) PRODUCED BY THE CONDITIONAL WGANs MODEL FOR THREE SUBAREAS OF PUMA	71
FIGURE 35. THE CHANGE IN NON-ORTHOGONALITY AS THE TRAINING EPOCHS INCREASE	72
FIGURE 36. THE CHANGE IN NON-ORTHOGONALITY AS THE TRAINING EPOCHS INCREASE	74
FIGURE 37. (A) FEMA SHELTERS ARE SCATTERED WITHIN AND OUTSIDE ANNE ARUNDEL COUNTY. (B) TRANSPORTATION NETWORK AND SHELTERS WITHIN ANNE ARUNDEL COUNTY	76
FIGURE 38. SPATIAL DISTRIBUTION OF SHORTEST PATH TO FEMA SHELTERS. GRAY AREAS SHOW DYSCONNECTIVITY FOR DIFFERENT SEA-LEVEL RISE SCENARIOS	77
FIGURE 39. ILLUSTRATION OF DIFFERENT OBJECTIVE FUNCTIONS IN EVACUATION PROBLEMS	79
FIGURE 40. DIFFERENT STRATEGIES THAT CAN BE USED TO ACHIEVE A REALISTIC BIMODAL EVACUATION PROBLEM	81
FIGURE 41. THE STRUCTURE OF THE BIMODAL EVACUATION PROBLEM FOR CENSUS TRACT LEVEL FLOOD EMERGENCY ALERTS	81
FIGURE 42. THE STRUCTURE OF THE NETWORK REDESIGNED FOR THE PROPOSED EVACUATION MODEL	85
FIGURE 43. BILEVEL DISCRETIZATION OF THE TIME	87
FIGURE 44. SCHEMATIC REPRESENTATION OF THE CONSERVATION OF FLOW AT NODE J	89
FIGURE 45. A SIMPLE EXAMPLE NETWORK DESIGNED TO EXAMINE THE PROPOSED FORMULATION	90
FIGURE 46. GRAPHICAL REPRESENTATION OF THE SOLUTION TO THE EXAMPLE TET MILP PROBLEM	93
FIGURE 47. SIMPLE NETWORK WITHOUT AN AUXILIARY SOURCE LINK USED IN DEVELOPING THE HEURISTIC ALGORITHM	95
FIGURE 48. A TYPICAL GENOTYPE AND THE INFORMATION STORED IN EACH GENE	97
FIGURE 49. CAPACITY AND TIMETABLES FOR THE SIMPLE NETWORK	98
FIGURE 50. STEPS TO CALCULATE THE ENTRY POINTS FOR A LINK PAIR IN A GENE PATH (STEPS 1 AND 2) AND LEAVE POINT FOR THE FIRST LINK IN THE LINK PAIR (STEP 3)	99
FIGURE 51. CROSSOVER AND OBTAINING SIX CHILDREN OUT OF TWO PARENTS	100
FIGURE 52. CONCEPTUAL DEMONSTRATION OF THE GINI INDEX CONCEPT	101
FIGURE 53. NSGA-II SORTING MECHANISMS FOR AN EXAMPLE OF MINIMIZATION OF A QUADRATIC FUNCTION WITH TWO OBJECTIVES, (A) PARETO-FRONT PROPAGATION, (B) NON-DOMINATED SORTING, AND (C) CROWDING DISTANCE	102
FIGURE 54. HEURISTIC NSGA-II ALGORITHM DESIGNED FOR THE EVACUATION OPERATION	103

FIGURE 55. UNI-OBJECTIVE GA ALGORITHM RESULTS FOR A SIMPLE EXAMPLE FOR VERY FEW ITERATIONS	104
FIGURE 56. (A) IMPEDANCE CALCULATION FOR THE TOY NETWORK EXAMPLE, AND (B) RESULTS OF GA ON TOY NETWORK IMPLEMENTING TIME AND CAPACITY UPDATES	105
FIGURE 57. PARETO-FRONTS OBTAINED FOR SEVEN ITERATIONS FOR MULTI-VARIATE NSGA-II APPLIED TO TOY NETWORK IMPLEMENTING TIME AND CAPACITY UPDATES	106
FIGURE 58. NETWORK FOR ANNE ARUNDEL COUNTY (FROM THE MSTM MODEL) AND DEMAND, SHELTER, AND AUXILIARY SINK NODES	107
FIGURE 59. RESULTS OF RUNNING THE ALGORITHM WITH A POPULATION OF 60 GENOTYPES FOR 50 ITERATIONS	110
FIGURE 60. NETWORK REDESIGN TO MODEL CONTRAFLOW	115

List of Tables

TABLE 1. THREE MOST POPULAR ARCHIMEDEAN COPULAS.	27
TABLE 2. ESTIMATED PARAMETERS FOR CLAYTON COPULA FOR CENSUS TRACTS WITHIN ANNE ARUNDEL COUNTY	37
TABLE 3. RESULTS OF BINOMIAL MODEL USED TO ESTIMATE CAR OWNERSHIP.	38
TABLE 4. THE ALGORITHM TO TRAIN THE GANS MODEL.	48
TABLE 5– DATA ATTRIBUTES AND THEIR CLASSIFICATIONS	54
TABLE 6– NON-ORTHOGONALITY MEASURED FOR COVARIANCES OF THE REAL AND THE DATA SYNTHESIZED WITH WGANs TRAINED WITH GEN-1, GEN-2, AND GEN-3.	72
TABLE 7– – NON-ORTHOGONALITY MATRIX FOR ALL POSSIBLE PAIRS OF COVARIANCES CALCULATED FOR REAL AND SYNTHESIZED DATA USING CONVOLUTIONAL WGANs TRAINED WITH GEN-2	73
TABLE 8. THE PARAMETERS, DECISION VARIABLES, AND NOTATIONS USED TO FORMULATE THE NON-LINEAR PROBLEM	82
TABLE 9. THE PARAMETERS, DECISION VARIABLES, AND NOTATIONS USED TO FORMULATE THE TET MILP	85
TABLE 10. THE TRAVEL TIME MATRIX FOR EACH LINK AT EACH TIME STEP	90
TABLE 11. THE PARAMETERS USED FOR THE EXAMPLE NETWORK EVACUATION PROBLEM	91
TABLE 12. THE SOLUTION TO THE EXAMPLE TET MILP PROBLEM	92
TABLE 13. FINAL UPDATED TIMETABLE OBTAINED FROM THE BEST SOLUTION FOR GA APPLIED TO THE TOY NETWORK EXAMPLE	105
TABLE 14. PACKET SIZE AND PACKET NUMBER CALCULATED BASED ON CAR-LESS PEOPLE PERCENTAGE AND VEHICLE TYPE	108
TABLE 15. MSTM MODEL BPR FUNCTION PARAMETERS	109

Chapter 1. Introduction

1.1. Motivation

Flooding is a significant disaster that threatens the lives of millions of people within the United States. The 100-year coastal floodplain in the United States accommodates over 60,000 miles of major roadways (Douglass et al., 2005). Moreover, United States Census Data reveals that more than 50% of the U.S. population lives close enough to the shoreline to be influenced by potential flooding. More than 37 million people and their associated businesses have moved and been established in the coastal area only within the last three decades. The migration rate grows due to increased development along the scenic coastlines and bay shipment terminals (U.S. Commission on Ocean Policy, 2004). With these developments, more infrastructure and communities will be exposed to severe coastal flood strikes and sea-level rise.

Different disasters require different levels and shares of mitigation, response, and recovery strategies, which are the main three phases of any disaster management program (see McLoughlin 1985; Bullock and Haddow, 2012; Bumgarner, 2008). Human-induced environmental hazards such as global warming are progressive exacerbations of an unknown situation, usually identified by its already emerging devastating effects. In the case of terrorist attacks, extremists typically commit intentional human-made threats without any notice. Also, the exact or even approximate occurrence time for some natural hazards is not identifiable beforehand. For example, only a very rough prediction of earthquakes is possible, given that the activity of tectonic plates under the ground surface is an uncertain event (Geller, 1997). Mitigation and recovery are the leading strategies and focus of the disaster research programs for these hazards.

On the other hand, some natural hazards are not necessarily strange random events. A natural hazard might be identifiable with science and technology aid before it triggers, which

justifies investments in a disaster management system's preparedness and response phases. For a flood, which is the focus of this study, the probability of identifying a potential inundation strike for a specific geographical area is high, even days before its onset (Jones, 2004).

However, a high-quality flood preparedness plan should involve a robust risk assessment procedure to determine the potential of a location for a disaster (Apel et al., 2004). Although a comprehensive risk assessment is most relevant when disaster mitigation strategies are being developed, a byproduct of such analysis is the identification of at-risk zones. The risk is typically defined as the probability of a disaster multiplied by its negative consequences (UNISDR, 2016; Mamuji and Etkin, 2019). From the economic point of view, this definition translates to the monetary value of the damage that an event will cause. Assuming that the results of proper and meticulous risk identification research are available, the remaining issue is identifying the demand, resources, and development of practical planning models to manage the hazard, which drives the motivation for this study.

According to the yearly reports on the number of significant disasters published by the US Federal Emergency Management Agency (FEMA, 1996), the number of natural or human-made disasters that exceed the communities' local capacities, require state and federal assistance, and require an evacuation is between 45 and 75 cases (Transportation Research Board of the National Academies). Therefore, transportation engineers, planners, and policymakers are rigorously working to identify successful evacuation strategies to resolve one of the most challenging problems that threaten today's societies. With the unprecedented increase in frequency and intensity of climatic disasters experienced in the U.S. and worldwide that cost thousands of lives lost each year, developing methods and tools to help identify effective evacuation plans remains a crucial research need. The conditions under which evacuations happen and the characteristics of

the impacted communities may vary significantly depending on many factors, including but not limited to the nature and the intensity of the hazard, geography, sociodemographics, institutional structures, and the preparedness level. However, it is an unfortunate fact that often, low-income and typically residents who do not own cars are the ones that take the most brutal hit, as observed in hurricanes Katrina, Rita, and Maria cases.

In 2005, Hurricane Rita (the 4th most intense Hurricane recorded in the Atlantic Basin) made landfall which caused severe to catastrophic flooding and wind damage. New Orleans and Houston officials were criticized and applauded, respectively, for their evacuation performance. The critical factor in Houston officials outperforming New Orleans officials was that Houston officials managed to efficiently use their bus transit services and resources to evacuate the people who did not own cars before this extreme event. The remarkable statistics are that nearly 8 percent of the 1.3 million New Orleans residents who were left helpless in the urban area had no access to automobile transportation. At the same time, half of the car-less people evacuated the area used the empty seats provided by their close friends and relatives (Wendell, 2006). Figure 1 shows the rescue efforts continued in New Orleans days after Hurricane Katrina had struck the area. The lack of accurate predictions for evacuation facilities had led to days of rescue efforts and inevitable fatalities (Brown, 2006). These observations suggest that matching the available resources (including the transit and car-sharing possibilities) with the trip productions during an extreme event is crucial in successfully evacuating people who do not own cars. Therefore, it is critical to develop specific strategies and plans that consider the location of vulnerable populations and whether they have the means to travel to safe zones or not.



Figure 1. Rescue efforts in New Orleans in the days after Hurricane Katrina (Source: <https://www.indiewire.com/2020/03/floodlines-podcast-host-vann-newkirk-hurricane-katrina-1202220318/>)

Demography is the study of a population utilizing the available characteristics such as age, race, gender, access to personal cars, and other social attributes. A set of either available or analytically obtainable population characteristics is usually referred to as sociodemographic characteristics. A population's sociodemographic features help model the travel demand and transportation systems optimized for pre and post-hazard operations. Under an evacuation operation, especially in the US, people with personal cars typically drive to destinations other than emergency shelters. In contrast, low-income people who do not own cars mostly rely on public transportation.

Additionally, in the presence of lead-time in evacuation, group activities (e.g., family members gathering or parents going to their children's house) have been observed during the evacuation period (Stern, 1989). These activities may be partially due to the lack of access to personal cars or the age of older adults who are too old to drive a vehicle in harsh situations. If authorities do not provide public transportation, the people who do not own cars can only rely on their relatives and friends. While not all low-income people suffer from a lack of access to personal vehicles, they constitute the majority of the public transportation users under normal conditions.

According to the US Census Bureau, a total of 21% of transit-using households are only paid less than \$15,000 as their annual income, and 46% of riders have no access to a personal vehicle. Their transportation to shelters or safe zones needs to be provided by transit agencies under emergency conditions. Such a case requires reliable predictions on the number of low-income people who do not own cars for small enough areas to distribute reasonable evacuation resources without significant accessibility issues.

Once the location and percentage of low-income people who do not own cars are identified, the crucial step is to develop a planning model that optimizes the utilization and allocation of the facilities, resources, and existing infrastructure to evacuees to efficiently evacuate the identified danger zones. During an evacuation, the existing infrastructure gets crowded with an inevitable traffic volume whose consequences are unprecedented. Hurricane Floyd (1999) and hurricanes Katrina and Rita (2005) required millions of people to evacuate, creating the most extensive traffic jams in U.S. history (Transportation Research Board of the National Academies, CNN LIVE, 2001; CNN LIVE at Daybreak, 2001). According to a 2006 report to the Texas House of Representatives (HRO, 2006), during the evacuation for Hurricane Rita (2005), at least 60 people died only due to heatstroke, cars catching on fire, and traffic accidents. Cars ran out of gas on the freeway, which blocked the lanes, and people spent 20 hours going 100 miles in very congested stop-and-go traffic. Figure 2 shows a view of congested traffic during Hurricane Rita. Since 2005, only Houston added over 1.5 million to its population, meaning that the same Hurricane can cause far more fatalities than observed in 2005. The collectors, arterials, and freeways will be more congested, and without a practical multimodal evacuation plan, a worse experience of failure will be inevitable.



Figure 2. Thousands of Houstonian evacuees in massive, congested traffic as Hurricane Rita approaches (Source: Public Domain)

It should be noted that it is almost impossible to eliminate fatalities and damages caused by intensive natural hazards, such as earthquakes, hurricanes, floods, and pervasive wildfires, as the capacity of the available infrastructures and availability of the resources are subject to restrictive constraints. According to the International Federation of Red Cross and Red Crescent Societies (IFRC), when a community's capacity to cope with widespread losses is not high enough to prevent the community's malfunctioning, the phenomenon is identified as a disaster. Malfunctioning of a community happens when the utilization of its resources is not enough to control actions required for the preparedness and response phases of a threatening hazard. Therefore, developing optimization and planning models is crucial in efficiently employing the available resources and facilities for evacuation purposes and identifying the required additional supplies to further conquer the disasters' destructive nature.

1.2. Scope of Study

Estimating and alleviating the consequences of a natural hazard to the human environment requires accurate information on the impacted areas' geographic and demographic characteristics and an efficient planning model for extreme situations. One necessary aspect of these models is

the accessibility of low-income, car-less, and older people to transportation resources. This issue's immediate requirement is to accurately estimate the number of people who strictly depend on publicly provided transportation facilities. The precise prediction of the number of people who do not own cars needs strong models validated through robust statistical techniques and tests. Having obtained the demand for publicly available evacuation transport systems, developing accurate emergency evacuation models to minimize these disasters' negative consequences is the next step. In terms of transportation and traffic congestion, these negative consequences can be translated into the required time for different actions during specific phases of an evacuation or defined as the number of people to whom the evacuation system fails to respond to their needs. Therefore, an efficient evacuation model aims to minimize the evacuation operation's temporal dimensions or maximize the total evacuees that can be successfully cleared from the risk zones. The focus of this study, therefore, is to propose reliable solutions for the following interrelated issues regarding evacuation:

- A) Developing a reasonable population synthesizer that can be utilized to identify low-income residents' location in conjunction with car-ownership status in a target area of interest for evacuation. The main objective is to identify low-income residents who do not own cars and impose many challenges in preparing for evacuation, as they solely rely on public transportation services.
- B) Developing an evacuation planning model that uses a bimodal transportation system (transit and personal car) for evacuation instances and efficiently employs available resources and possible strategies.

The bridge between these two goals is the public transportation demand, which indirectly determines the need for roads within the flooding area by personal cars. This research suggests a

robust method to generate a synthetic population and identify low-income people with no access to a vehicle in an area with a medium to high risk of a natural hazard. The model's output paves the road toward the development of efficient evacuation planning for transportation services. The next chapter (Problem Statement) will discuss the specific features of the synthetic population required for evacuation problems.

1.3. Dissertation Outline

This dissertation is organized as follows. Chapter 2 provides a clear problem statement on specific aspects of the evacuation operation and highlights the contributions and methods used in this research. Chapter 2 further clarifies how our study differs from the other analyses. Chapter 3 reviews the literature on this study's three critical components: population synthesis approaches, generative models, and evacuation optimization models. The chapter will mainly focus on evacuation data needs, existing population synthesis methods in transportation studies, and the available literature on optimization models.

Chapter 4 discusses the copula approach previously used to synthesize the population. Simple binomial choice models that can be utilized to estimate the number of people who do not own cars from the synthesized population are mentioned using a simple example from a previous study. The copula model's deficiencies and shortcomings are discussed using a case study for Anne Arundel County in Maryland. Chapter 5 focuses on developing multiple multivariate generative adversarial networks (GANs) models to overcome the copula model's deficiencies. The chapter further discusses the results of the Conditional Wasserstein GANs model in generating data for sub-areas of a larger geographical area. In addition, a couple of additional statistical methods are suggested to understand the applicability of synthesizers in different problems. Chapter 6 first demonstrates the complexity and nonlinearity of the bimodal evacuation problem and then presents an optimization model that aligns with the public transportation demand estimations from the

GANs model introduced in Chapter 5. Chapter 7 develops a genetic algorithm-inspired heuristic solution for real-size networks, which creates Pareto-Fronts for simultaneous minimization of total travel time and overutilization of the shelters. Finally, Chapter 8 concludes this research by summarizing the achievements and proposing future research suggestions.

Chapter 2. Technical Problem Statement

2.1. Small Area Population Estimation

People's behavioral patterns and characteristics affect their car-ownership status and indirectly influence their decision to own a personal car. These effects can be captured by deriving the statistical correlation between individuals' characteristics and car-ownership status by utilizing various data sources. It is possible to identify people who do not own cars based on their socio-demographic characteristics, given that a strong correlation is obtained between their characteristics and car ownership status. The biggest challenge is that the whole population's characteristics are usually unknown due to practical reasons, or data are not accessible to the public due to privacy and security reasons. Only the attributes of a random sample are usually available for public use. These samples are typically obtained from a population that belongs to a larger geographic area than the area experiencing the hazard, usually a tiny nest within the sampled area.

An example of the available data is the Public Use Microdata Area (PUMA) level information that is available from American Community Survey (ACS) data (United States Census Bureau, 2016). The PUMA level information (PUMS) is gathered and updated annually for each PUMA in the United States. Each PUMA consists of several areas with different demographic characteristics (e.g., income levels, race, employment). Income values for each PUMA vary within an extensive range. An appropriate sublevel of PUMA that may reduce this variation is the census tract, which is geographically small enough to show low variability in its residents' income levels. Fortunately, census tracts are perfectly nested within PUMAs. However, they are minimal, and very few samples (or none) from ACS observations might fall into each census tract. Therefore, these samples are not adequate to represent the whole population of a census tract.

To be qualified as reliable synthetic data, a set of synthesized agents must satisfy two essential criteria. First, the synthetic data must preserve the dependencies between the attributes

for each geographical area. This criterion relies on the multivariate characteristics of a dataset. For example, it is reasonable to assume that individuals who share the same sociodemographic features (e.g., level of education, household income, marital status) tend to have the same tendency to have children or buy a car. Second, the cumulative distribution of the variables for all attributes of the interest should be satisfied. The satisfaction of these two criteria guarantees the validity of a synthetic population. In purely mathematical terms, these two criteria ensure that the correlation between the cumulative distribution of the variables and their joint distribution is the same for the synthesized population and the actual population.

The main problem regarding the synthetic population generation for a small area is the availability and quality of sample data that adequately represent the sociodemographic features of the people living in that area. When it comes to small area estimation, these data are usually neither enough nor reliable to be used to generate a synthesized population. Moreover, the actual population data is not available due to privacy reasons and difficulties and limitations in data acquisition. Figure 3 illustrates a geographical area in Anne Arundel County, Maryland, comprised of two layers: the large area with an ocean shoreline on the east (P) and multiple small areas within this large area (Cs). Layer P (PUMA 1203) is a large geographical area where the available data samples are enough to describe the whole community. Layers that are labeled C (Census Tracts), on the other hand, are the sublayers of layer P. Each layer C only includes a very small portion of the data samples from layer P. This small data is insufficient to obtain the dependencies between different attributes of the sublayer's total population and generate the synthetic population.

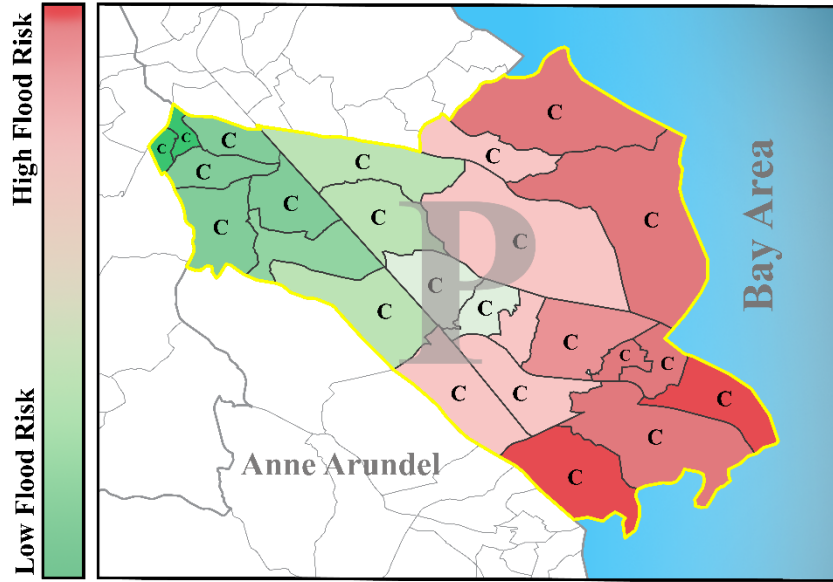


Figure 3. Two layers of geographical area, PUMA level (P) and Census Tract level (C)

Although the sample data is adequate to synthesize the population for area P, the resulting synthetic population is not suitable for evacuation planning for smaller areas. The most critical problem is that not all the subareas in layer P need to be evacuated before, during, or after any natural hazard. For example, in Figure 3, the sublayers in the eastern border of the whole area (layer P) are more prone to flood, and they need to be evacuated before the flooding to prevent possible casualties and fatalities.

A potential solution to this problem is contingent upon making reasonable assumptions. Suppose the dependencies between the characteristics of the people living in area P are the same for the sublayers. In that case, the sample from layer P can be used to obtain the data's structural dependency. Then, the cumulative distribution function (CDF) of each sublayer C's attributes can be utilized to generate synthesized data for each of them.

In summary, the structural dependency of the data for each census tract cannot be obtained from a few samples that are inherited from PUMA-level sample data. However, assuming that dependency is a nonvariant of each PUMA, one may use this dependency to generate a synthetic population from the CDF of each census tract's attributes. The feasibility of capturing the structural

dependency of data by using two different generative models is described in Chapter 4. The deficiency of the previously used Copula Generative model is illustratively described with an example.

2.2. Generative models and transportation data

Generative models are probabilistic approaches capable of making comparisons, data synthesizing, data cell imputation, signal denoising, inpainting, texture and image synthesis, semi-supervised learning, unsupervised feature learning, and many other tasks (Theis et al., 2015). The lack of utilizable statistical data for evacuation programs motivates studies to find practical solutions to generate synthetic populations that are highly representative of the people's characteristics within a risk-prone small area. Many questions have yet to be addressed before the successful employment of generative models in transportation operations.

In this regard, the most critical issue is the nature of the generative models that are more consistent with continuous variables rather than discrete data. However, discrete and categorical data is expected in characterizing the users of transportation facilities regardless of the conditions being an evacuation or daily routine commutes. For example, each person's activity in Activity-Based Models is derived from their characteristics, including but not limited to income groups, age, and household statistics. Therefore, any utilized generative models for generating a synthetic population should be universal and handle categorical, bimodal, and continuous variables and attributes.

2.3. Operational Evacuation Programming Models

During an evacuation, the traffic congestion is entirely different from regular daily traffic. Without accurate planning, the evacuation would occur with all people exiting the hazard zone simultaneously. This would result in excessive congestion in significant arterials, which increases

the catastrophic consequences of the flood. A practical solution to this problem is combining evacuation operations scheduling with the demand extracted from the synthetic population generated using population synthesizer models. This information can be used to estimate the share of public transportation in evacuation operations for each small evacuation area. This paves the road toward a mathematical model consistent with the generated demands. The resulting bimodal scheduling problem can also benefit from implementing strategies such as partial contraflow operations and collaborative lane usage with the expense of extra computational costs.

Although the mathematical models can be easily formulated for evacuation problems, the nature of evacuation operation is highly non-linear (with respect to time). The resulting solutions for current formulations are not computationally justifiable. As its final goal, this research aims to employ the statistically validated demand for personal cars and public transportation (bus service) to achieve a practical model that meets the needs for a successful evacuation. The following chapter provides a detailed literature review on the synthetic population generation and evacuation models.

Chapter 3. Literature Review

3.1. Data needs for evacuation of small areas

Bailey et al. (2007) surveyed the emergency response and evacuation plans in 20 metropolitan regions selected among areas with minorities, low-income populations, and households that do not own cars. They found that only a few agencies had included the population of minorities (e.g., low-income people who do not own cars) in their emergency plans. They suggested that disadvantaged populations' characteristics should be identified by examining all the individuals within a geographical area of interest. The necessity of sociodemographic data is highlighted in emergency evacuation literature in the sense that either the individual-level data (disaggregated information) for micro areas are not available, or only a small sample from the total population is available. Smith and McCarty (2009) stated that, unfortunately, no data sources include comprehensive and reliable information on the demographic and socioeconomic effects of hurricanes and other natural disasters.

According to a technical report published by the Disaster Technical Assistance Center of SAMHSA, not only does the individuals' sociodemographic status affects their response to the evacuation orders, but their location and other variables matter as well (SAMHSA, 2017). Masozera and Bailey (2007) and Thiede and Brown (2013) showed that the level of poverty and people's educational background directly influences individuals' access to evacuation transportation facilities. Bian and Wilmot (2017) reviewed the level of analysis for vulnerability measurements. They stated that, as a critical factor in evacuation studies, vulnerability has usually been assessed at a relatively macro level of spatial detail (e.g., see Chakraborty et al., 2005; Wang and Yarnal, 2012; Cutter et al., 2000; Cutter et al., 2003; Cutter and Finch, 2008; Emrich and Cutter, 2011; Zhou et al., 2014; Borden et al., 2007). However, the key factors influencing individuals' vulnerability to an extreme event vary significantly for two neighboring micro-level

areas. Among the available literature, only a few studies used small area variables for evacuation analysis purposes (e.g., see Cavusoglu et al., 2013). However, these studies focused on a limited number of informational factors and only from specific year data. For example, Cavusoglu et al. (2013) used the 2010 census bureau data for census tract level car-less people estimates. Therefore, it is necessary to identify the characteristics of the vulnerable and low-income people and their car-ownership status through a reproducible approach for any year and for a geographical level that is small enough to allow car-less people to access the allocated evacuation transit resources.

In addition, small area predictions allow the authorities and urban officials to better predict the pick-up points for vulnerable people. For example, Bian and Wilmot (2018) focused on analyzing transit pick-up points for the people with low mobility during Hurricane Katrina, In New Orleans. They stated that the analytical zones used for evacuation studies should be small enough to help the vulnerable people who typically have low mobility, by providing higher accessibility to evacuation resources.

3.2. Synthetic population

Due to the lack of disaggregated data, a synthetic population needs to be generated for an area of interest and used for various purposes such as microsimulation, optimization, and travel behavior models for evacuation planning purposes. For example, to create a transit-based evacuation model, Naghawi and Wolshon (2010) used a module in TRANSIMS (TRansportation ANalysis and SIMulation System), an agent-based microsimulation transportation modeling software, to generate a synthetic population based on the 2000 U.S. Census aggregated data and disaggregated data from Public Use Microdata Samples (PUMS). Later, they expanded their work to investigate a multimodal transport system's operation during a mass evacuation. They include

numerous factors such as auto ownership rates, income, and household size in developing the synthetic population (Naghawi and Wolshon, 2015).

Cuellar et al. (2009) emphasized the data gap in synthetic population generation and proposed a simple approach that generates a statistically representative population to model the population response. Lämmel et al. (2010-1) used a synthetic population obtained from census data (Germany) to simulate people's evacuation in the hypothetical dam-break and inundation of Zurich. Their synthetic population embraced 165,571 agents and included all people whose workplace was within the evacuation area. They used Multi-Agent Transportation Simulation (MATSim) framework to microscopically model evacuees as individual agents who try to optimize their routing toward a safe location. They extended this work to the case of a Tsunami in the Indonesian city of Padang and used Indonesian census data (BPS, 2005) to generate the synthetic population (Lämmel et al. (2010-2)).

Yin et al. (2014) generated a disaggregated household and individual level synthetic population using the TRANSIS package to develop an agent-based travel demand model for hurricane evacuation simulation. Their model consists of several econometric and statistical models for mode-choice, destination choice, departure time choice, and other demand model components. These studies are only a few examples among the vast literature emphasizing the need for the synthetic population in emergency evacuation problems.

Synthetic populations are used commonly as input for activity-based models where the travel patterns of individuals and their transportation choices are modeled at a granular level (Arentze and Timmermans, 2004; Salvini and Miller, 2005; Pinjari et al., 2006; Guo and Bhat, 2007; Ye et al., 2009; Bradley et al., 2010; and Javanmardi and Mohammadian, 2012). In addition, synthetic populations have been used in establishing measures of accessibility (O'Kelly and

Horner, 2003; Dong et al., 2006; and Ziemke et al., 2018). A few population-synthesizing approaches generate population while preserving the general characteristics of the population and masking individuals' identities for privacy purposes. One widely used method to create synthetic populations is the Iterative Proportional Fitting (IPF) technique, first used by Beckman et al. (1996) in transportation studies. They utilized IPF for the U.S. census data structure and implemented it into TRANSIMS (Choupani and Mamdoohi, 2016). Although Beckman et al. (1996) were the pioneers who proposed IPF in transportation-related studies, the original method dates far back to the 1940s (Deming and Stephan, 1940). IPF updates the weights of the counts from a sample until the sum of the weights for all the attributes matches the marginal values available for the area of interest. This process includes a fitting and an allocation step that consists of cross-tabulation, integer conversion, and selection. For detailed properties of the IPF, we refer the readers to Deming and Stephan (1940), Ireland and Kullback (1968), Fienberg (1970), and Mosteller (1968).

Although the IPF method is well established in transportation studies, multiple problems can obscure the validity of the synthetic population generated by IPF. One of the issues is the generation of zero cells, zero marginal, and table sparsity caused by zero or small values. Another problem is that IPF can synthesize the population at one level, either the households or individuals (not both simultaneously) (Pritchard and Miller, 2012; Pukelsheim, 2014).

An algorithm known as Iterative Proportional Updating (IPU) that generates synthetic populations where household-level and person-level characteristics of interest are matched by adjusting and reallocating weights was developed by Ye et al. (2009). IPU was further improved to accommodate control variables at multiple spatial resolutions by Konduri et al. (2016).

A few statistically driven population-synthesizing techniques have been proposed. Bayesian inference system method proposed by Schafer (1997), combinatorial optimization method proposed by Openshaw and L. Rao (1995) and Williamson et al. (1998), simulated annealing presented by Voas and Williamson (2000), and Monte Carlo Markov Chain (MCMC) proposed by Farooq et al. (2013), are among these methods. One of the latest studies in this area is the work performed by Sun et al. (2018), where they integrated three statistical methods to generalize the associations of within- and cross-level attributes and reproduction of structural relationships among household members to construct the synthetic population.

3.3. Generative models

Another approach to synthesizing population data is the utilization of generative models. In approaches that employ generative models, a generator is trained to reproduce samples from the complex dependency matrix of the data attributes. Extracting the dependency between different data set attributes does not necessarily require a generator. For example, In the specific context of synthetic population generation, Kao et al. (2012) proposed a dependence preserving approach toward synthesizing household characteristics by fitting multivariate normal distributions and modifying the covariance matrix. In general, the generative models can capture the dependency structure of the dataset and, at the same time, preserve the distribution of the variables in a joint cumulative manner.

In this research, two generative models are discussed: Copula Generative Models and Generative Adversarial Neural Networks (GANs) Models. Copulas are robust methods helpful in capturing the dependency between the different attributes of a data set with continuous variables. In the context of econometric model estimation, a copula-based approach has been proposed to test the dependency among parametric structures by Bhat and Eluru (2009). The method has seen

several empirical applications, including the case of sample selection in connection to building environment effects on travel behavior (Bhat and Eluru, 2009), and when considering multiple events, e.g., type and duration of participation, and accompaniment, in a simultaneous equations model system (Born et al., 2014). Copulas are valuable tools for examining the validity of cross-tabulation results or other data augmentation approaches. For example, Jeong et al. (2016) used copulas to validate the dependencies that have been captured by the IPF method.

This research discusses the limitations of the copula approach, and a novel research avenue is proposed to overcome these challenges by using GANs integrated with a couple of statistical components. GAN is a machine learning approach that has been widely used in image, video, and sound generation but is barely used for one-dimensional numerical data generation. GANs have been first proposed by Goodfellow et al. (2014) as a framework that the competition between a generator and a discriminator neural network results in a stable model that understands the distribution and dependency of the attributes within a dataset. They proposed the most straightforward framework for successfully generating new images from a training set in an unsupervised learning system. Since then, GANs have been on an evolutionary journey towards their current maturity, and semi-supervised (Salimans et al., 2016), fully supervised (Isola et al., 2017), and reinforcement learning (Ho and Ermon., 2016) methods have been developed to harness its fruitful capabilities better. The application of GANs in image and video processing is countless and cannot be highlighted thoroughly in a dissertation. To name a few, Radford et al. (2015) proposed a convolutional GAN that can be used in generating new image datasets. The convolutional nature of capturing the dependency between pixel data made this type of GANs very popular. Later, these generative models have been extended to create fake images of human beings (Karras et al., 2017), Generating automatic anime characters (Jin et al., 2017), unpaired image to

image translation (Zhu et al., 2017), text to photo image synthesis (Zhang et al., 2016), and animating still images (Zakharov et al., 2019).

GAN has excellent potential for transportation data analysis, where only a population sample is available. However, to date, only a very few attempts have been made to use this technique in transportation engineering applications. For example, Yazdizadeh et al. (2019) used Convolutional GANs to infer travel mode from GPS trajectory data and derive trip information from smartphone-based travel surveys. The same type of analysis has been performed by Günthermann et al. (2020). James (2020) used the idea to develop a set of four deep neural networks that work in a proxy-label-based semi-supervised learning algorithm to infer the travel modes.

Lv et al. (2018) proposed using GANs in parallel transportation systems for traffic data generation, modeling, prediction, and control. Lin et al. (2019) proposed a new traffic flow prediction model (called a pattern-sensitive network) that automatically handles different traffic patterns. They used adversarial training (GAN model) to make accurate predictions in unusual traffic states without compromising its performance in usual states.

In the study area of driver behavior, Kuefler et al. (2017) proposed a GAN model to predict and simulate (imitate) human driving behavior accurately. Their proposed approach is helpful in generating realistic driving behaviors that are robust to trajectory perturbation. Such studies are crucial in the development of intelligent transportation systems and realistic highway simulations.

In the study area of synthetic population, Xu and Veeramachaneni (2018) proposed a GAN model to synthesize tabular data that may consist of discrete and continuous variables. Badu-Marf et al. (2020) presented a Composite Travel Generative Adversarial Networks to generate synthetic agents for agent-based models. Their synthetic data includes the sociodemographic of each agent

and the trajectory data. Although their synthetic population is helpful for modeling travel demand models, the problem of estimating synthetic population for small areas where the number of samples is not enough for training has not been tackled yet. This study introduces a Conditional Wasserstein GANs approach that successfully generates a synthetic population for a geographical level suitable for evacuation modeling.

3.4. Bimodal evacuation models

Researchers have proposed numerous mathematical evacuation models to solve shelter assignment problems (e.g., Yamada, 1996; Han et al., 2006; Kimms and Seekircher, 2015), lane-based routing, and crossing-elimination strategies (e.g., Cova and Johnson, 2003; Xie et al., 2010; Kimms and Maassen, 2011; Zhao et al., 2015), contraflow strategy (e.g., Wang et al., 2013), and scheduling problems (e.g., Tüydeş and Ziliaskopoulos, 2014) deterministically or by employing the robust/stochastic approaches. Rui et al. (2009), Abdelgawad et al. (2010), Bish (2011), An et al. (2013), Goerigk et al. (2013, 2014), Kulshrestha et al. (2014), and Zheng (2014) proposed models and methodologies on mass transit evacuation planning. These models aim to solve the same shelter location, routing, and scheduling problems for the massive utilization of public transit systems. Three different methods have been utilized to implement these models: 1-Network design optimization (e.g., Kalafatas and Peeta, 2006; Tüydeş and Ziliaskopoulos, 2005), 2-Network flow optimization, and 3- Cell Transmission Model, which was originally developed by Daganzo (1994) based on hydrodynamic theory (Lighthill and Whitham, 1955; Richards, 1956). This research will use the famous network flow optimization models and propose a network design strategy to deal with possible non-linearity due to lane numbers.

The network flow problems are np-hard by nature. The high traffic volumes during the evacuation operations add to the complexity and nonlinearity of the problem, especially when

realistic assumptions are made. Sbayti and Mahmassani (2006) mathematically stated the high nonlinearity in evacuation problems while proposing an optimal scheduling evacuation solution using the simulation approaches. Adding new modes of evacuation would escalate the problem difficulty due to the interactions between different travel modes.

The literature review in the study area of network and flow evacuation planning models is vast, and it is impractical to highlight all in this limited review section. Therefore, this section will focus mainly on bimodal optimization models recently proposed to account for car-less evacuees. Abdelgawad et al. (2010) proposed a multi-objective optimization framework for multimodal evacuation operation, which combines vehicular traffic and mass transit. They first used optimal spatiotemporal evacuation (OSTE) formulation to optimize the evacuation scheduling problem where the decision variable is a staging vector for evacuation. Each element of this vector is the percentage of demand released at time interval t . This vector is then fed into a multiple-depot, time-constrained pickup and delivery vehicle routing problem (MDTCPD-VRP) for detailed scheduling results. Goerigk et al. (2014) proposed a comprehensive evacuation model. In this model, each bus within the evacuation operation is considered a standalone commodity, and private vehicles are modeled as a flow of cars. They assumed that each bus could stop at any node to pick up evacuees until the bus's available capacity was used. Although they named it a comprehensive evacuation model, this model is not sufficient for flood occasions where the mass that evacuates the area cannot be picked up from the public transit services' preexisting regular schedules. In this study, Goerigk et al. (2014) used the same approach Bretschneider (2012) proposed and added a risk parameter to each link within the network. In a simultaneous study, Goerigk and Grün (2014) and Goerigk and Deghdak (2015) suggested a robust stochastic method to account for the uncertainty in demand, travel times, and the number of buses during the

evacuation operation. Amideo and Scaparra (2017) proposed an integrated model that combines a system-optimal and a user-optimal approach. They assumed that evacuation planners have complete control over the bus routes, whereas the car-based evacuees aim to reach the shelters following the shortest available path. Yang et al. (2018) Considered evacuees who choose between car and bus station and can use the regular bus transit scheduled for peak hours to arrive at a shelter. This is not realistic as the number of people close to the shelters and who have access to personal cars is unknown.

In summary, there is no reliable method to accurately estimate the number of people who do not have access to personal vehicles. The main reason is that the structure of the existing data samples in the US is not suitable for car-ownership estimations for the geographical levels that require emergency evacuation during natural hazards. This is a research gap, and this research proposes an integrated statistical and machine learning approach to produce micro-level population data from high-level samples. Another gap in the existing literature is that the evacuation models are either non-realistic or employ resources that are not suitable for flood cases. The most helpful strategy in the case of a predicted flood strike is to use a bimodal transportation system that employs a bus transit system and personal vehicles to evacuate the people who do not have access to private cars. This study develops a MILP TET model and a heuristic solution that optimally uses these systems in the case of flood striking in areas close to shorelines.

Chapter 4. Generative Copula Model and its Limitations

4.1. Copula Model

4.1.1. Introduction to the generative copula models

The statistical model, namely *Generative Copula Model*, works based on preserving the attributes' marginal distributions' dependency by fitting the sample data's rank to specific types of non-linear functions (called copula functions). The data's rank is simply the sorting index of the data within the dataset (e.g., the variable's rank with the smallest value within a data column is 1, and that of the next smallest value is two, and so on). This statistical model's performance in generating a synthetic population for evacuation purposes is evaluated in this research. The critical factor is whether the synthesized population is good enough to represent the population characteristics, including people's income within small geography. With this accurate data in hand, the necessary information, such as car ownership, can be modeled for evacuation planning. Specifically, data from large-scale samples (e.g., ACS and IRS data) for Anne Arundel County, Maryland, has been used to generate synthetic populations for smaller subareas (census tract). The generated synthetic population is then used to estimate low-income residents' car-ownership status in the subareas, based on transportation-related data extracted from the National Household Travel Survey (NHTS). The Copula Generative Model application issues are discussed, and the alternative approach, which is a machine learning generative model, is proposed to overcome these issues.

4.1.2. Generating synthetic population using Copula (methodology)

The mathematician Abe Sklar first introduced the copula theory and its unique feature into probability theory in 1959 (Sklar, 1959). Based on his definition, copulas are any kind of mathematical functions that perform a mapping between marginal distributions of a d -dimensional multivariates and their joint cumulative distribution:

$$C: \underbrace{[0,1]^d}_{\substack{\text{marginal distributions} \\ \text{of } d \text{ dimensional} \\ \text{multivariate data}}} \rightarrow [0,1]^1 \quad (1)$$

Where C is a copula function. In other words, the inputs of any copula function are the marginal distributions of the multivariate data, and the output is a unique value between zero and one, which is the joint cumulative distribution of the random variables. Assuming that H is the joint cumulative distribution function (CDF) of any d -dimensional continuous random variable, Sklar Theorem states that:

$$H(x_1, x_2, \dots, x_d) = C\{F_1(x_1), F_2(x_2), \dots, F_d(x_d)\} \quad x_1, x_2, \dots, x_d \in \mathbb{R} \quad (2)$$

Where F_d is the marginal distribution of the d^{th} variable. These marginals can be defined as the components of the pseudo copula samples and can be expressed as a function of the rank (R_d) of the observation in multivariate data. Assume that the multivariate data consists of n number of observations. For the i^{th} observation within d^{th} dimension of a multivariate data, F_d^i and R_d^i are calculated as follows:

$$F_d^i = \frac{R_d^i}{n} \quad (3)$$

$$R_d^i = \sum_{j=1}^n 1(x_d^j < x_d^i) \quad (4)$$

Based on equations 1 to 4, an empirical copula can be constructed as the empirical distribution of the rank-transformed data. This empirical distribution can perform the mapping mentioned in equation 1. For any realization of the rank-transformed multivariate data, the empirical copula function can be written as follows:

$$C_n(f_1(x_1), \dots, f_d(x_d)) = \frac{1}{n} \sum_{j=1}^n 1(F_1^j(x_1) < f_1, \dots, F_d^j(x_d) < f_d) \quad (5)$$

Sklar (1959) showed that, for any multivariate set of continuous variables, there exists at least one functional (parametric) form of the copula that can perform the mapping mentioned in Eq. (1). Several families of copulas have been described in the literature; each family consists of multiple copula functions. Copulas that belong to one family usually share some common property. For example, Archimedean copulas admit an explicit formula and allow modeling dependency of multivariate random variables with only one parameter (θ), which is also the reason for the popularity of this family of copulas. Table 1 shows the formulation for the three Archimedean copulas used in this study.

Table 1. Three most popular Archimedean copulas.

Copula	Multivariate Copula C_θ	Range of the θ
Clayton	$[\max\{f_1^{-\theta} + f_2^{-\theta} + \dots f_d^{-\theta} - d + 1; 0\}]^{-1/\theta}$	$\theta \in [-1, \infty] \setminus \{0\}$
Frank	$-\frac{1}{\theta} \log \left[1 + \frac{\prod_{i=1}^d (\exp(-\theta f_i) - 1)}{\exp(-\theta) - 1} \right]$	$\theta \in \mathbb{R} \setminus \{0\}$
Gumble	$\exp \left[- \left(\sum_{i=1}^d (-\log f_i)^\theta \right)^{1/\theta} \right]$	$\theta \in [1, \infty)$

There are deterministic approaches for estimating the copula parameters for bivariate random samples that rely on dependency measurements. For multivariate samples, maximum log-likelihood estimation is a popular method. For big sample data, maximum log-likelihood estimation usually converges to a solution, although this method is computationally exhausting. To reduce the longitudinal dimension of the problem (i.e., the number of observations), one can write the log-likelihood in its weighted form. In this form, the repeated observations within the

data appear only once in the log-likelihood with their associated weights. Let's denote the vector of pseudo-observations (rank-transformed observations) for a d dimensional data by $\mathbf{U}$:

$$\mathbf{U} = (F_1(x_1), \dots, F_d(x_d)) \quad (6)$$

Assuming that n^* is the modified sample number after deleting the identical observations, and ω_i is the weight for each observation, one can write the log-likelihood function as follows:

$$\mathcal{L}(\theta) = \frac{1}{n^*} \sum_{i=1}^{n^*} \omega_i \log[C_\theta(\hat{U}_i)] \quad (7)$$

The concept of rank is utilized to introduce the marginal distribution of the data to the copula. The variables' rank preserves the dependency among the variables for a multivariate problem. This is because the rank of the data remains unchanged under any strictly decreasing or increasing transformation function. The rank of the variables divided by the number of observations augmented by one result in the normalized rank for each realization within the sample. Note that the copula maps the normalized ranks of the variables to the joint distribution of the variables.

Once the data is fitted to a copula, it is necessary to evaluate how well the chosen copula family describes the dependency structure of the multivariate random variable being modeled. The statistical test can be initiated with a null hypothesis that the parametric copula (C_θ) that preserves the dependency structure of a given data belongs to the chosen family of copulas. The statistical test to evaluate the null hypothesis can be constructed using the empirical copula given in Eq. (5), as proposed and examined by Genest et al. (2008, 2009). Genest et al. (2009) proposed testing the null hypothesis based on distance measures that are functions of the difference between empirical copula and empirical distribution function defined in Eq. (5). The most common form of this

empirical distance had been considered by Fermanian (2005) in the form of $\mathbb{C}_n(u) = \sqrt{n}(C_n - C_{\theta_n})$ who commented that the tests based on the empirical process are impractical, except for bootstrapping techniques. Genest et al. (2009) proposed rank-based versions of the familiar Cramér-Von-Mises statistics to deduce approximate P-values from its limiting distribution.

The continuous form of the Cramér-Von-Mises statistics that measures the statistical distance between distributions of the empirical copula and the parametric copula can be expressed as follows:

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\mathbf{U})^2 dC_n(\mathbf{U}) = \sum_{i=1}^n \left(C_n(\mathbf{U}_i) - C_{\theta_n}(\mathbf{U}_i) \right)^2 \quad (8)$$

The Null hypothesis is rejected with large values of the S_n as n grows to infinity (Genest et al., 2009). Although the limiting distribution of the S_n is unknown as $n \rightarrow \infty$, it can be approximated by using parametric bootstrap. Figure 4 shows the bootstrap flowchart to calculate the P-values using the Cramér-Von-Mises statistics.

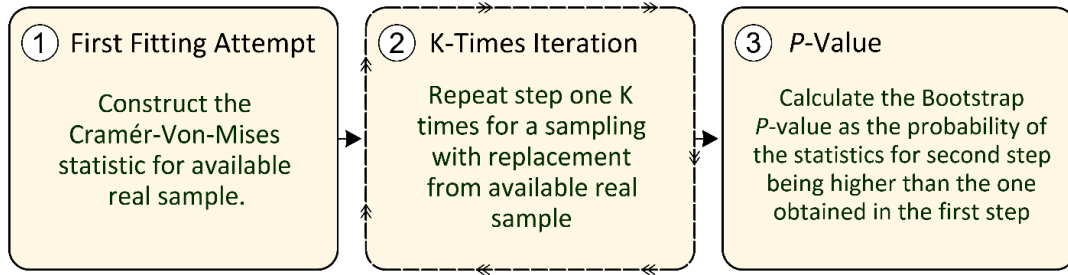


Figure 4. Bootstrap flowchart to calculate the P-value for a selected copula.

The initial step of the bootstrap method is to fit the available data to a targeted copula. The Cramér-Von-Mises statistic is then constructed and calculated as the sum of squares for differences between empirical and fitted copulas. This distance (S_n) is calculated for the available sample and the estimated copula first. This procedure is repeated for k times (a high number of iterations needs

to be selected), each time with the data obtained by sampling with replacement from the initial dataset, which leads to a statistics (S_n^*) for each iteration. The following equation can be used to calculate the bootstrap P-value:

$$P = \frac{1}{K+1} \left[\sum_{i=1}^K 1(S_n^* \geq S_n) + 0.5 \right] \quad (9)$$

Note that this P-value is an approximate criterion. A P-value higher than a desired statistical level of significance indicates that there is insufficient statistical evidence to reject the null hypothesis. The fitted parametric copula belongs to the copula family that has been chosen. This intuitively means that the copula successfully captures the true dependency structure of the multivariate data.

Once a copula passes the statistical test for the goodness of fit, there are statistical methods to generate synthetic data from the fitted copula. It should be noted that estimated copulas produce pseudo-observations, which are realizations from the cumulative distribution function of the population. The last piece of the problem that completes the procedure is utilizing the inverse cumulative distribution function of each variable. Each pseudo-observation drawn from the copula is transformed into a synthetic observation using that variable's inverse cumulative distribution function. Within the mentioned procedure, the copula preserves the dependency of the variables, while the CDF of the variables evokes the distribution and realization of the observations independent of the copula. Figure 5 shows the flowchart of the entire procedure for the generation of the synthetic population. The process starts with a sample of observations as input. Then, the rank of variables in the data set is determined, followed by the pseudo-observations computation. The pseudo-observations are then fitted to a copula, and the goodness of fit test is performed using

a bootstrapping algorithm. After the test, n observations are drawn from the pseudo-observations to represent the population, and they are transformed into a synthetic population using CDFs.

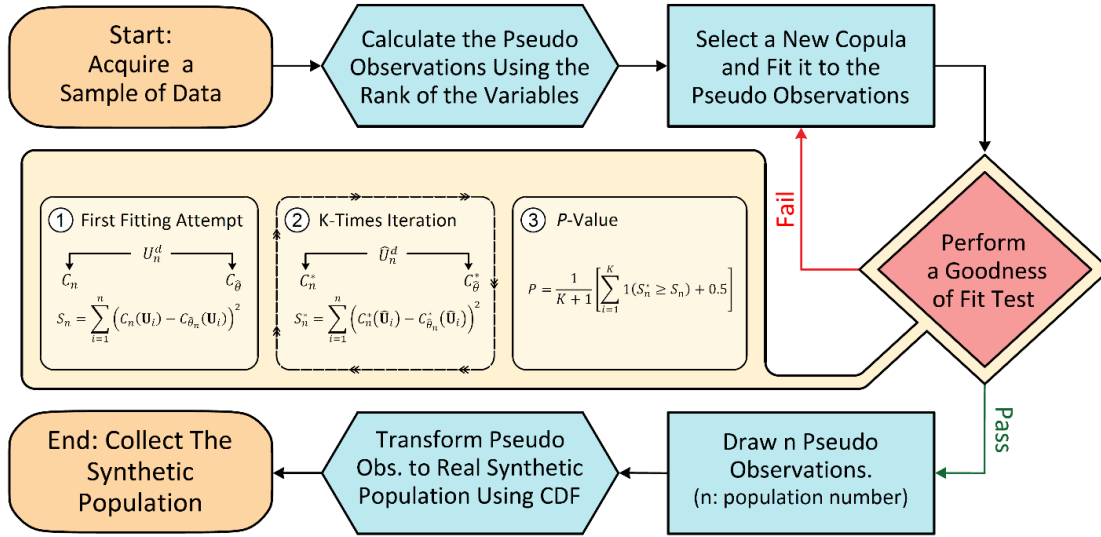


Figure 5. Flowchart of the population-synthesizing algorithm.

4.1.3. Anne Arundel County Application

Anne Arundel County in Maryland is located south of the City of Baltimore and west of Chesapeake Bay. The county's east border is surrounded by water, including the Chesapeake Bay and its numerous tributaries, the various rivers, creeks, streams, and inlets, covering 29% of its area. The county's geo-spatial location, combined with changing climatic conditions, has made it vulnerable to natural disasters like hurricanes, storm surges, and flooding. The county has experienced significant damages from, e.g., tropical storm Isabel (September 2003), hurricanes Ernesto (September 2006), Irene (August 2011), and Sandy (October 2012). The tragic examples of hurricanes Katrina (2005) and Maria (2017) that cost thousands of lives have taught us that the evacuation of typically low-income populations who do not own cars is crucial. Therefore, the need for a tool to help planners, first responders, and others ensure these vulnerable populations' safety is essential. The population synthesis method developed in this study is applied to Anne Arundel County as a foundation for such a tool.

Figure 6 shows the median household income for census tracts within Anne Arundel County in Maryland. Anne Arundel County includes four PUMAs that are almost equal in terms of area. Within each PUMA, there are a few census tracts that are occupied by people with low median household incomes. The number of observations for some of the census tracts is too low. Therefore, the implementation or validation of any model would be difficult or impossible using these few observations. Copula solution is consequently proposed to generate a synthetic population provided that a sample of attributes of interest and their marginal distributions for the study region are available.

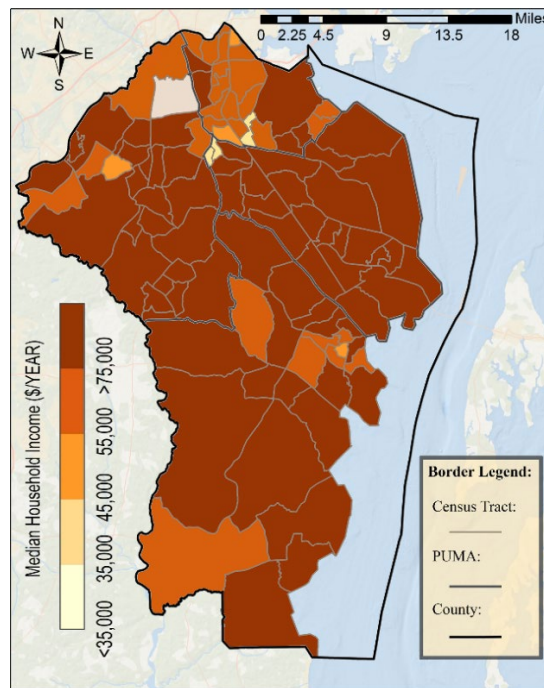


Figure 6. The median income for different census tracts within Anne Arundel county data (note that the data was obtained from Interactive maps available at the United States Census Bureau website in 2018, currently available in a different format at <https://www.census.gov/programs-surveys/geography/data/interactive-maps.html>).

First, the study area's spatial data was acquired, consisting of three layers, each finer layer nested within the larger layer. The geographic structure and the resolution of the data are depicted in Figure 7.

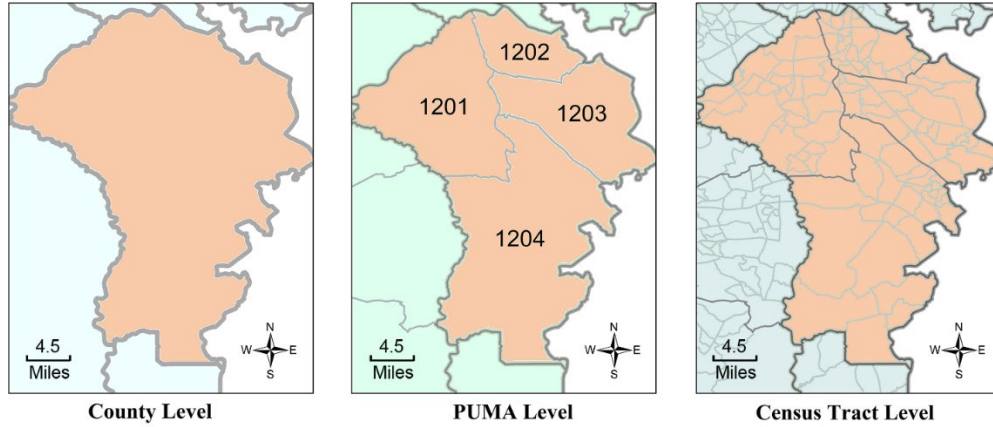


Figure 7. Illustration of PUMAs and census tracts in Anne Arundel County.

Anne Arundel county is divided into four PUMAs, and each PUMA consists of several census tracts. Census tracts, PUMAs, and the county are perfectly nested within each other, ensuring that no data is mistakenly duplicated. In this study, the data from the PUMA and census tract levels combined with IRS data is used to generate the synthetic population. To the best of our knowledge, they have not been consolidated before for population synthesis purposes. The three sources of data linked in this study are (1) American Community Survey (ACS), (2) Decennial Census Data, and (3) IRS Income Data from tax returns.

American Community Survey provides a sample of information for Public Use Micro Areas (PUMA), called Public Use Micro Samples (PUMS). PUMS consists of two different levels of the data: household level and individual level. These two levels of data share a common attribute (i.e., serial number) that is suitable for data integration and tabulation purposes. The serial number indicates the attachment of each individual to a specific family. Each of the two data sets benefits from a weight column that determines the observation's weight within the sample. The weights can be used to replicate the data before the copula fitting process. ACS suggests that the household-level data can be integrated with individual-level data for individual-level studies. Then the combined data can use the individual level weights to replicate the observations. The attributes used in this study from the household level (HH) and individual level (IL) are listed below.

1. NP: Number of person records following the housing record (HH),
2. HHT: Household or family type (HH),
3. HINCP: Household income (past 12 months) (HH),
4. HUPAC: Household presence and age of children (HH),
5. WIF: Workers in the family during the past 12 months (HH),
6. AGEP: Age of the person (IL),
7. SEX: Sex of the person (IL),
8. ESR: Employment status of the person record (IL),
9. RAC1P: Recorded detailed race code (IL).

In this study, the individual-level data is constructed by the data tabulation method discussed above. The rank-based pseudo-observations for these attributes can be fitted to a copula as a multivariate problem. The rank of the data for cardinal variables has been calculated by replacing each category with a numerical value. The copula is capable of capturing the dependency between these nine variables. The synthesized pseudo-observations can then be generated from the estimated copula. The generated data holds the structural dependencies of the initial sample data.

As mentioned before, preserving the dependency between the variables is not sufficient for synthetic population generation. Another critical piece of the puzzle is each variable's Cumulative Distribution Function (CDF). Inverse CDF of the variables transforms the pseudo-observations generated by the copula into actual synthetic observations. In this study, two data sources have been used to construct CDF functions for the variables of interest. The first data is Decennial Census Data 2010, including tables required to generate CDF functions for six out of nine attributes mentioned above. These attributes are NP, HHT, HUPAC, AGEP, SEX, and RAC1P. Unfortunately, CDF functions for WIF and ESR are not currently available, and therefore, the CDF

for these two variables has been obtained from the PUMA samples. IRS data has been used to generate the CDF function for the income variable, which is the critical attribute for identifying the low-income population. Figure 8 shows an example of IRS data that has been used to generate CDF functions for Income.

ZIP	INCOME	CLASS	N_return	N_single	N_joint	N_head	N_dpnd	N_TOT
20601	Total	0	13160	6120	3950	2520	8390	24930
20601	\$1 under \$25,000	1	3720	2790	230	620	1390	5260
20601	\$25,000 under \$50,000	2	2820	1450	450	750	1940	5040
20601	\$50,000 under \$75,000	3	2100	890	510	560	1560	4030
20601	\$75,000 under \$100,000	4	1640	540	660	340	1120	3320
20601	\$100,000 under \$200,000	5	2440	450	1690	250	1940	6020
20601	\$200,000 or more	6	440	0	410	0	440	1260
20602	Total	0	13280	6390	3180	3190	9210	25150
20602	\$1 under \$25,000	1	4100	2880	240	890	1930	6180
20602	\$25,000 under \$50,000	2	3080	1560	390	980	2300	5620
20602	\$50,000 under \$75,000	3	2280	1000	490	650	1610	4240
20602	\$75,000 under \$100,000	4	1620	590	530	410	1270	3330
20602	\$100,000 under \$200,000	5	1990	360	1340	260	1880	5180
20602	\$200,000 or more	6	210	0	190	0	220	600
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 8. A sample view of IRS data used to generate CDF functions for Income.

IRS reports the number of returns for different types of income return forms. The report includes the number of returns, the number of single returns, the number of joint returns, the number of head of family returns, and the number of dependents recorded. The number of joint returns records was doubled and summed with single, head of a family, and dependents records to obtain the CDF function for income. The last column in Fig. 6 represents the calculated total returns for individual-level data. It should be noted that since the individual-level data has been fitted to the copulas, the CDF function of the variables must be at the individual level. The reported value for the variables integrated from the household is a characteristic of people's families. This means that these variables' CDF function should be obtained from individual-level distributions while reflecting people's family characteristics (e.g., income).

Another issue associated with IRS data is geographical resolution: the data is reported for zip codes; therefore, it cannot be used to generate the CDFs for census tract levels directly. Fortunately, the Office of Policy Development and Research provides a crosswalk between zip code level and census tract-level data that indicates the percentage of each zip code that belongs to a census tract. The data consists of the ratio for residential, business, and other addresses. The total ratio is reported. In the IRS data case, since individuals tend to file their tax return forms based on their residential addresses, the residential ratio was used as a base to generate income CDFs for different census tracts. Figure 9 shows an example of the crosswalk for two census tracts. The first three columns in Figure 9 represent the zip code, the tract number that partially includes the zip code, and the state of the zip code, respectively. The remaining columns represent the residential, business, other, and finally, the total ratio of the zip code level addresses that belong to the desired census tract.

ZIP	TRACT	STATE	RES_RATIO	BUS_RATIO	OTH_RATIO	TOT_RATIO
21054	24003702204	24	0.013561211	0.03500761	0.029126214	0.016115904
21114	24003702204	24	0.146610815	0.043153527	0.013071895	0.132833958
21114	24003702205	24	0.292650419	0.546058091	0.575163399	0.325260092
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 9. A sample view of crosswalks between ZIP code and Census Tracts.

4.1.4. Identification of Low-income People who do not own cars

After calculating the rank-transformed pseudo-observations for the nine input variables available in ACS, these pseudo-observations have been fitted to four Archimedean copulas (i.e., Clayton, Gumbel, Frank, and Joe Copulas). Clayton copula showed the best fit; however, except for the Frank copula, p-values for all other copulas considered were significant within the 99% confidence interval. Therefore, the Clayton copula was used for population synthesis. The Clayton copula results for the four PUMAs in Anne Arundel county to generate the synthetic population

are shown in Table 2. There are identical observations within the sample, usually referred to as "ties."

The ties within the sample need to be accounted for in calculating the rank and pseudo-observations. The ties' rank can be set to the minimum, maximum, or the average of the index assigned to the identical observations. In this study, the ranks of the variables are obtained by replacing the index set of the ties with their mean value (called the "average" method), and the pseudo-observations have been calculated accordingly. This method keeps the weighted distribution of the pseudo observations close to the distribution of the real observations. In other words, training a copula using this method will correlate the joint distribution of the data set to half of the frequency of each category for each attribute. This method ensures that the generated pseudo observations will be distributed around the average of the frequencies and well represent the dependencies between all the variables regardless of being of cardinal or ordinal nature.

Table 2. Estimated parameters for Clayton copula for census tracts within Anne Arundel county

PUMA	Copula	Statistics	P-value	Estimated Parameter (θ)
1201	Clayton	156.74	0.9995	0.30082
1202	Clayton	139.65	0.9995	0.33478
1203	Clayton	173.25	0.9995	0.28861
1204	Clayton	158.7	0.9995	0.29214

Having obtained the estimation parameters, a set of pseudo-observations is generated for each census tract. The size of each set is equal to the total population of the corresponding census tract. The CDF functions, produced using 2010 Decennial Census Data and IRS Income Return Data, are utilized to obtain the synthetic population. The synthetic population distribution is compared with that of the real data from the Decennial Census and IRS. The error is statistically zero within the 99% confidence interval.

Once the synthetic population is ready, different models can benefit from the available anonymous data for the entire population of a census tract. Model-based estimates are widely applied in small area estimation when only small sample sizes are available for the geographical area of interest and when design-based estimators do not result in sufficiently reliable estimates (Rao and Molina, 2015). In this research, a simple binomial logit model that predicts if individuals belong to a household with no car or those with access to one or more cars is estimated using Biogeme (Bierlaire, M., 2018) on data extracted from the 2017 National Household Travel Survey (NHTS) for the State of Maryland. A series of models have been tested to find the best predictors. The final formulation includes the following attributes: number of children (CH1) and number of workers (WIF) in the family, income (on three levels of variation, below poverty (inc1), low-middle income (inc2), and middle income (inc3)), and race (African Americans (R1) and Asian (R2)). The utility of the base alternative (no-car) has been set to zero. Equation 10 expresses the final utility function defined for the choice of having access to a personal car.

$$U = \beta_0 + \beta_{CH1} + \beta_{inc2} + \beta_{inc3} + \beta_{inc4} + \beta_{R1} + \beta_{R2} + \beta_{WIF} \quad (10)$$

Table 3 shows the estimated binomial logit and associated statistical test results for a 98% confidence interval.

Table 3. Results of Binomial Model used to estimate car ownership.

Name	Value	Robust std. err	Robust t-test	P-value
Alternative Specific Constant - 0 car	0.00	Fixed		
Alternative Specific Constant – at least one car	0.583	0.180	3.24	0.00
Number of children in the household (CH1)	1.01	0.802	1.26	0.21
Personal income < \$25k (inc2)	1.55	0.369	4.20	0.00
Household income \$25K - \$50k (inc3)	3.65	0.468	7.81	0.00
Household income \$50K - \$75k (inc4)	3.93	0.588	6.68	0.00
Race - African American (R1)	-1.35	0.562	-2.40	0.02
Race – Asian (R2)	-2.06	0.795	-2.59	0.01
Number of Workers (WIF)	0.745	0.262	2.85	0.00

Coefficient estimates are consistent with results available in the literature (Liu et al., 2014). The presence of children in the household increases the likelihood of owning cars, although this effect is not very significant. As income increases, the utility of having a car also grows. Asian and African Americans are less likely to own a car, while the number of workers in the household positively impacts the utility of having at least one car.

Having estimated the car ownership model's coefficients, the model estimates are applied to the synthetic population. The probability of having zero or one or more cars for each individual within the synthetic population is derived using Equation 11.

$$P(U_{1+car(s)} | x) = \frac{e^{U_{1+car(s)}}}{1 + e^{U_{1+car(s)}}} \quad (11)$$

This probability has been aggregated over the entire synthetic population for each census tract in Anne Arundel county to find the percentage of the people who do not have access to a personal car. Figure 10 illustrates the percentage of low-income individuals, defined as those who belong to a household with a total income of less than \$25,000. The darker the color, the higher the percentage of the low-income population. Figure 10 also shows the percentage of people in each census tract who do not have access to a household vehicle; the magnitude of the dark green spots indicates the share of the population who might need help to evacuate to a safe place during evacuation.

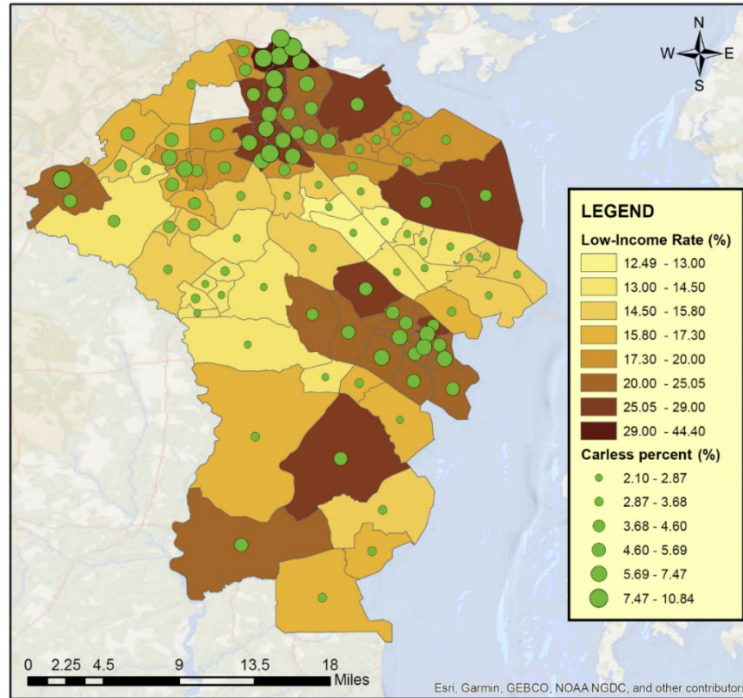


Figure 10. Car ownership and income level for Census Tracts within Anne Arundel County.

It is clear from Figure 10 that as the percentage of low-income people increases within a census tract, the percentage of people with no access to a car also increases. It is observed that most people with no car are located within the northern part of the county, which is adjacent to the south of Baltimore City. The percentage of people with no car varies between 2 and 11 percent, while the low-income percentage varies between 12 and 44. The percentage of low-income people obtained appears to be higher than the percentages found in official statistics (see Figure 6). The reason for this discrepancy is in our simplified definition of low-income households, based on a threshold of \$25,000/year. Different poverty measures are defined based on family size and their composition and used for mapping poverty. These measures can easily be incorporated into our method if more detailed income and family composition data become available.

4.1.5. Limitations of the copula model

As shown in the previous section, Copula can estimate a population's characteristics by preserving the dependency between the different attributes of the data as a joint cumulative

distribution of marginal distributions of the attributes. Although the results can be used to estimate the real population, the accuracy of the results can be questioned for two distinct reasons.

The first reason is inherent in the copula's nature. The copula theory is proven to be valid for continuous variables, and for tabular/cardinal discrete data, it is impossible to measure the accuracy level for the results. Genest and Nešlehová (2007) highlighted statistical evidence that copulas are not suitable for discrete data, especially when the nature of the data includes ties that result from atoms in the probability distribution. Later, through a series of simulation tests, Trivedi and Zimmer (2017) provided more evidence that proves Genest and Nešlehová's (2007) conclusions. They also offered cases where the copula theory might result in better results for discrete count data.

The second reason is due to the nature of the problem that we have in estimating smaller areas from samples of larger areas. Our estimations' governing assumption is that the marginal distributions of the attributes are the same for different sub-levels of the PUMAs. This means that the population in a PUMA is evenly distributed within all Census Tracts. If this assumption does not hold for drawing the synthetic population from a trained Copula function, then the results' validity is compromised. Figure 11 schematically illustrates this issue. Area A is an area divided into two smaller areas. Assume that a parametric copula function has been fitted to the samples of area A. In this imaginary area, people between the ages of 20 and 40 own a personal car. A logical realization (R_c) of the fitted copula can be 0.3 for Age and 0.15 for an individual's car ownership status. Assume that subarea A_1 does not accommodate any person between the ages of 0 and 20. Using the pair R_c (0.3, 0.15) and CDFs for subarea A_1 to draw a sample for this small area results in an invalid sample of a person who is 44 years old and own a car.

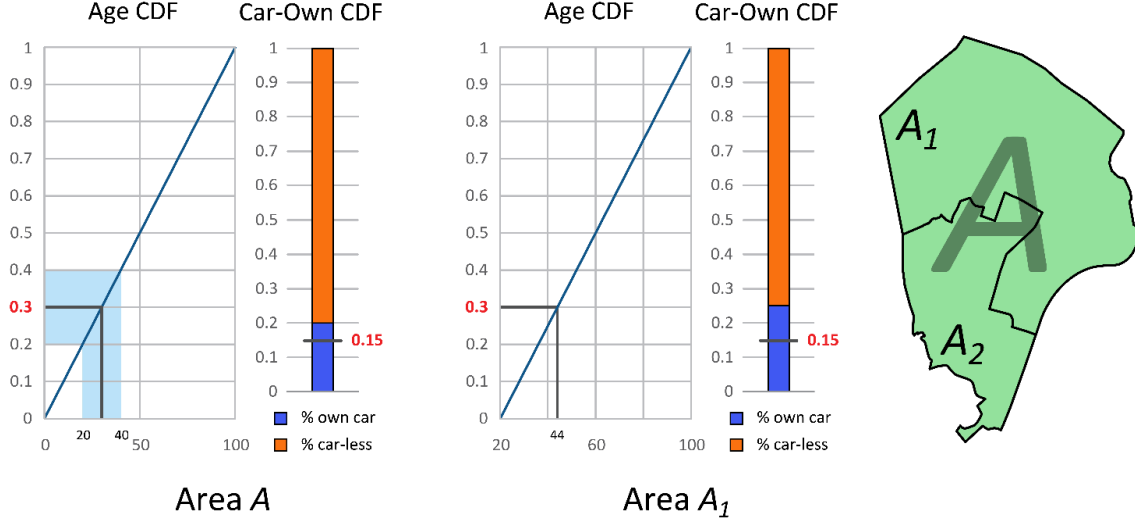


Figure 11. Illustrative representation of the second issue with small area estimations using *Copula*

In the mathematical form, the issue can be stated as the probability of the sample belonging to subarea A_1 . At the same time, it has been drawn from a copula fitted to the samples from area A . Since the fitted copula preserves the samples' joint distribution, the condition in this conditional probability is equivalent to stating that the sample belongs to the distribution of A .

$$P(R_c \in A_1 | R_c = C_\theta(\mathbf{U})) = P(R_c \in A_1 | R_c \in A) \quad (12)$$

The realization for area A_1 from the copula and CDFs of A_1 is relevant only if the following equation holds:

$$P(R_c \in A_1) = P(R_c \in A_1 | R_c \in A) = \frac{P(R_c \in A_1 \cap R_c \in A)}{P(R_c \in A)} \quad (13)$$

Equation 13 holds only if the distributions and the dependency of variables are the same for area A and subarea A_1 .

The next chapter develops a GAN-based approach integrated with statistical methods to tackle these two problems.

Chapter 5. GANs and Synthetic Population Generation

5.1. Introduction to GAN model for sociodemographic data generation

Generative Adversarial Neural Network (GAN, in short) is an advanced machine learning approach that has been widely used in image, video, and sound (signal) generation but is barely used for structured-data generation. This chapter first discusses the reasoning behind the higher reliability of GANs model over the Copula method. Then several GANs-based models are developed for data synthesizing.

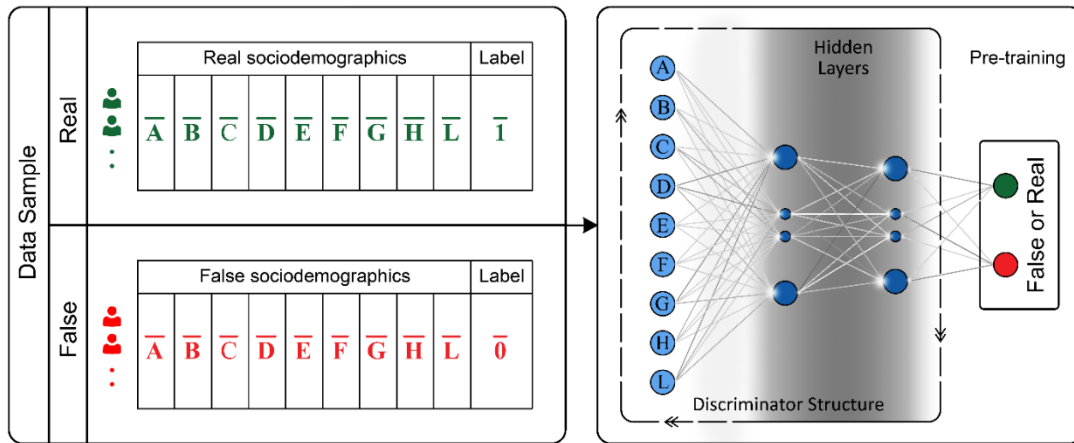
Once GAN generates an accurate synthetic population, the mapping of the results for the car-ownership model (i.e., the percentage of the people who do not own a car) can then be used in a variety of transportation and optimization models for planning, operations, design, and logistics of evacuation strategies.

5.1.1. Basic GANs model for synthetic population generation

A basic GANs model consists of two competing neural networks, a discriminator and a generator. The discriminator is trying to learn if a given sample is fake or real. Simultaneously, the generator tries to fool the discriminator with artificial samples, which the discriminator cannot recognize their fakeness. The term adversarial in GANs originates from this battle between the discriminator and the generator. A successful GANs model balances the tradeoff between the discriminator's ability to recognize the fake samples and the generator's ability to generate synthetic samples. This makes sure that the generator successfully generates synthetic samples that belong to the real data distribution.

One can use different layouts and structures to construct this adversarial battlefield between a discriminator and a generator. Before discussing the GANs model as a standalone training unit, the different components of the GANs will be highlighted. The discriminator and the generator are

the indelible components of any GANs structure. The discriminator is a neural network that is trained using two datasets of fake and real data. These fake and real datasets (in our case, the sociodemographic of the people within a small area) are labeled by a Boolean index indicating whether the observation is fake or real. A trained discriminator can recognize if a new observation comes from the same distribution of the training data set. A discriminator can be pre-trained for a very few initial iterations. Depending on the nature of the problem and the GANs' structure, pre-training can be beneficial or detrimental for the successful convergence of the GANs model as a single unit and should be performed only for a handful of iterations. Figure 12 schematically shows the structure of the discriminator in the proposed model in its pretraining step.



The generator is the second inseparable component of a GANs model. The primary task of the generator is to take as an input a vector of a predefined number of noises drawn from a latent space and generate a new fake observation as an output. A noise vector can be drawn from different symmetric or nonsymmetric distributions with different parameters. The trial-and-error process defines the size of the noise vector and the shape of the distributions. Figure 13 schematically shows a generator's structure with a noise vector with three components, each element being drawn from a gaussian distribution with a different mean and standard deviation.

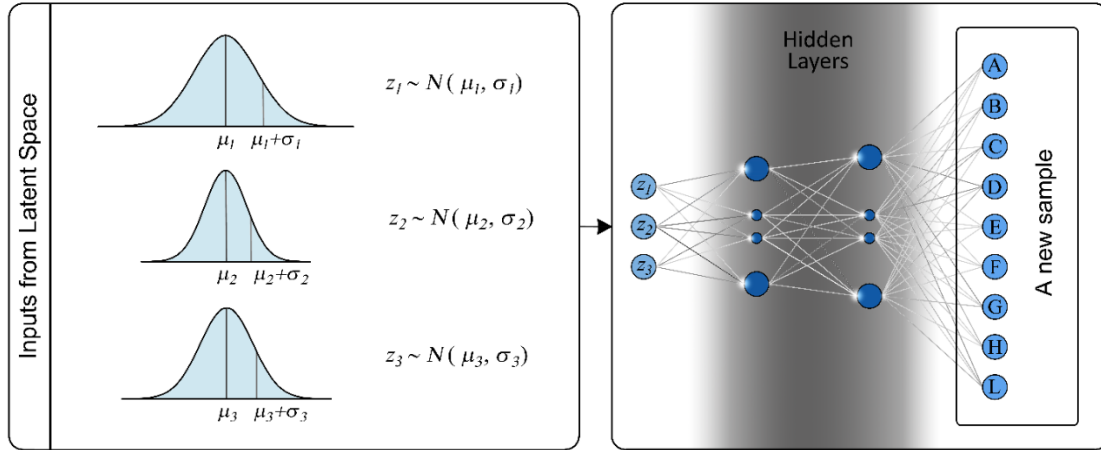


Figure 13. The structure of an example generator

In any neural network model, the goal is to tune each neuron's weights for every layer to minimize the difference between the model's predictions and the samples' real labels. Regardless of how it can be formulated, this difference is usually called loss function. The fundamental difference between the generative adversarial neural networks and a standalone neural network is that in a GANs model, a discriminator and a generator are integrated into a composite model where the neurons' weights for the generator are trained based on the performance (loss function) of the discriminator. Figure 14 schematically shows the structure of a compiled GANs model.

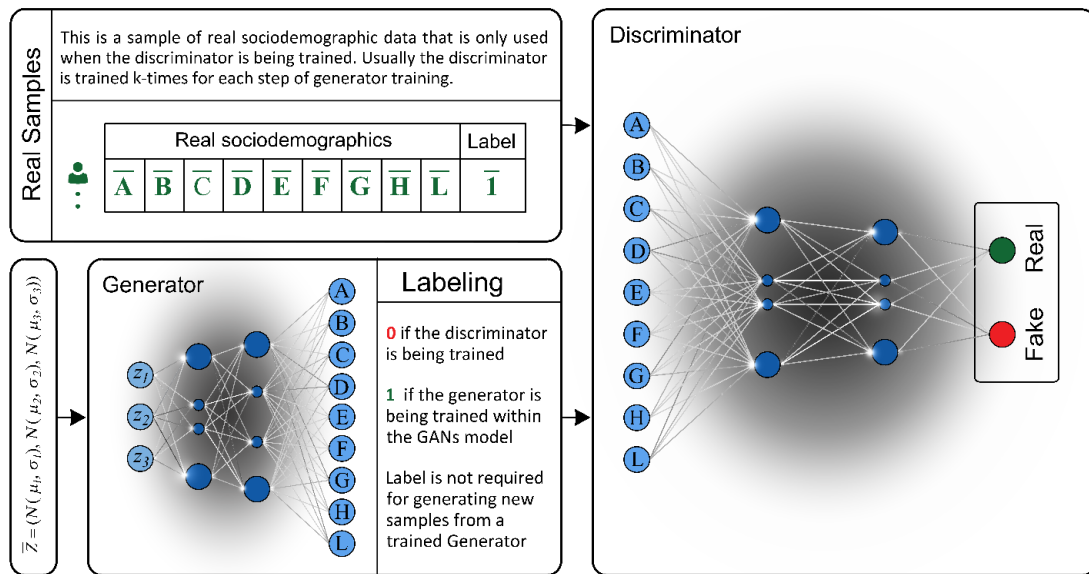


Figure 14. The structure of a compiled GANs model

Within the process of GANs training, the discriminator tries to minimize its error in identifying the fakeness of the samples received as input. At the same time, the generator attempts to maximize the discriminator's error in determining the fakeness of the samples generated by the generator. This action is called a two-player minimax game in which one player tries to maximize its probability of winning (performance), and its competitor tries to minimize this probability. The mathematical expression for this minimax performance function is called Value Function and has been formulated by Goodfellow et al. (2014) as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}} [\ln D(x)] + \mathbb{E}_{x \sim p_{data}} [\ln (1 - D(G(z)))] \quad (14)$$

Equation 14 expands the binary cross-entropy loss function where the natural log has been used instead of the least-square loss function to penalize the model further when the differences between the predicted and actual outcomes are high. The expression for binary cross-entropy function for predictions over n number of observations is as follows:

$$J = - \sum_{i=1}^n [y_i \ln \hat{y}_i + (1 - y_i) \ln(1 - \hat{y}_i)] \quad (15)$$

For any binary classification using the sigmoid function, the predicted outcome ($\hat{y}$) is a value between 0 and 1 (cannot be exactly 0 based on the definition of the sigmoid function). The actual class of the variable is defined by either 0 or 1. The binary cross-entropy loss function ensures that three important criteria have been satisfied for any binary classification. First, when the value of y and $\hat{y}$ is equal, the value of the loss function (J) should be zero. Second, J should be preferably very high for misclassification instances. Finally, J should have a non-negative value for any feasible combination of y and $\hat{y}$.

$D(x)$ and $D(G(z))$ in equation 14 are realizations from the actual data and the fake data drawn from the generator. For any sample i , when the data label is 1 (a sample from the dataset, or when the generator wants to fool the discriminator), equation 15 changes to:

$$y_i = 1, \quad \hat{y}_i = D(x) \rightarrow J = [\ln D(x)] \quad (16)$$

Similarly, if the data label is 0 (a sample is drawn from the generator, then equation 15 will be expanded to result in the following expression:

$$y_i = 0, \quad \hat{y}_i = D(x) \rightarrow J = \ln(1 - D(G(z))) \quad (17)$$

Summing equations 16 and 17 and taking the expectation over the distribution of the actual data sample and that of the generator will result in equation 14. This is the objective function for the GANs model that is solved using different heuristic optimization techniques such as stochastic gradient descent, RMSprop, and Adam optimization algorithms.

Figure 15 illustrates the process of GANs training for a compiled GANs model. The available actual data samples are divided into multiple samples of the same size, each called a batch. A complete training cycle over the entire batches is called an epoch.

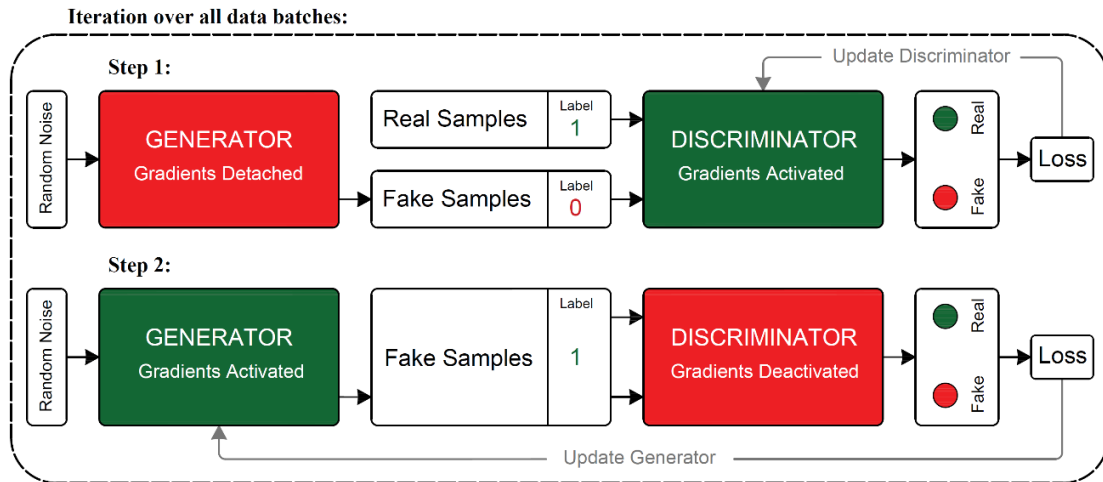


Figure 15. The training process for a compiled GANs model

The first step in training a GANs model with a batch of data is to generate a set of fake samples using the generator. The generated samples are detached from the generator, meaning that the generator's weights will not be updated. The fake data (labeled 0) is combined with real data (labeled 1) and inputted to the discriminator. This way, we only train and update the discriminator's weights through the backpropagation of the gradients calculated for each neural network layer. In the second step, the fake samples from the generator are labeled as real samples (labeled 1) and fed to the discriminator. In this step, the gradients are calculated and backpropagated to the generator. However, the discriminator is not allowed to tune its weights. This way, the generator finds the opportunity to tune its weights to produce more realistic results. The algorithm for GANs training is detailed in Table 4 for a single batch.

Table 4. The algorithm to train the GANs model.

Pre-training the discriminator
Train the discriminator using samples of real sociodemographic data and false data
Main Loop
Discriminator Loop (k-times)
<i>Fix the generator</i>
<i>Label the fake data from the generator and samples of real data by 0 and 1, respectively</i>
<i>Train the discriminator using the labeled fake and real data</i>
Label a sample of fake data from the generator by 1
Fix the discriminator
Train the generator weights by maximizing the error made by the discriminator

5.1.2. Challenges to be addressed

A few challenges must be tackled before implementing the GANs model. In this research, a couple of statistical methods are proposed to be integrated with the basic GANs model to generate a synthetic population and estimate the demand for evacuation models. These challenges are highlighted below:

- 1- How to provide false data that is from the null space of the real sociodemographic data?

- 2- GANS is applied in image and video processing and is rarely used on 1D data. Is it applicable to transportation data? Are the CDFs of the data trainable with GANS?
- 3- If all the challenges mentioned above are tackled, how can we associate the data derived from GANs to micro areas?

The solutions to each challenge will be discussed in the following sections, which are supported by shreds of evidence from actual examples.

5.1.3. Generating fake data feed for discriminator pre-training

Generating false information from outside the real data distribution is only helpful if we would like to perform a pretraining for the discriminator network. This pre-training should not be done to perfection as it may increase the possibility of overfitting for GANs model. If overfitting happens, the discriminator will recognize the fake data and not allow the generator to get closer to the real data distribution. One approach is to use a limited number of data samples from inside and outside the real data distribution and continue training alongside the generator in GANs model.

Generating data from outside the real data distribution (Null space) can be performed using the singular value decomposition concept. In our case, we have a large number of observations for nine sociodemographic attributes extracted from PUMS data and IRS income data. To generate data that is entirely different in terms of dependencies, we can randomly draw samples using the inverse Variance Covariance matrix:

$$[VC]^{-1} = \left[\frac{1}{N-1} [A^T A] \right]^{-1} \quad (18)$$

Singular value decomposition on inverse Variance Covariance matrix gives:

$$[VC]^{-1} = [U][S][V^T] \quad (19)$$

U and S are nine-by-nine eigenvector and eigenvalue matrices for the inverse Variance Covariance matrix. A sample from outside the distribution of the actual data (Null space) can be drawn from a distribution (here we assume a normal distribution) with the standard deviation of S , using the following equation:

$$fake\ sample = [U][random\ value(var = S)] \quad (19)$$

5.1.4. Applicability of GANs to one-dimensional data

As mentioned previously, GANs are widely used in image, video, and sound processing, and the utilization of the GANs for structured data is minimal. In addition, to the best of our knowledge, the application of GANs in modeling the dependency between the marginal distribution of the variables in transportation data is not explored yet. In this section, we generate three attributes from a simple spiral function to examine the applicability of GANs to structured datasets. Figure 16 shows the shape of this function.

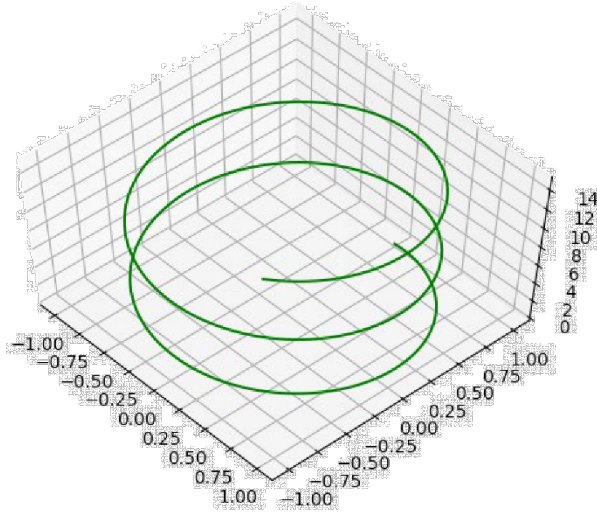


Figure 16. A simple Spiral function used to explore the applicability of GANs to generate the samples from marginal distributions of the variables

The formulation for each attribute of this function is as follows.

$$z = [0,15] \quad (20)$$

$$x = \sin(z) \quad (21)$$

$$y = \cos(z) \quad (22)$$

To create a GANs model for this function and generate synthetic samples from the data distribution, a large number of true values are created using the exact function, and a tiny noise value is added to defer slightly from the exact numbers. This could be analogous to the Heteroscedasticity effect or heterogeneity in people's taste. These values are considered real data samples. The false data is generated using the value decomposition method mentioned in the previous section. The latent space is chosen to come from three Gaussian distributions with mean and standard deviations chosen to be $N(1,5)$, $N(5,10)$, and $N(10,15)$. The discriminator and the generator are trained once for 128 observations in each iteration. Figure 17 shows the synthetic data generated by trained GANs after 50, 200, and 10,000 iterations.

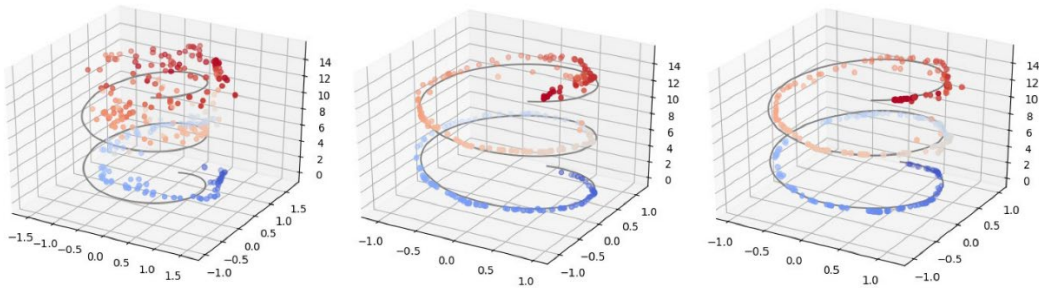


Figure 17. synthetic data generated by trained GANs after 200, 5000, and 10000 iterations

Figure 18 shows the discriminator's accuracy in identifying the fake and real samples for each set of inputs for each iteration. It shows that the average accuracy is somewhere between 70 and 80 percent, which is a good sign. The reason is that if the accuracy is 100 percent, it indicates that the generator cannot capture the dependency structure (joint distribution) of the data, and the discriminator quickly identifies the fake data for 100 percent of the situation.

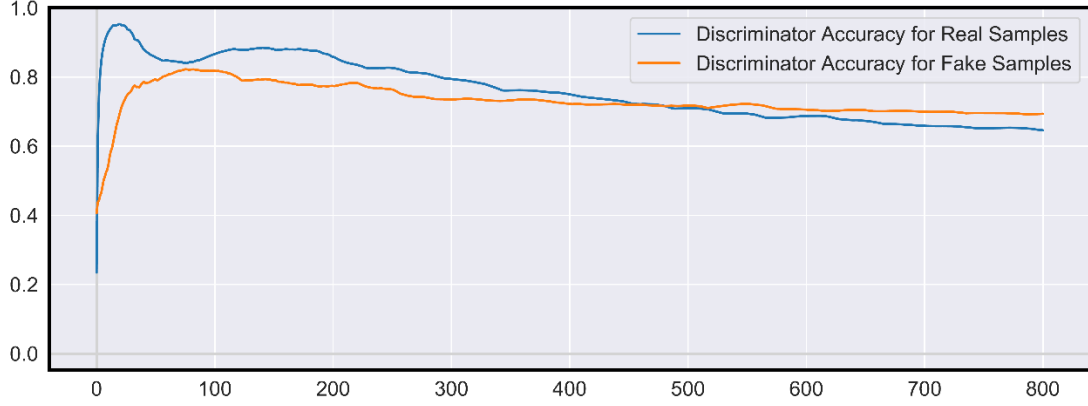


Figure 18. Accuracy of the discriminator in identifying the fake and real samples

Figure 19 shows the discriminator loss for actual and generated fake data and the generator loss for generated fake samples. The discriminator loss for real and fake data hovers around 0.5, a Nash Equilibrium for a two-player minimax game.

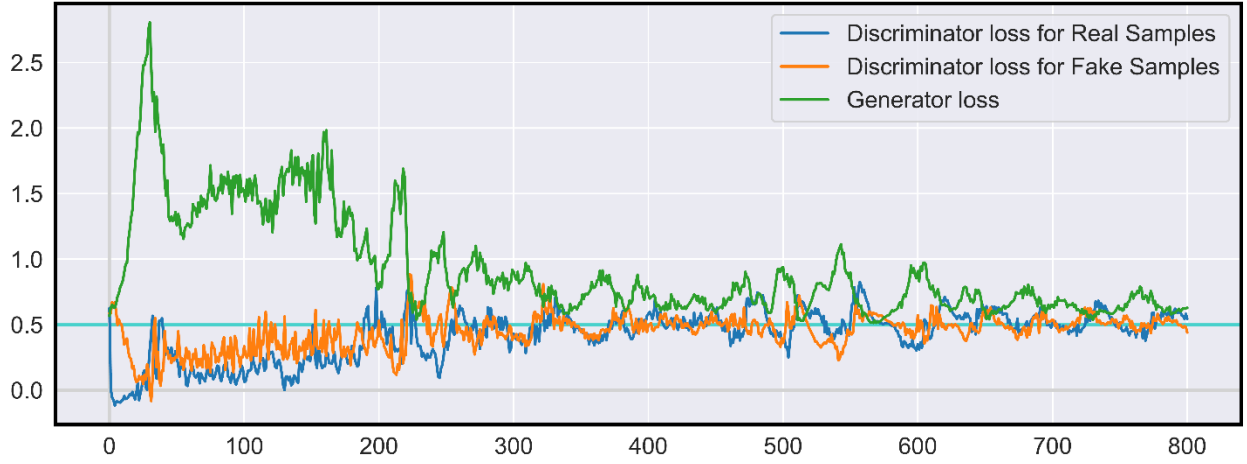


Figure 19. Accuracy of the discriminator in identifying the fake and real samples

The same GANs structure is used to examine the ability of GANs to preserve the dependency between the marginals of the variables. Ten thousand data points are drawn from the spiral function (see equations 20-22). The samples' CDFs for each attribute are calculated, and the CDF realizations associated with each data point are introduced to the GANs model. The distribution of the real inputs (marginals) is shown in figure 20 (a).

Once GANs model is trained, the inverse CDFs are used to reconstruct real samples and plotted as shown in figure 20 (b) for iterations 50, 200, and 10,000. The generator can capture the dependency between the variables when inputting the marginals of the data, but the distribution of the data is not well captured. This problem is currently being investigated for potential data or GANs structural issues or mistakes.

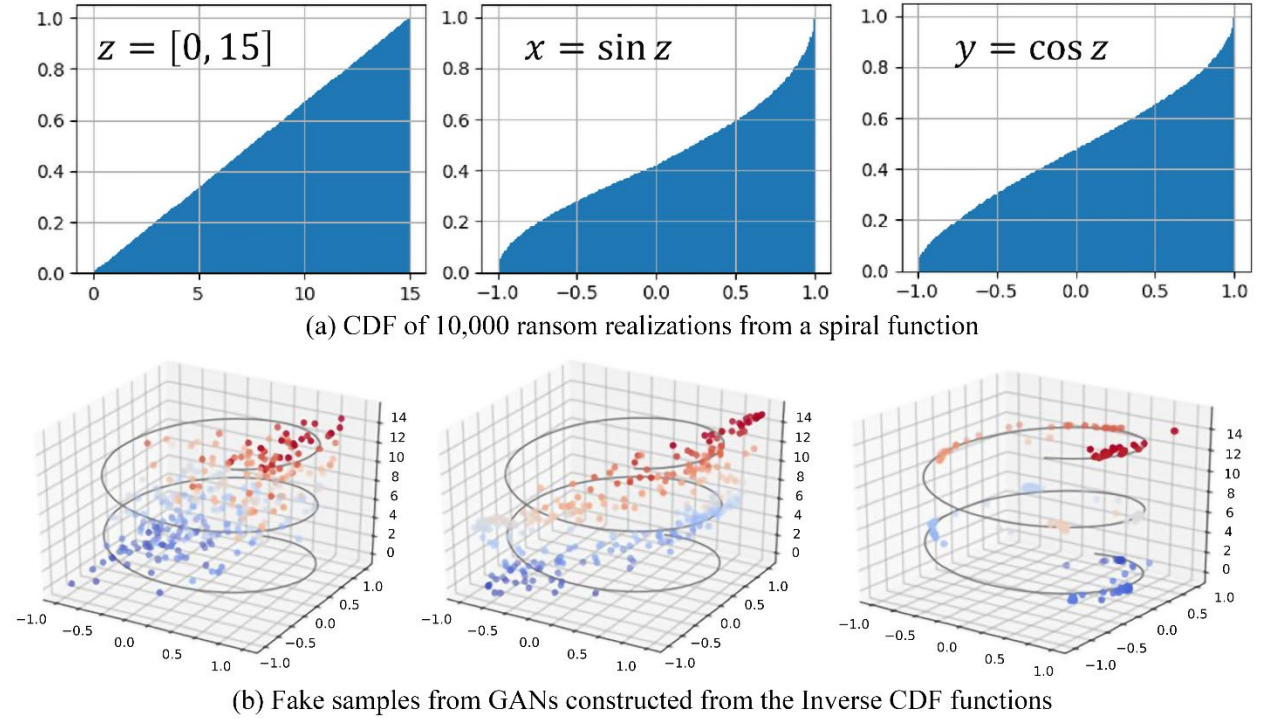


Figure 20. (a) Generating CDFs and inverse CDFs for 10,000 random samples from spiral function (b) The generator samples after 50, 200, and 10,000 training iterations

The following section discusses the data preparation steps for GANs' inputs, and different data variants used in the proposed models are illustrated.

5.2. Data preparation

Since this research is part of the evacuation study project where the car-ownership of the entire population is of interest, the attributes of the sample data are limited to the decisive variables in identifying the people's car-ownership status. In Chapter 4, we highlighted the critical variables in capturing the car-ownership status of the people in Maryland, USA. We found that Race,

Income, number of workers in the family, and the number of children play an essential role in their decision to own a car. A series of binomial models are developed to identify each attribute's best class aggregation. We also include the age of the people in our data to increase the data dimension and variability. Table 5 summarizes the data attributes and classes used in this study.

Table 5– data attributes and their classifications

Attribute	classes
Age	Used ten classes of 10-year-interval between 0 and 100 years old
Race	White and Others (1)
	African American (2)
	Asian (3)
Family Income	<\$25k (1)
	\$25k< and <\$50k (2)
	\$50k< and <\$75k (3)
	\$75k< and <\$100k (4)
	>\$200k (5)
Status of children in the family	From a family with children under six years old (1)
	From a family with children between six and 17 years old (2)
	From a family with children from both two categories, 1 and 2 (3)
	From a family with no children (4)
	Single (5)
Workers in the family	No Workers in the family (1)
	Possessing a vacant house (2)
	One worker in the family (3)
	Two workers in the family (4)
	Three or more workers in the family (5)

The data for one PUMA (PUMA 1201) in Anne Arundel County, Maryland, US, is obtained from the American Community Survey and aggregated to the classes mentioned in Table 1. In addition, a one-hot vector has been created for each attribute, and these one-hot vectors are concatenated. The concatenated one-hot vectors add up to 28 variables; the first ten indicate age status, the next three indicate race, the following five indicate income status, and the following five are workers' status in the family. Finally, the last five variables indicate the status of the children in the family. For example, an Asian person with the age of 28, an income of \$74,000,

two workers in the family, with one child below six years old, is represented by a one-hot vector shown in Figure 21 (a).

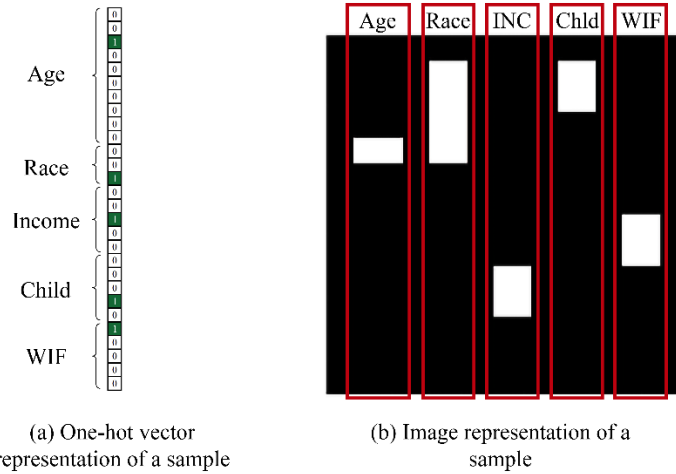


Figure 21. Two samples represented as (a) a one-hot vector, and (b) an image

We also prepared a second version of the data in a different structural format. The purpose is to show the application of Convolutional Neural Networks in generating synthetic data in transportation engineering. For this set of data, we transformed each data into an image. Each image codes the data attributes and their classes in a specific location of the image. Figure 21 (b) shows an example of data observation coded in an image. The data point, shown in Figure 21 (b), represents a person identified as white, aged between 30 and 40 years, with a family income of greater than \$200k, with at least one child under six years old and two workers in the family. The images are created by specifying a column for each attribute. The vertical position of the white pixels in each column indicates the class for each attribute. The columns are concatenated, and a band of black pixels is used to separate the attributes. The remainder of this chapter will discuss the models that utilize these data sets.

5.3. Wasserstein GANs

The Basic GANs model can handle many available problems, such as producing handwritten digits from a dataset. However, numerous problems may occur if the dimension and

distribution of the data become complicated. The most relevant issue is that the basic GANs are not always stable, and an unstable model would result in a problem called Mode Collapse. Mode Collapse happens when the generator learns to generate only one mode of data in the data distribution instead of the whole distribution. A mode is a group of data in the sample distributed around a central point, and each dataset may have multiple modes. For example, in generating handwritten digits, each digit is a mode, and there is a central handwritten that is probably most common in handwriting. To overcome this problem, Arjovsky et al. (2017) suggested defining the GANs model's loss function as the horizontal distance of the actual data distribution from the distribution of the data generated by the generator. Since WGANs try to criticize how far the actual distribution is from the generated distribution, the evaluating component of the WGANs is called the critic instead of the discriminator. The challenge here is that, for the WGANs model to converge, the loss function should be 1-Lipschitz continuous, meaning that the gradient of the loss function should be less than 1 for every point (every sample). This could be achieved by clipping the weights to limit the updating extent of the weights. Weight clipping is easy to implement but may limit the learning ability of the critic. A more stable and practical way is to add a regularization (penalty) term to the loss function of the WGANs model that penalizes the loss function if the gradient is greater than 1. The regularization term calculates the distance of the norm-2 of the gradient for an intermediate sample from unity. This intermediate sample randomly shares a portion of the actual data and a portion of the data generated from the generator. The final expression of the WGANs' loss function is as follows:

$$\min_G \max_C V(D, G) = \mathbb{E}[c(x)] - \mathbb{E}[C(G(z))] + \pi \mathbb{E} [\|\nabla c(\hat{x})\|_2 - 1]^2 \quad (23)$$

The C and G are the critic and the generator, respectively. π is a parameter to regulate the magnitude of the penalty. x is a real observation and $\hat{x}$ is an intermediate sample. An intermediate

sample should cover all possible samples between the output space of the generator and real data. A practical solution is to sample it uniformly along straight lines between the real and fake samples:

$$\hat{x} = \epsilon x + (1 - \epsilon)G(z) \quad (24)$$

The training process is no different from the process for the basic GANs model except for the number of times that the critic is trained before updating the generator. Studies showed that training critic a couple of times for each generator training produces the most stable model. In what follows, we will discuss the model structures used in this study to generate synthetic populations for subareas of a larger area.

5.4. The structure of WGANs models examined in this study

Many variants of the WGANs model in this study use an MLP neural network in its simplest form to model the critic. The critic consists of two layers of linear transformation. A Leaky ReLu activation function follows the first linear layer. The discriminator in the basic GANs model uses a sigmoid function in its last layer to model the probability of being fake or real. In the case of WGANs, the distance between real and fake distributions can be any value, and the sigmoid function must be avoided. Figure 22 illustrates the structure of the critic used in this study.

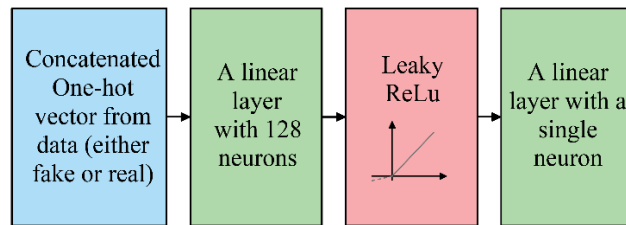


Figure 22. Critic structure used for most of the models in this study

We examined a couple of different structures to model the generator. The first and simplest generator takes six random numbers from the normal distribution as a noise vector and generates 28 outputs. Hereafter, we call this model GEN-1. The last layer of the network uses a sigmoid

function to map the previous layer's outputs to a value between 0 and 1. During the training phase, the exact outcome of the generator is passed to the critic for evaluation. However, during the generation of samples from the trained generator, the 28 outputs are deconcatenated to resemble the one-hot vector for each class (i.e., age, race, income, number of workers in the family, and the status of the children in the family). A max function is then used to determine the class of the attribute. Figure 23 illustrates the structure of this generator.

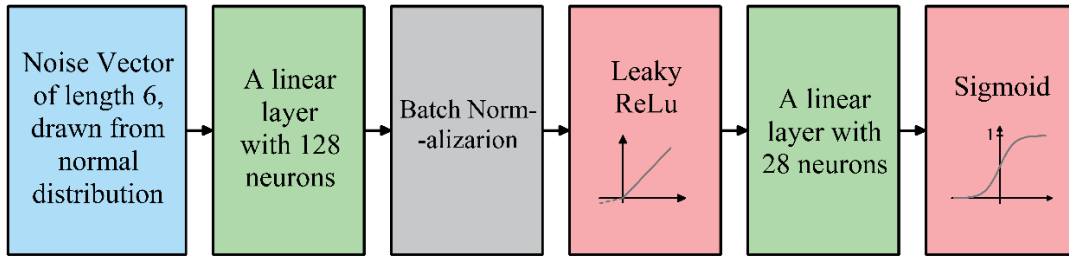


Figure 23. Generator's structure for GEN-1 model

The second structure used to model the generator (GEN-2 model) benefits from the decomposition of the outputs of the last linear layer. The first block of this network includes a linear layer with 128 neurons, a batch normalization, a Leaky ReLU activation, and a linear layer with 28 neurons. In this structure, the output of the last linear layer is divided into five vectors of lengths 10, 3, 5, 5, and 5. Then a softmax activation function is used to model the probability of falling in each class for the components of each decomposed vector. These probabilities are concatenated again and directly introduced to the critic during the training phase. The class with the highest probability is assigned to the associated attribute during the sample generation. Figure 24 illustrates the structure of this generator.

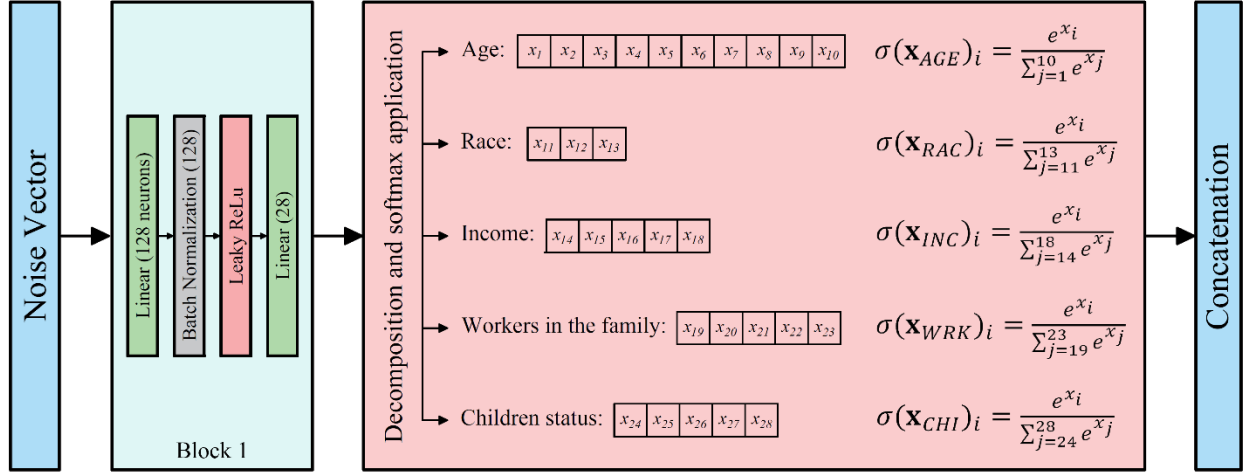


Figure 24. Generator's structure for GEN-2 model

Block 1 in Figure 24 can take a more complicated neural network structure if the data structure and attributes are complex. A complex deep neural network structure for deep learning that is established recently uses deep convolutional layers to learn the structure of the input data and the complex dependencies in data structures. Convolution is a technique to use filters to detect the features in the dataset. A filter (kernel) is a convolution matrix of any size that skims over the dataset and captures the features that are not only dependent on one input but multiple inputs in the dataset. Each kernel consists of learnable parameters that find features with simple mathematical operations. The research in generating synthetic images used convolutional GANs in its mastery level; therefore, we also prepared an image version of the data set and designed a 2D convolutional generator and discriminator. The process of producing the dataset was mentioned in section 5.2. We named this model GEN-Conv2D. GEN-Conv2D model uses a discriminator with convolutional layers, which takes a batch of fake and real samples coded as images and outputs a single value that determines the probability that the data is real.

On the other hand, the generator starts with a noise vector and produces an image through transposed convolutional layers. In GEN- Conv2D, the generator tries to produce the images that

code the distribution of the real dataset. The structure of the discriminator and generator for GEN-Conv2D is shown in Figure 25.

The GEN-Conv2D model is illustrated to visually describe the feature extraction step in convolutional GANs. The problem with applying 2D convolutional layers is that there is no dependency on the height of the images produced from the data. The data dependency is between the attributes, which varies on the horizontal axis of the images. In other words, the classes for each attribute are independent of other classes. Therefore, the generator may produce images that have more than one class for a single attribute. More details are provided in section 5.6.

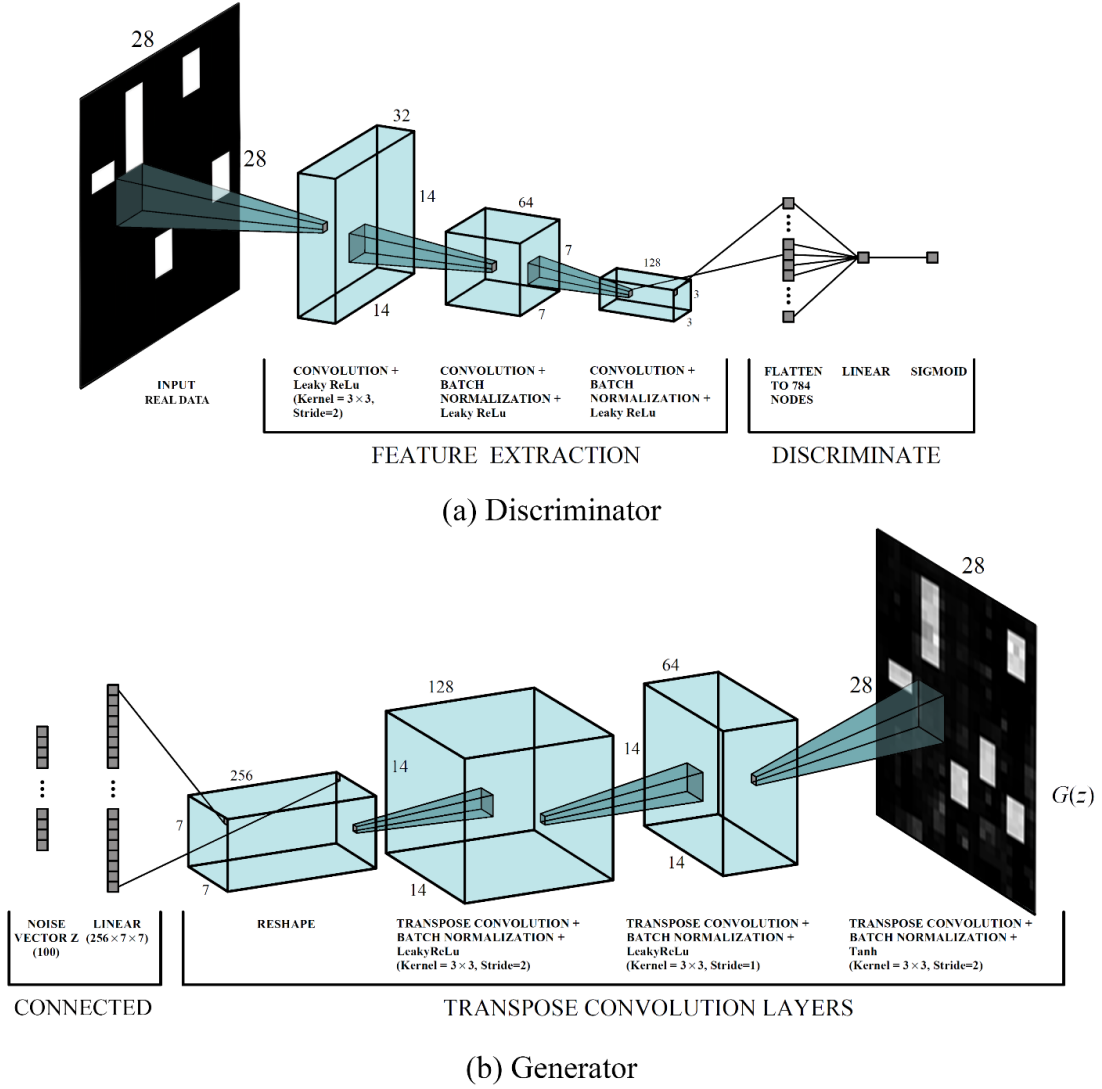


Figure 25. The structure of the discriminator and generator for GEN-Conv2D

To benefit from the feature extraction in convolutional GANs, one can use a 1D convolutional network as a generator and a simple MLP model as a critic in WGANs. Hereafter, we call this model GEN-3. The generator in GEN-3 is the same as GEN-2, where block 1 is substituted with four 1D transposed convolutional layers. The first three layers include a batch normalization layer followed by the ReLu activation function. The last layer's output is passed to the decomposition and softmax block mentioned in the GEN-2 model. The kernel size and stride pairs are (3,2), (4,1), (3,2), and (4,2) for the first, second, third, and last layer, respectively. The

generator's input is a noise vector with 64 elements drawn from the normal distribution. The numbers of channels for the first to the last layers are 256, 128, 64, and 1, respectively. This leads to 28 outputs for the output of the model. The critic in GEN-3 is the same as the critic model used for GEN-1 and GEN-2.

5.5. Sample generation for each micro-area

As previously mentioned in the Introduction, the central issue in generating synthetic populations for small areas is that the joint distribution for the variables is not available for small areas. A mathematical solution to this problem is to estimate the covariance matrix for each micro-area. After training the GANs model for PUMA samples, One can assign these samples to each micro-area using the maximum likelihood (ML) function:

$$P = \frac{\prod_i^{N_v} P_i}{\max(P_i)} \quad (25)$$

Maximum likelihood assignment assumes that the attributes of the data are entirely independent, which is contrary to the fundamental assumptions of the problem. Therefore, the error associated with this assignment approach has to be modeled to achieve better estimates for the covariance matrix of the smaller areas. The goal is to minimize the difference between the covariance matrix of the total data and small areas' contribution in the covariance matrix. This approach models the error resulting from the independence assumption in maximum likelihood assignment mentioned in equation 26:

$$\min_{e_1, e_2} \left\| C - \sum_{i=1}^{N_{MA}} \left[\left(\frac{N_i}{\sum N} \right) \times C_i(e_i) \right] \right\|_f^2 \quad (26)$$

where C is the data covariance and C_i is the modified covariance matrix for subarea i .

$$C_i(e_i) = \sum_{x \in \Omega_i} (x - \mu_1 - e_1)(x - \mu_1 - e_1)^T \quad (27)$$

Ω_i is a set of samples for subarea i that is assigned by the maximum likelihood approach. The errors (e_i) are obtained and used to estimate the covariance matrices for each micro-area. Then these covariance matrices can be used to measure the probability of each observation generated by GANs model belonging to a small area.

There is a limiting restriction with this model. The model only works for the cases where the number of sub-areas is smaller than the number of attributes. Otherwise, equation 26 results in an underdetermined system of unsolvable equations. This approach is not practical for the PUMA level data. The subareas of the PUMA level are Census Tracts, and a PUMA consists of tens of census tracts. Another solution is to assume that the contribution of each census tract in the multivariate distribution of the data is the same. This assumption is weak and may result in false sample assignments. In this approach, the cumulative distribution of the data is used to train a WGANs model. Then the generated samples for each census tract are transformed into synthetic samples using each attribute's inverse cumulative distribution function for each census tract.

The last two approaches are either impractical or introduce significant errors to the model. In this study, we propose the application of conditional WGANs to produce synthetic samples for different classes of data. Assume a second pair for the spiral function mentioned in equations 20-22 obtained by adding 3, -0.8, and -0.8 to z , y , and x . Conditional GANs can generate the samples for each of the spiral functions, given that the real data is labeled to indicate the belonging spiral function for each sample. Figure 26 shows the samples generated from conditional GANs for each class. The generator and the critic losses are shown in Figure 26 as well. The generator loss on WGANs does not imply any intuition for the model's performance. On the other hand, the critic's

loss provides the difference between the distribution of the fake and real samples. As this value gets closer to zero, the model approaches the global optimum of the loss function.

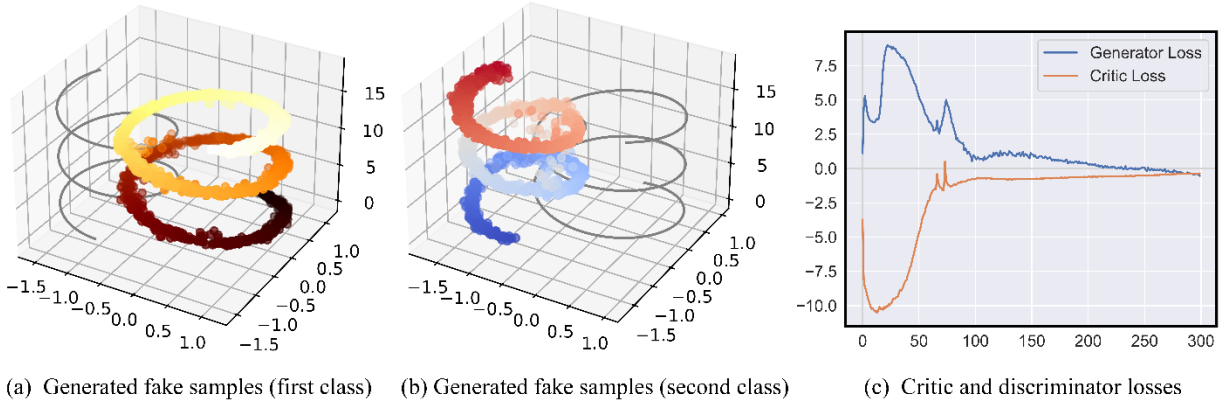


Figure 26. An example of conditional WGANs using the data from two different spiral functions, (a) Samples generated for the first spiral function, (b) Samples generated for the second spiral function, (c) losses for the critic and the generator for the first 6,000 iterations (each data point is the mean for the last 20 iterations)

In modeling the conditional WGANs to generate synthetic populations for micro areas, at least a few samples for each micro-area should be labeled with the correct micro-area number. Otherwise, another data source should provide additional labeled data. Here, for the sake of simplicity, we assume that the entire dataset has a record indicating the micro-area to which each data point belongs. The generator's input noise vector is concatenated with a one-hot vector that indicates the class (micro-area) for which the generator produces a sample. The real data is also concatenated with a one-hot vector coded with each sample's micro-area. The fake data produced by the generator is also concatenated with the original inputted one-hot micro area vector. The generator knows for which class a sample has to be generated, and the critic considers the location of the data point in discriminating the fake and real samples. The approach can be applied to all the abovementioned models except for GEN-Conv2D. For a critic with 2D convolutional layers, additional channels can be introduced to the first layer. The number of additional channels is equal to the number of subareas, and one of the channels (a matrix with all-one elements) represents the

subarea of the sample. Figure 27 shows the concatenation of the noise vector for different models for the critic and the generator.

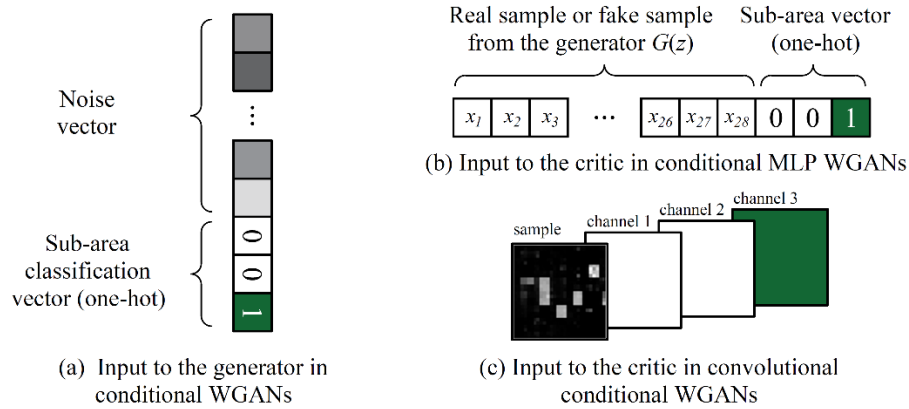


Figure 27. Inputs to conditional WGANs (a) Generator Input (b) critic input in the MLP model (c) critic input in the convolutional model

If the labels are only available for a few samples in the data or a labeled dataset is available from another data resource, one can first train a WGANs model with zero labels for the entire dataset. This helps utilize more data to capture the dependency and distribution among the entire data attributes. Once the training is complete, a new training step with new labeled data is initiated. During the second round of the training, the model will calculate the gradients and tune the neurons' weights to capture the label information.

As highlighted above, for the sake of convenience in this study, we assumed that we have the subarea records for the entire dataset. It should be noted that this study aims to show the potential of the WGANs model in generating the synthetic population for transportation research, and the applicability of the models depends on the data availability. Here, we assume that the interested PUMA consists of three micro-areas. We assumed that only people between ages 1 to 30 live in one of the areas, The second and third micro-areas accommodate people between ages 30 to 60 and 60 to 100, respectively. Although this assumption is not realistic, the comparison of the distributions is easier with this data segmentation.

We used two measures to compare the results from GANs model and the actual samples from the PUMA dataset. The first measurement in comparing the generator results with real data is the difference between the distribution of both data. We generated the same number of samples using the trained WGANs and calculated the percentage difference from the real distribution for each attribute class.

The second measurement is the difference between the covariance matrix of the real samples and fake samples produced from the GANs. We used a variant of the Correlation Matrix Distance (CMD) proposed by Herdin et al. (2005). Assuming that C and $C_{G(z)}$ are the covariance matrices of the real and fake samples from GANs, the non-orthogonality (NO) between two correlation matrices is defined by:

$$NO = \frac{tr\{CC_{G(z)}^T\}}{\|C\|_f \|C_{G(z)}\|_f} \in [-1, 1] \quad (28)$$

Equation 28 measures the distance from the case where the two correlations are orthogonal. Therefore, a value of 0 means these two matrices are not correlated, and a value of 1 indicates that the covariance matrix for both samples is the same. A value of -1 indicates that these two matrices are negatively correlated.

Section 5.6. discusses the results for all the GANs models mentioned above. The conditional WGANs are integrated only with the GEN-2 model, which appeared to be more stable and easier to implement.

5.6. Results of the GANs Models on Sociodemographic Data

Among the models mentioned in the previous section, GEN-Conv2D generates images that can be visually judged. First, the GEN-Conv2D model is used to show possible problems that may arise with basic GANs. Figure 28 shows the mode collapse problem in which the generator is

trapped into a local optimum and generates only one sample from the distribution. The average of the pixel values for each attribute class is reported in the images.

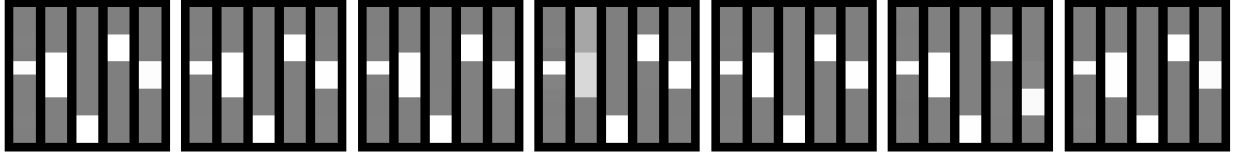


Figure 28. Mode collapse after training convolutional GANs with GEN-Conv2D model

Employing WGANs, the result of the model is improved, and the model is capable of capturing more samples from the distribution. Figure 29 shows the results of the GEN-Conv2D model within WGANs training structure after 100, 1200, and 3000 iterations. Although, based on further exploration of the results, the distribution of attribute classes is close to the actual data distribution (not reported here); the model sometimes assigns two classes to an attribute or leaves the class unidentified for some rare cases. This problem is observed for attributes with more classes as there is more chance of multiple correlations with other attributes. Therefore, it is necessary to employ probabilistic approaches to assign classes to attributes after sample generation.

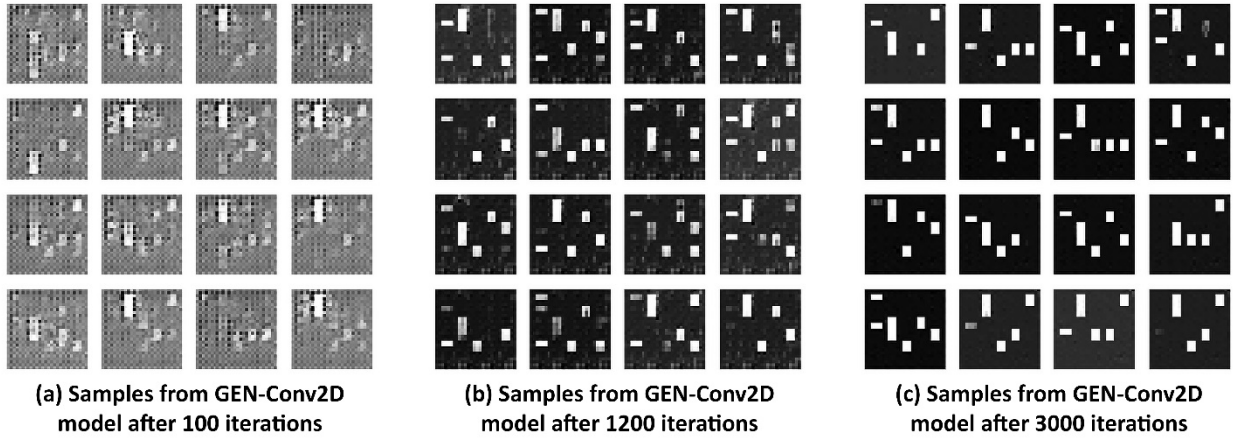


Figure 29. WGANs results with GEN-Conv2D model after (a) 100 (b) 1200, and (c) 3000 iterations

Using GEN-1 (Figure 23) as a generator and MLP critic (Figure 22), a WGANs model is trained to generate a synthetic population. GEN-1 model uses a single value to identify each class in a one-hot vector, making it easier to generate synthetic samples after the training phase using a

simple max function. Having trained the WGANs model, a sample with the same size as the real data size is generated. Figure 30 (a) shows the distribution of the real data and a fake sample (i.e., $G(z)$) generated by the generator that is trained for 1750 iterations. There are 170,756 observations in the real data for PUMA 1201. We generated 170,756 samples 100 times from the generator and produced a probability density function for each of the 100 sample sets. Figure 30 (b) shows the percentage of deviation from the true distribution. The percentage deviation from the true distribution is plotted to examine the range and magnitude of the deviation. Note that the same iteration number is used to train all the models in this study. It is clear from Figure 30 that the generator could not capture the density of the attribute classes that have a small density in the entire dataset. For example, the generator did not generate any samples aged 81 to 100 (class 9 and 10).

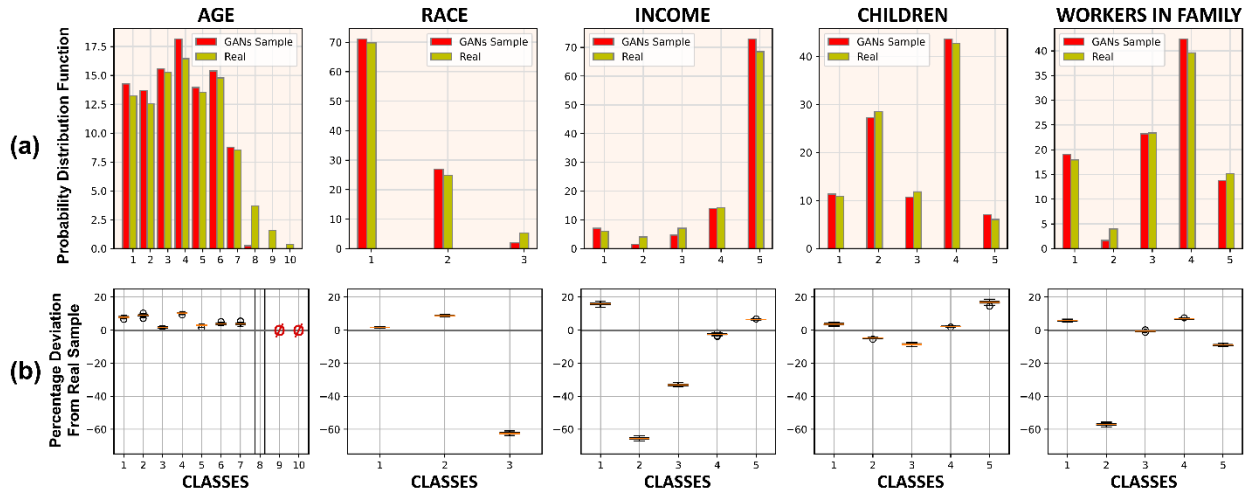


Figure 30. (a) The probability distribution function for a set of 170,756 samples from trained GEN-1 model (b) Percentage deviation from real data for 100 sets of 170,756 samples from trained GEN-1 model

Figure 31 shows the percentage deviation from real data for 100 sets of 170,756 samples obtained by training WGANs models with GEN-2 and GEN-3 generators. The models use the softmax activation function to simulate the probability of each class belonging to a specific sample. As expected, these two models perform better than GEN-1 for the same number of iterations (1750

iterations). For both models, the deviation from real data is less than 5% for most of the classes for all attributes. The deviation range appears to be higher for the classes with a lower density over the entire dataset. One surprising outcome is that both models performed almost the same in capturing the real data distribution. Despite the complex structure of the convolutional layers in GEN-3, it did not outperform the GEN-2 model within the same training iterations.

The convergence of the three models is explored using the critic's loss function, which indicates the difference between the distributions of the real and fake data. Figure 32 shows the track of the loss function for WGANs trained with GEN-1, GEN-2, and GEN-3 as the training iterations grow. The horizontal axis indicates the number of training epochs. Each epoch is a complete circle of training over the entire batches of data. Based on Figure 32, GEN-3 performs better than the other two generators, even though it shows more oscillations during the first three training epochs.

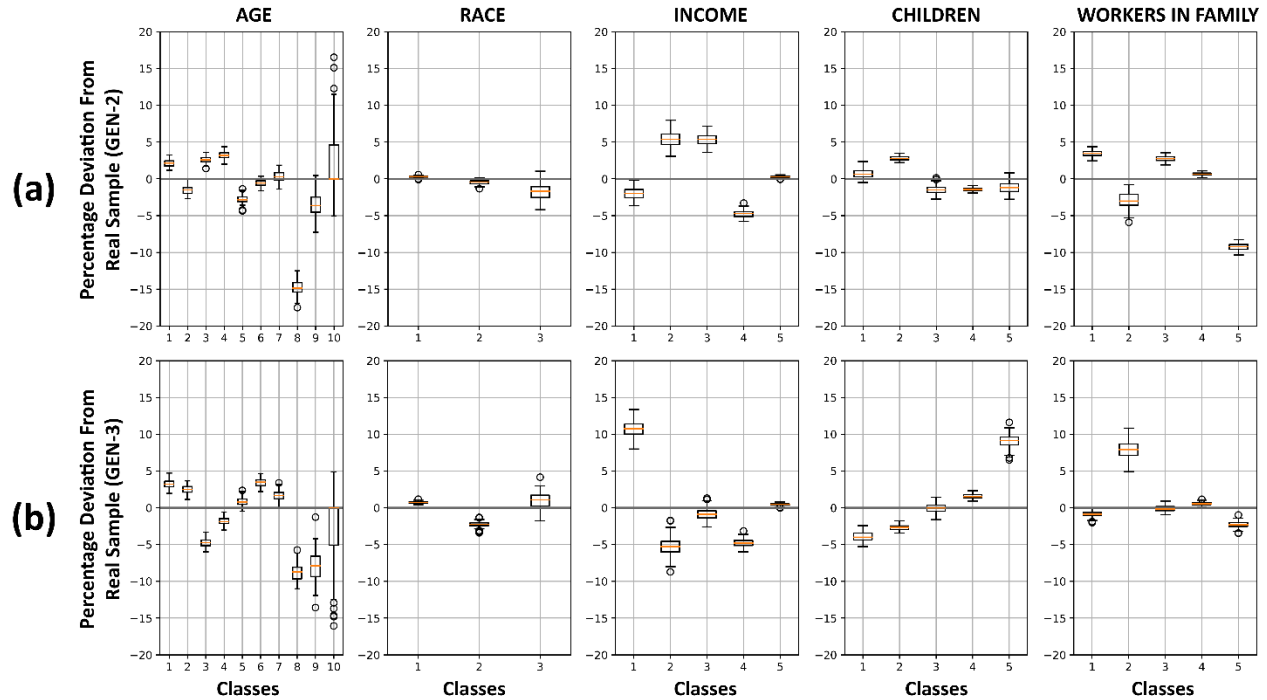


Figure 31. percentage deviation from real samples for 100 sets of 170,756 samples from (a) WGANs trained with GEN-2 generator (b) WGANs trained with GEN-3 generator

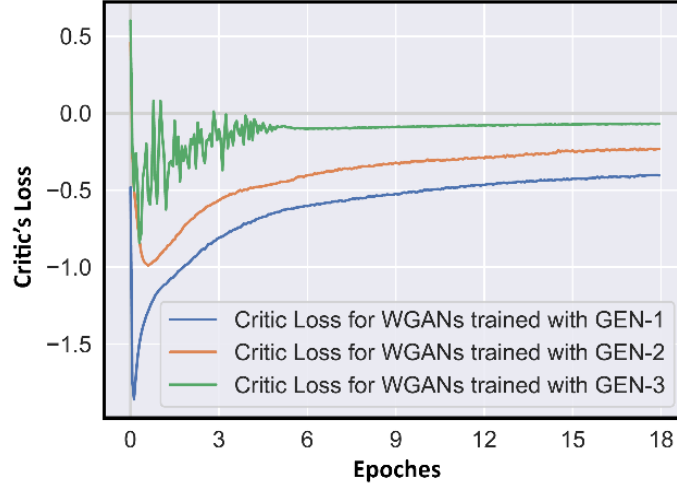


Figure 32. The loss function for WGANs trained with GEN-1, GEN-2, and GEN-4

Figure 33 shows the results for conditional WGANs trained with the GEN-2 generator. We integrated the GEN-2 generator with the conditional GANs as the result of this model is relatively close to the convolutional GANs, while the run time is relatively low. The model captured the distribution of the data for all three subareas. Interestingly, the generator did not generate samples for the classes with zero density in the real data distribution.

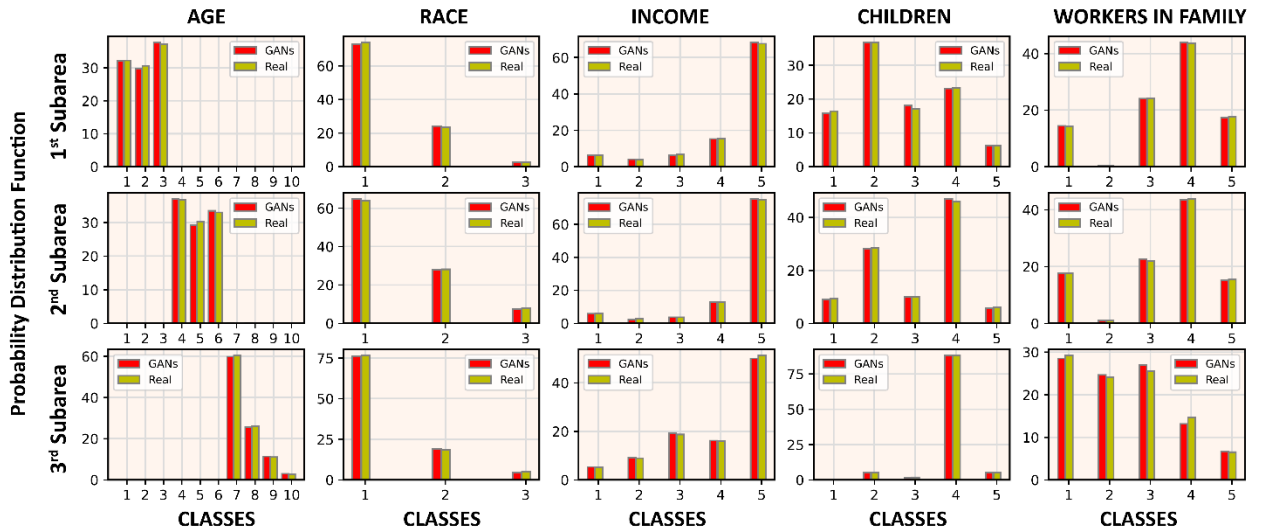


Figure 33. Probability Distribution Function for synthetic data produced by Conditional WGANs model for three subareas of PUMA

Figure 34 depicts the percentage deviation from real data for 100 sets of 170,756 samples obtained by training conditional WGANs model with GEN-2 generator. The data generated for

each subarea is extracted, and the deviation from real data is calculated separately. The deviation variance is higher for the attribute classes with a very low density in the distribution function. We expected this since the absolute value of the counts for these classes is so low that a slight variation in the absolute value would significantly affect the percentage deviation. Overall, the mean percentage deviation is below 5% for the majority of the classes, which indicates the model's good performance in synthesizing the data and capturing the distribution of attributes classes for each subarea.

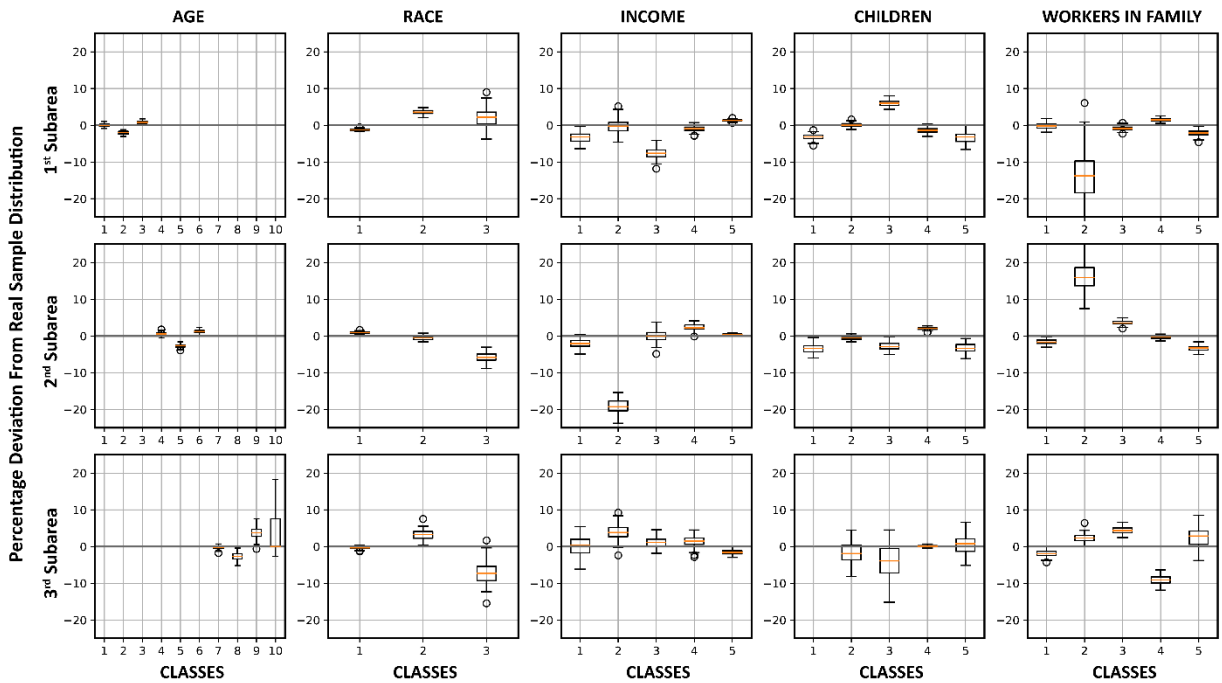


Figure 34. Percentage deviation from the real sample distribution for synthetic data (100 sets of samples) produced by the Conditional WGANs model for three subareas of PUMA

Sociodemographic information is the building block of many significant transportation engineering problems. Such data are widely used in predictive models to estimate the behavioral decisions and choices of travelers and transportation service users and customers. In such cases, the multivariate distribution of the data and the dependencies between the attributes are critical. Therefore, a synthesized dataset must preserve the dependencies between the attributes. We calculated the Non-Orthogonality (N-O) measurement for covariances of the real data and a single

set of synthesized data. Table 6 shows the N-O for WGANs trained with GEN-1, GEN-2, and GEN-3 models. Both GEN-2 and GEN-3 perform better than GEN-1 in capturing the multivariate nature of the real data. Although the GEN-3 model performs slightly better than GEN-3, we experienced higher run times for training the GEN-3 model. The reason relies on the fact that in addition to weights in its fully connected layers, a convolutional neural network also has to train kernel parameters.

Table 6– Non-Orthogonality measured for covariances of the real and the data synthesized with WGANs trained with GEN-1, GEN-2, and GEN-3.

Generator Model	GEN-1	GEN-2	GEN-3
Non-orthogonality measure	0.99443	0.99976	0.99984

Figure 35 shows the improvement of the non-Orthogonality measurement as the training epochs increase. GEN-3 performs quite volatile in capturing the dependencies until the 6th epoch of training but is quicker in stabilizing around 1.0 after the 6th epoch. GEN-1 and GEN-2, on the other hand, show more stability in convergence but requires more training epochs to achieve the same result as GEN-3.

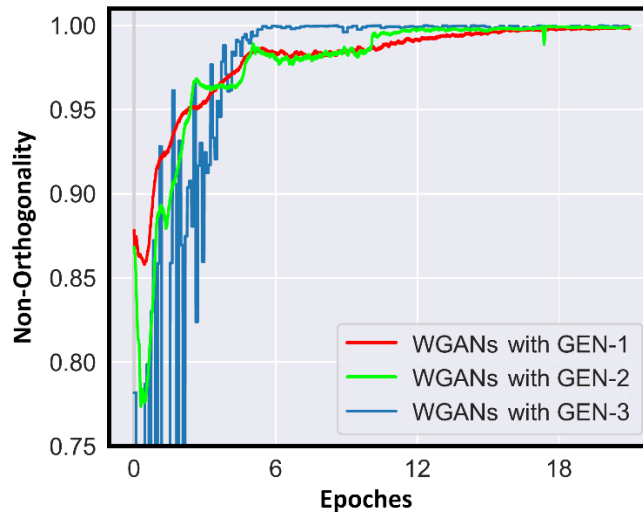


Figure 35. The change in non-Orthogonality as the training epochs increase

To measure the performance of the conditional WGANs, we constructed a square matrix that measures the non-orthogonality for covariances of the real and synthetic data for every pair of

subareas. This matrix is shown in Table 7. The diagonal elements of the matrix have the highest non-orthogonality, meaning that the synthetic sample set for each subarea preserves the multivariate nature of the data for that specific subarea. Subarea 3 shows the most difference in dependency with other subareas, which was expected from the data distribution (see Figure 33).

Table 7— Non-Orthogonality matrix for all possible pairs of covariances calculated for real and synthesized data using convolutional WGANs trained with GEN-2

Real \ Synthetic	Subarea 1	Subarea 2	Subarea 3
Subarea 1	0.99924	0.96656	0.88605
Subarea 2	0.96643	0.99974	0.88068
Subarea 3	0.88579	0.879313	0.99960

The primary goal of this research is to highlight the potential of the GANs model in generating synthetic sociodemographic data that represents the distribution and multivariate nature of the real data. Once the multivariate distribution of the synthetic data is examined and the robustness of the model is confirmed, augmented data for the population level can be generated using the generator, and the population counts from different data sources.

To apply and examine the performance of the proposed conditional Wasserstein GANs in generating synthetic data for the census tracts of the PUMAs in Anne Arundel County, we assumed that the synthetic population derived from Copula (See chapter 4) model is the real data. The Census Tracts can be partially labeled with the IDs obtained from the Copula model. A subsample of each census tract was derived using bootstrapping technique (sampling with replacement), and a portion of the data is labeled with the Geo-IDs. The derived sample was augmented using Iterative Proportional Factoring to represent the true distribution of its census tract. WGANs with GEN-2 were trained for the entire sample with empty (zero values) subarea classification vector. The model parameters were then transferred to a new model where the labeled data was introduced to the generator, and the entire GANs model was trained to learn the conditions (census tract IDs

of samples). The binomial model mentioned in Chapter 4 was used to obtain the car ownership status of the synthetic population generated by GANs. Note that each PUMA was trained independently by a standalone model to reduce the number of conditionals in Conditional Wasserstein GANs and increase the robustness of the model. Figure 36 compares the results of the GANs model with the results of the copula model. Figure 36 (a) shows the results of the copula model that was used to generate a synthetic population for each census tract using the PUMS samples. While copula results cannot be compared with benchmark data, assuming that the copula-generated samples are real, the performance of the GANs model with partially labeled data can be evaluated. Figure 36 (b) shows an excellent match between the pseudo-real data from copula and the synthetic data from GANs.

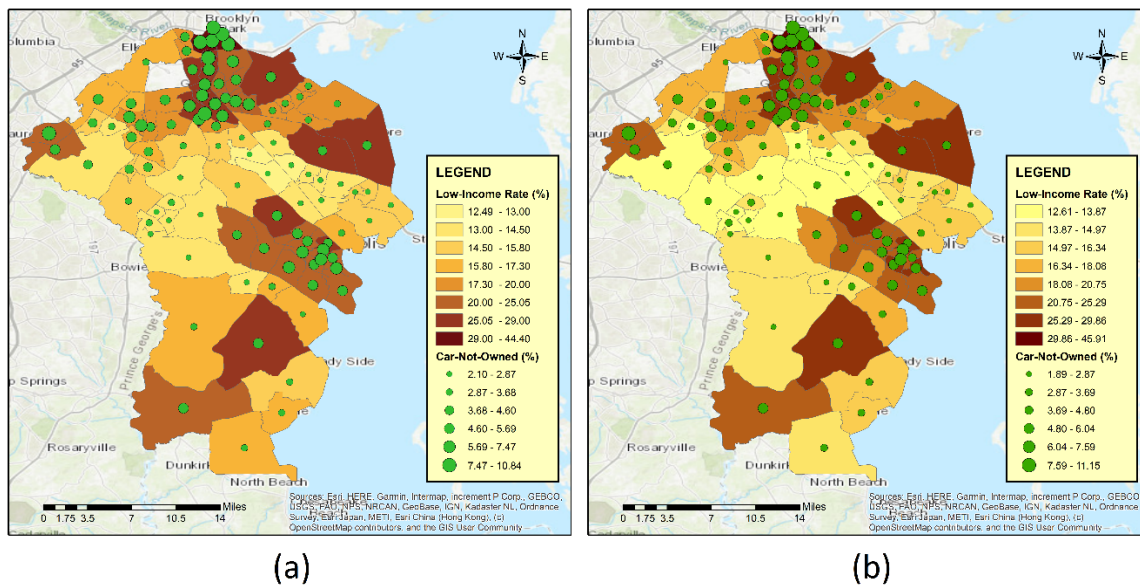


Figure 36. The change in non-Orthogonality as the training epochs increase

Chapter 6. Bimodal Evacuation Planning model

6.1. Understanding the requirements for flood evacuation planning

Comprehensive disaster management must address proper actions for four specific phases to minimize the risks and fatalities caused by a disaster: Mitigation, Preparedness, Response, and Recovery [4-5]. Preparedness and Response are crucial puzzle pieces that require implementing efficient strategies and actions to plan for potential disasters and manage the evacuation operation successfully. A well-structured flood evacuation planning and management needs the agents to properly understand the risk zones' geographical conditions and the availability of evacuees' data in desirable geographical levels to develop response strategies suitable and feasible for the area under study. For Anne Arundel County, Maryland, which is the case study of this research, there are insufficient data samples to estimate the demand for evacuation and transportation facilities. The previous sections proposed GANs approach to generate such data. The GANs model combined with a simple binomial model estimates the number of evacuees who have access to personal cars and the number of evacuees who mainly depend on public transportation systems at the census tract level. Figure 37 shows the number of FEMA shelters available to support evacuees' short-term accommodation during an emergency.

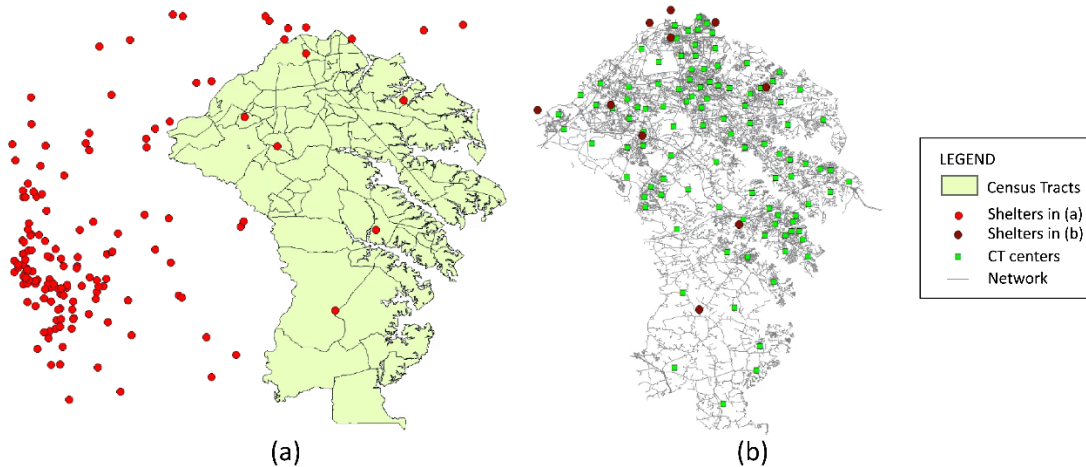


Figure 37. (a) FEMA shelters are scattered within and outside Anne Arundel County. (b) Transportation network and shelters within Anne Arundel County

Figure 38 shows the accessibility of census tracts to FEMA shelters based on shortest path criteria. From the figure, it is clear that the census tracts closer to the shoreline have longer distances from the shelters, making the evacuation even more difficult. In this analysis, sea-level rise (SLR) projections presented in a recent study by Boesch et al. (2018) are used to explore the possibility of road closure for two Representative Concentration Pathways (RCPs that indicate atmospheric concentrations of Growing Emissions (RCP 8.5) and Stabilized Emissions (RCP 4.5). The probability that the sea level rise meets and exceeds the presented road closure is one percent, indicating the worse scenario. However, it should be noted that the flooding situation can result in much higher exposure to high levels of water exposure than the sea level rise. We analyze the sea-level rise due to easy access to sea-level rise projection data. The gray areas on the map are census tracts that will be inaccessible due to sea-level rise.

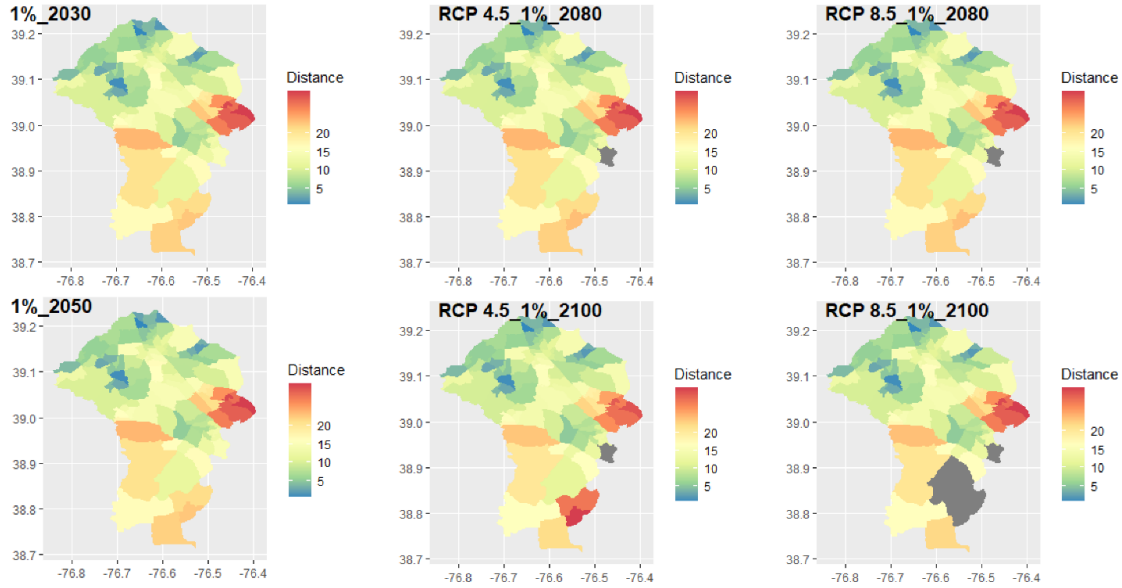


Figure 38. Spatial distribution of shortest path to FEMA shelters. Gray areas show dysconnectivity for different sea-level rise scenarios

6.2. Evacuation optimization problems

6.2.1. Flood evacuation problem specifications

Evacuation in motorized transportation problems can be formulated using static or dynamic approaches. In the static method, evacuation demand and hence traffic flow is fixed. Different User Equilibrium (UE), System Optimal (SO), Constrained System Optimal (CSO), and nearest allocation optimization formulations can be used to optimize the problem. On the other hand, the dynamic approach considers temporal changes in flow rates. Dynamic traffic assignment (DTA) can be solved with SO and UE formulations as well as Cell Transmission Models. Although it is easier to formulate and solve, System Optimal formulations are not realistic as they are intended to minimize the total travel time for all transportation facility users. This comes from the notion that users would usually not help reduce the total travel time at the expense of sacrificing their benefit in achieving short travel times. Also, operators can't control the travel destinations and routes due to the high dynamics of drivers' behavior and the considerable diversity in travel purposes. However, in evacuation problems, an area must evacuate, and the destinations are

predetermined. Therefore, System Optimal solutions are easier to be implemented using radio and mobile communication devices. The utilization and implementation of the System Optimal solutions favor transportation time-sensitive problems such as flood evacuation.

An evacuation optimization can be formulated to consider either deterministic or stochastic parameters. Deterministic methods assume fixed conditions for parameters such as demand, occurrence time, and availability of resources and facilities. On the other hand, Stochastic methods consider uncertainties in one or a combination of these parameters. Including stochasticity would add to the complexity of the evacuation problems. The use of GANs in predicting reliable demand for small areas allows for using deterministic approaches while still getting reliable solutions.

Another critical factor in developing accurate evacuation models is considering every available mode of travel. The evacuation problem can be formulated for private vehicles, which is not realistic because many people would not have access to personal cars. On the other extreme side, we can assume the mass transit system will evacuate everyone. A reliable and realistic approach assumes a multimodal evacuation system in which different modes of travel are being considered for people who own a car and people who do not have access to any personal vehicle. In the case of a flood using multimodal tours for a single trip from an origin to destination may create bottlenecks in the system. For example, using the urban bus system's routine schedule to access the subway is not possible. Besides, subway stations may not be available or exposed to high water volumes and may not be operational. A logical and realistic system is a bimodal private-public system where the bus schedules have been set for an optimal evacuation operation.

The last consideration for flood evacuation is the objective function. Different objective functions can be considered for evacuation problems. Choosing the right objective function depends on acquiring a thorough knowledge of the type of disaster and the lessons learned from

the previous experiences. Figure 39 illustrates the different objective functions that can be used for evacuation problems. The objective function for evacuation planning can be formulated to minimize the total evacuation time (TET) or the net clearance time (NCT). TET is an approach that tries to minimize the total evacuation travel times for evacuees.

On the other hand, NCT tries to minimize the time the last evacuee exits from the risk zone. NCT formulation is suitable for sudden disasters, and short notice is given due to a lack of decent predictions. On the other hand, TET is ideal for cases where the authorities are well acquainted with the situation, and there is no need for rapid evacuation.

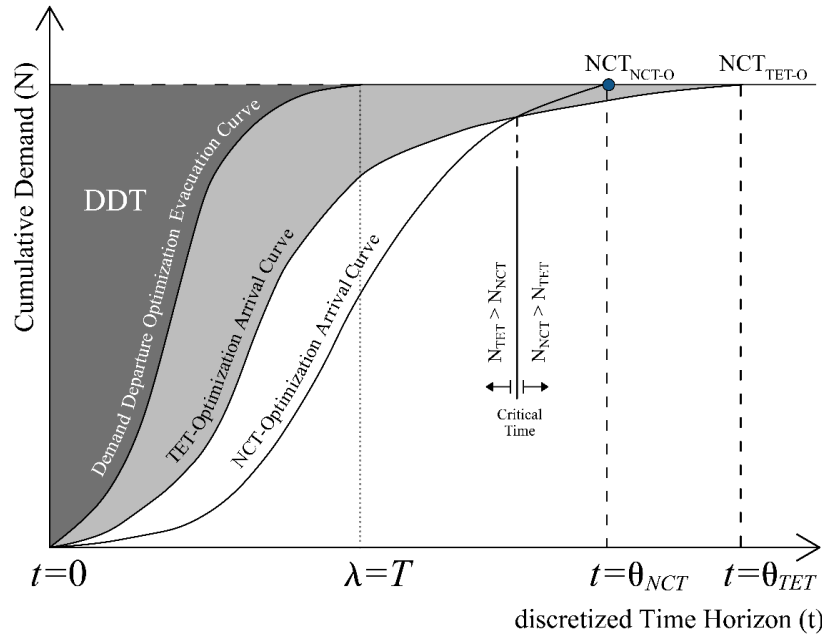


Figure 39. Illustration of different objective functions in evacuation problems

TET is superior to NCT because it reduces the travel times for individuals and keeps the traffic volumes low by releasing the evacuees gradually, but at the expense of a longer net clearance time. The cumulative demand arrival distribution curve for TET and NCT optimization has a critical point where the two curves cross each other (Xie et al., 2010). Before this crossing point, the number of evacuees who have arrived at the safe zone is higher for the TET optimization

approach. Therefore, in elusive disaster identification or when a disaster is expected to happen sometime before the critical time, TET-optimization is preferred over NCT-optimization. In the case of flood inundation evacuation, the potential for high-resolution predictions of the severe flood is high. Usually, there is enough time to let the evacuees leave the risk zone successfully. Therefore, the TET optimization approach is preferable where threatening flood inundation strike risks are well identified. Another method is to maximize the number of evacuees within a given evacuation time. This approach yields the same TET optimization results if there is enough time to evacuate the people. In this research, a bimodal, deterministic TET optimization approach is suggested that meets the demand parameters that can be obtained by using GANs generative model capabilities.

6.2.2. Problems associated with flood evacuation operations

Static and dynamic routing problems are subproblems of any traffic flow problems where new approaches are employed to route evacuees to destinations (Shelters) while minimizing a temporal or integer objective function. In the case of evacuation, shelter location and shelter assignment problems can be solved simultaneously or in a bi-level design. In bi-level design, two high-level and low-level mathematical optimizations are used to identify the best shelters (High-level problem) and route the evacuees to the shelters (Low-level problem). Signal and lane management strategies (e.g., Cross elimination, intersection control, lane closure, and contraflow traffic) are usually introduced as constraints to these problems. In multimodal problems, optimal bus scheduling, departure time, pick-up location, fleet size, bus routing for public vehicles, contraflow, and collaborative lane utilization are possible strategies that can be employed to ameliorate the evacuees' immediate escape from the risk zone that contains an imminent threat.

Figure 40 shows the strategies that can be utilized to achieve a robust model for successful evacuation.

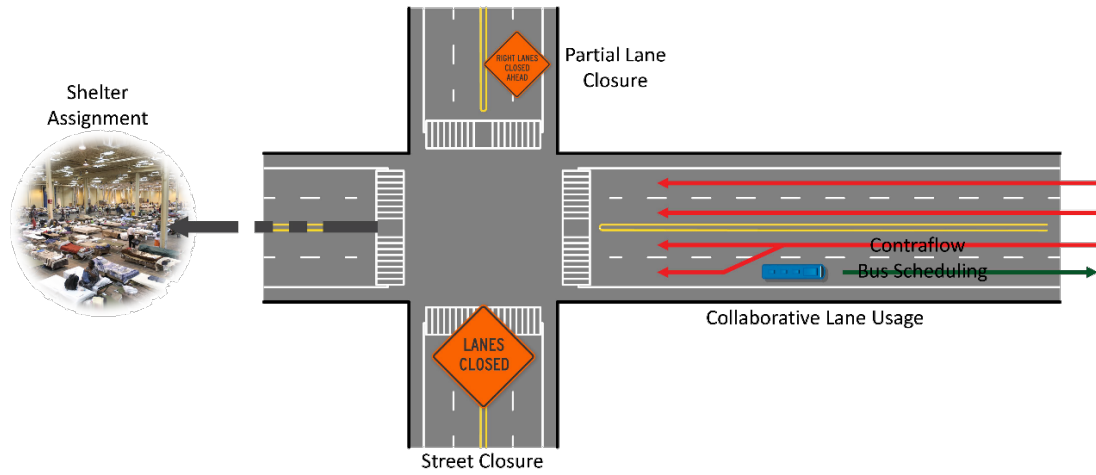


Figure 40. Different strategies that can be used to achieve a realistic bimodal evacuation problem

6.2.3. Highly non-linear nature of the evacuation problems

Traffic conditions in evacuation situations are highly volatile, and the flow within the road network can easily pass the links' capacities. Also, in the case of bimodal evacuation, dynamic scheduling of private and public transportation facilities escalates the issue. Figure 41 schematically shows the structure of the bimodal evacuation problem for census tract level flood emergency alerts.

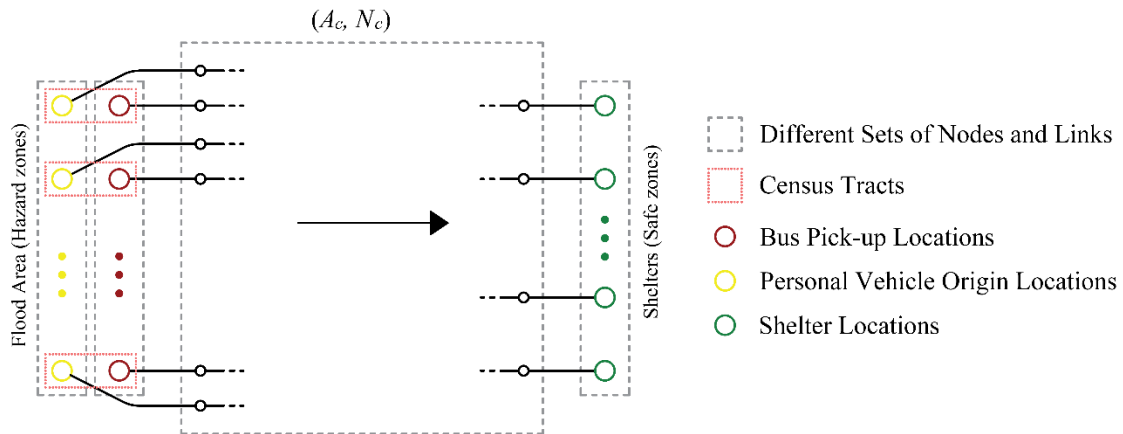


Figure 41. the structure of the bimodal evacuation problem for census tract level flood emergency alerts

A single pick-up location is used for each census tract. A single central node is assigned to each census tract as the origin of travel intended using personal cars. These nodes are connected to a road network in hazard zones and finally attached to shelter nodes located in the safe area. The time is discretized into small enough steps to represent travel times within links with the shortest distance and lowest travel time. Here, we formulate the scheduling problem for this bimodal problem to show the highly non-linear nature of the realistic evacuation operation. Table 8 summarizes the parameters, decision variables, and notations used to formulate the non-linear problem.

Table 8. The parameters, decision variables, and notations used to formulate the non-linear problem

Decision Variables	
λ_i^*	Departure time for zone i
$T_{i,j,k}^{\lambda_i^*, vt}$	Travel time by vehicle time vt from origin i to shelter j using route k , departure at time λ_i^*
$V_{i,j,k}^{\lambda_i^*}$	Personal vehicle flows from origin i to shelter j using route k , departure at time λ_i^*
$B_{i,j,k}^{\lambda_i^*}$	Public vehicle flows from origin i to shelter j using route k , departure at time λ_i^*
x_a^t	The flow at link a at time step t
xi_a^t	The flow entering link a at time step t
xo_a^t	The flow exiting link a at time step t
O_j^t	The vehicles arrive at node j and exit the network
$\delta_{i,j,k}^{\lambda_i^*, t, a, v}$	Indicates if the vehicle is in link a or not
Parameters	
D_i^v	Personal vehicle demand in node i
D_i^b	Public vehicle demand in node i
C_j	The capacity of shelter j
Δ	Length of time step t
Other Notations	
t	Time step
I_v	Set of origins (nodes) for personal vehicles
I_b	Set of origins (nodes) for public transport vehicles
$G(n)$	Set of links entering to node n
$E(n)$	Set of links exiting to node n

The problem is formulated without any lane management strategies. It only provides a staging system in which each vehicle is scheduled to depart at a specific time to minimize the total evacuation time. The objective functions and constraints are as follows:

$$\min \text{TET} = \sum_{\lambda_i^*=1}^T \sum_{i \in I_v} \sum_{j \in J} \sum_{k=1}^K (\lambda_i^* + T_{i,j,k}^{\lambda_i^*,v}) V_{i,j,k}^{\lambda_i^*} + \sum_{\lambda_i^*=1}^T \sum_{i \in I_b} \sum_{j \in J} \sum_{k=1}^K (\lambda_i^* + T_{i,j,k}^{\lambda_i^*,b}) B_{i,j,k}^{\lambda_i^*} \quad (29 \text{ a})$$

$$\sum_{\lambda_i^*=1}^T \sum_{j \in J} \sum_{k=1}^K V_{i,j,k}^{\lambda_i^*} = D_i^v \quad \forall i \in I_v \quad (29 \text{ b})$$

$$\sum_{\lambda_i^*=1}^T \sum_{j \in J} \sum_{k=1}^K B_{i,j,k}^{\lambda_i^*} = D_i^b \quad \forall i \in I_b \quad (29 \text{ c})$$

$$\sum_{\lambda_i^*=1}^T \sum_{i \in I_v} \sum_{k=1}^K V_{i,j,k}^{\lambda_i^*} + \sum_{\lambda_i^*=1}^T \sum_{i \in I_b} \sum_{k=1}^K B_{i,j,k}^{\lambda_i^*} = \sum_{t=1}^{\Theta} O_j^t \quad \forall j \in J \quad (29 \text{ d})$$

$$\sum_{\lambda_i^*=1}^T \sum_{i \in I_v} \sum_{k=1}^K V_{i,j,k}^{\lambda_i^*} + \sum_{\lambda_i^*=1}^T \sum_{i \in I_b} \sum_{k=1}^K B_{i,j,k}^{\lambda_i^*} \leq C_j \quad \forall j \in J \quad (29 \text{ e})$$

$$x_a^t = \sum_{\lambda_i^*=1}^T \sum_{i \in I_v} \sum_{j=1}^J \sum_{k=1}^K (V_{i,j,k}^{\lambda_i^*} \times \delta_{i,j,k}^{\lambda_i^*,t,a,v}) + \sum_{\lambda_i^*=1}^T \sum_{i \in I_b} \sum_{j=1}^J \sum_{k=1}^K (B_{i,j,k}^{\lambda_i^*} \times \delta_{i,j,k}^{\lambda_i^*,t,a,b}) \quad \forall t, a \quad (29 \text{ f})$$

$$x_a^t = x_a^{t-1} + xi_a^{t-1} - xo_a^{t-1} \quad \forall t, a \quad (29 \text{ g})$$

$$\sum_{c \in G(n)} xi_c^t + O_n^t = \sum_{e \in E(n)} xo_e^t \quad \forall t, n: \text{nodes in the network} \quad (29 \text{ h})$$

$$\delta_{i,j,k}^{\lambda_i^*,t,a,v} = f(V_{i,j,k}^{\lambda_i^*}) \quad \forall i \in I_v, j, \lambda_i^*, t, a, k \quad (29 \text{ i})$$

$$\delta_{i,j,k}^{\lambda_i^*,t,a,b} = f(B_{i,j,k}^{\lambda_i^*}) \quad \forall i \in I_b, j, \lambda_i^*, t, a, k \quad (29 \text{ j})$$

$$T_{i,j,k}^{\lambda,v} = \sum_{t=1}^{\Theta} \sum_a \delta_{i,j,k}^{\lambda,t,a,v} \times \Delta \quad \forall i \in I_v, j, \lambda, a, k \quad (29 \text{ k})$$

$$T_{i,j,k}^{\lambda,b} = \sum_{t=1}^{\Theta} \sum_a \delta_{i,j,k}^{\lambda,t,a,b} \times \Delta \quad \forall i \in I_v, j, \lambda, a, k \quad (29 \text{ l})$$

$$xi_a^t, xo_a^t = f(V_{i,j,k}^{\lambda_i^*}, B_{i,j,k}^{\lambda_i^*}) \quad \forall t, a \quad (29 \text{ m})$$

$$O_j^t = f(V_{i,j,k}^{\lambda_i^*}, B_{i,j,k}^{\lambda_i^*}) \quad \forall t, j \quad (29 \text{ n})$$

$$\delta_{i,j,k}^{\lambda,t,a,v}, \delta_{i,j,k}^{\lambda,t,a,b} = 0 \text{ or } 1 \quad \forall i, j, k, \lambda, t, a \quad (29 \text{ o})$$

$$\text{All variables} \geq 0 \quad (29 \text{ p})$$

Equation (29 a) is the objective function that aims to minimize the demand departure time as well as the travel time for flow in all routes from each census tract to the shelters. Equations (29 b – 29 e) satisfy the demand and supply within the census tracts and shelters, respectively. Equation 29 f defines the flow in each link at time-step t . Equations 29 g and h describe the flow conservation within the network nodes and links. Equations 29 i-j defines a decision variable that indicates whether a vehicle is in link a of route k from census tract i to census tract j as a function of existing

temporal demands. The travel time is, in turn, a function of this indicator (Equations 29 k-l). Equation 29 m indicates the dependency of temporal demand and the flows that enter and exit a node. These dependencies clearly indicate the non-linearity inherent to the evacuation planning problem, especially when advised intelligent scheduling is of interest to decision-makers. The mathematical optimization formulation described above cannot be solved for even small networks, considering the available machines' computational limitations. In the next section, we propose a formulation in which the capacity of each link in the network is restricted to an acceptable range, and the specifics of our small area evacuation problem have been embedded within a few linear constraints.

6.3. Proposed evacuation model

6.3.1. Formulating the problem

Figure 42 schematically illustrates the network used to formulate a mixed-integer linear programming model for census tract level evacuation for flood instances. Each census tract includes two nodes that are served as origins for bus and personal vehicles. Two auxiliary super nodes are assumed, which can serve as a single super origin node (D) and a single super sink node (S). Super origin node (D) is connected to all origins for bus and personal vehicles in each census tract through a set of artificial connectors (A_D). This node can serve as a depot or just as an auxiliary point that provides the required resources for bus pick-up points ($z_{h,b}$) as well as the demand for personal vehicle origins ($z_{h,p}$). The travel times for links in A_D can be set to zero, or it can represent the travel times for buses to arrive at the pickup locations. Table 9 summarizes the parameters and decision variables used to formulate the problem.

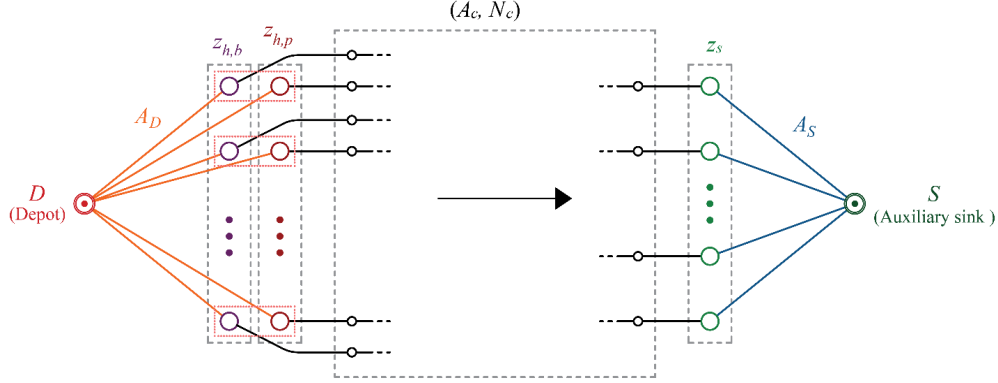


Figure 42. the structure of the network redesigned for the proposed evacuation model

Table 9. The parameters, decision variables, and notations used to formulate the TET MILP

Decision Variables	
y_i	Determines whether shelter i is open or not
$x_{v(i,j)}^t$	Number of vehicles originating their travel at node i at time step t towards node j
Parameters	
D_i^p	Personal vehicle demand in node i (public vehicle origin)
D_i^b	Public vehicle demand in node i (bus pick up point)
u_i	The bed capacity of the shelter i
U_i	The parking capacity of the shelter i
t	High-level time-step to represent flow at each link
λ	Low-level time-step to represent departure times
ω_j	The weight assigned to census tract j to prioritize its evacuation
$\bar{\omega}_v$	The weight that prioritizes the arriving time based on the vehicle type
$\bar{\delta}_{Dj}^{n_t \lambda}$	A parameter that allows modeling the waiting time at bus pick up points
δ_{ij}^t	Travel time from node i to node j at time-step t
ω_i^s	This variable allows modeling the shelter preparation costs
$\bar{C}_v$	Average occupancy for vehicle type v
Γ_j	Service rate for bus pick up location j
R_{eq}	Bus to car equivalency ratio factor
κ_{ij}	Flow capacity of link ij
Other Notations	
$v \in \{p, b\}$	Vehicle type set, which includes personal vehicles (p) and bus (b) transportation system
D	The Auxiliary super demand node
S	The auxiliary super Sink node
θ	Set of all time steps at a high level
σ	Set of time-steps in which node j of a link ij is accessible for a vehicle at time step t
n_t	Number of smaller time-steps within departure time-steps
τ	Set of time-steps in which a vehicle can stay in link ij if entered the link at t
$A \in \{A_c, A_D, A_S\}$	Set of network links, including network and auxiliary links.
N_c	Network nodes
$z \in \{z_{h,b}, z_{h,p}\}$	Set of hazard zones, including bus pick-up locations and personal vehicle origins
z_s	Set of shelter nodes

It is assumed that the buses are not allowed to pick up any evacuees in nodes other than the census tract centroids selected for passenger pick-ups. A bus can serve only one census tract for each trip from a census tract to the super sink node. On the supply side, a set of nodes (z_s) serve as available shelters. These nodes are connected to the super sink, which serves as a node that simplifies the formulation of the TET objective function. The Demand ($z_{h,b}$, $z_{h,p}$, and D) and supply nodes (z_s and S) are connected through a graph $G(A_c, N_c)$ which is the original highway network, where A_c and N_c are the links and nodes. The total evacuation demand is calculated from the following equations:

$$D_i^p = \frac{P_i - P_{ib}}{4 * SF} \quad (30)$$

$$D_i^b = \frac{P_{ib}}{40 \sim 80} \quad (31)$$

where D_i^p and D_i^b are the demand for node i for personal cars and buses, respectively, and P_i is the population of the census tract i . P_{ib} and P_{ip} are the population who do not have access to personal cars and the population who own a personal car, respectively. These two values are obtained from the proposed GANs model. Equation 30 assumes that four people occupy each personal car, and a safety factor (SF) is used to modify this value based on available data for each geographical area. SF takes a value less than or equal to 1. The time has been discretized into two different time steps to evacuate the demand dynamically over time. At a high level, the smallest time-unit, t , has been used to indicate the travel time and flow status for each link and node at each time step. A practical choice for each time-step length must be smaller than the travel time for the shortest link with the highest free-flow speed. Θ represents the set of all time-steps discretized at a high level. At the lower level, the departure time (λ) consists of T number of steps; each step includes n_t number of smallest time unit t . In addition, the departure time unit is chosen to be

evenly devisable by t . Figure 43 illustrates the concept of time discretization that is mentioned above and demonstrates the notations used.

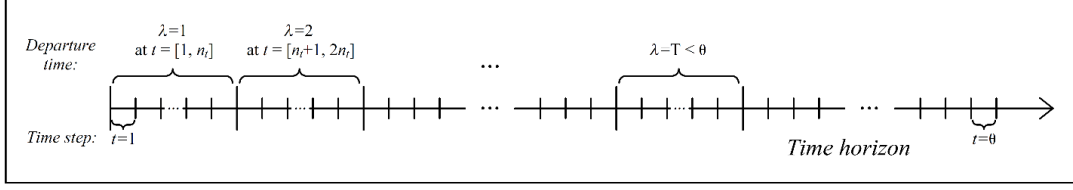


Figure 43. Bilevel discretization of the time

The objective function for the TET MILP optimization problem is stated in equation 32.

$$\min TET = \sum_{v \in \{p, b\}} \sum_{j \in Z_{h,v}: (D,j) \in A_D} \sum_{\lambda} \omega_j(n_t \lambda) x_{v(D,j)}^{n_t \lambda - \bar{\delta}_{Dj}^{n_t \lambda}} + \sum_{v \in \{p, b\}} \sum_{t \in \Theta} \sum_{i \in Z_s} \omega_v t x_{v(i,S)}^t + \sum_{i \in Z_s} \omega_i^s y_i \quad (32)$$

This objective function includes three terms. The first term intends to minimize the weighted departure time for both cars and buses. $x_{v(D,j)}^t$ is a decision variable that determines the number of vehicle type (v) that leave super origin D at the time $n_t \lambda - \bar{\delta}_{Dj}^{n_t \lambda}$. The term $\bar{\delta}_{Dj}^{n_t \lambda}$ can be used to model the travel time from the actual depots or the bus boarding time at each dedicated zone. ω_j is the weight assigned to census tract j to prioritize its evacuation if a higher hazard risk threatens the area. The second term minimizes the total travel time considering all the individuals ($x_{v(i,S)}^t$) who leave shelters toward the super sink at all time steps. ω_v is a weight that prioritizes the arriving time based on the vehicle type. Finally, the third term minimizes the total number of weighted shelters ($\omega_i^s y_i$) that have been used to accommodate the evacuees. The problem constraints are formulated as follows.

$$\sum_{\substack{i \in N: \\ (i,j) \in A; \\ t - \sigma = \delta_{ij}^\sigma}} x_{v(i,j)}^\sigma = \sum_{\substack{k \in N: \\ (j,k) \in A}} x_{v(j,k)}^t \quad \forall v \in \{p, b\} \& j \in N \setminus \{D, S\} \& \{\sigma \in [1, t) \mid t - \sigma = \delta_{ij}^\sigma\} \quad (33)$$

$$\begin{aligned} & Req x_{b(i,j)}^t + x_{p(i,j)}^t \leq \kappa_{ij}^t \\ \rightarrow & Req x_{b(i,j)}^t + x_{p(i,j)}^t \leq \kappa_{ij} - \sum_{\tau} (Req x_{b(i,j)}^\tau + x_{p(i,j)}^\tau) \quad \forall (i,j) \\ & \in A, t, \{\tau \in [1, t) \mid \delta_{ij}^\tau - (t - \tau) > 0\} \end{aligned} \quad (34)$$

$$\sum_{t \in \lambda} x_{p(D,j)}^t = D_j^p \quad \forall j \in z_{h,p} \quad (35)$$

$$\sum_{t \in \lambda} x_{b(D,j)}^t = D_j^b \quad \forall j \in z_{h,b} \quad (36)$$

$$\sum_{v \in \{p,b\}} \sum_{t \in \Theta} \sum_{j \in z_h} x_{v(D,j)}^t = \sum_{v \in \{p,b\}} \sum_{j \in z_h} D_j^v \quad (37)$$

$$\sum_{t \in \Theta} \sum_{i \in z_s} x_{v(i,S)}^t = \sum_{j \in z_{h,v}} D_j^v \quad \forall j \in z_s, v \in \{p,b\} \quad (38)$$

$$\sum_{v \in \{p,b\}} \sum_{t \in \Theta} \bar{C}_v x_{v(i,S)}^t \leq u_i y_i \quad \forall i \in z_s \quad (39)$$

$$\sum_{t \in \Theta} x_{p(i,S)}^t \leq U_i y_i \quad \forall i \in z_s \quad (40)$$

$$x_{b(D,j)}^t \leq \Gamma_j \quad \forall t \in \lambda \text{ and } j \in z_{h,b} \quad (41)$$

$$\text{All variables} \geq 0 \quad (42)$$

Equation 33 is a constraint that preserves the conservation of flow at node j . The dynamic nature of the conservation of flow constraint is illustrated in figure 44. σ in equation 33 is the time-step in which a vehicle of type v can traverse the link (i, j) with travel time δ_{ij}^v and arrive at node j at time-step t . In other words, only the links with travel time equal to the difference between the time-steps of nodes i and j appear on the left-hand side of equation 33. Since the right-hand side of the equation includes all the links that start at node j at time-step t , the entry to j is limited to the links that satisfy the travel time limitation.

Equation 34 limits private and public transportation flows at each link to its capacity at its time-step (κ_{ij}^t). The capacity of a link at each time step is defined as the total capacity of the link minus the vehicles that have already entered and are traveling on the link. A vehicle is still traveling on the link if the difference between its entry time step and the current time step is less than the travel time at the entry time step. R_{eq} is a bus to car equivalency ratio factor. Equations 35 and 36 match the demand with the total flow that exits the auxiliary depot summed over the departure

times. Equation 37 prevents any vehicle departures that might be allocated to the time-steps other than departure time-steps. Equation 38 matches the supply in shelters with the total demand. Equations 39 and 40 make sure that the bed capacity of each shelter i (u_i) is higher than the arriving demand occupied by average $\bar{C}_v$ people for vehicle type v , and that the parking capacity (U_i) of each shelter i is respected by arriving personal vehicles, respectively. Finally, equation 41 limits the number of buses that can be served at each time-step t for each pickup point to Γ_j .

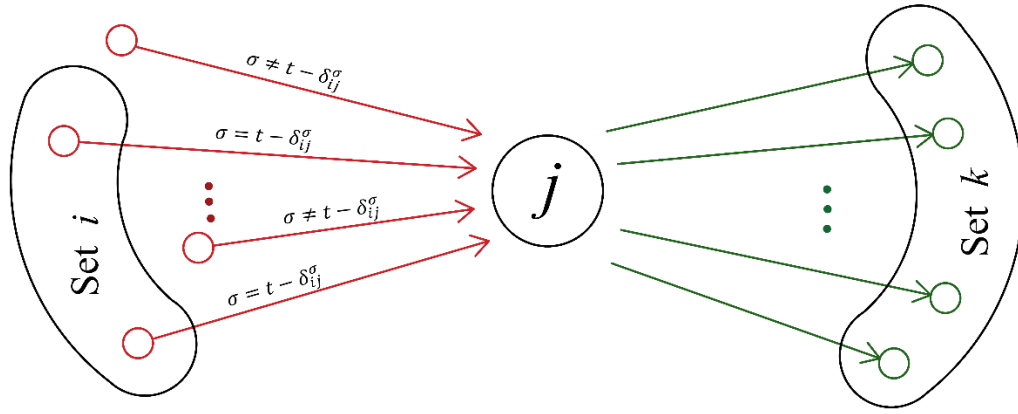


Figure 44. Schematic representation of the conservation of flow at node j

6.3.2. Application of TET MILP on a simple network

A simple toy network (Figure 45) has been designed to show the application of the formulated TET MILP problem. This network consists of 4 real links and two auxiliary links that connect a zone with one bus pick-up point and car origin to an auxiliary super demand node. There are two shelters in the network that are connected to a super sink node through two auxiliary links with zero travel time. All four auxiliary links have a capacity of 100.

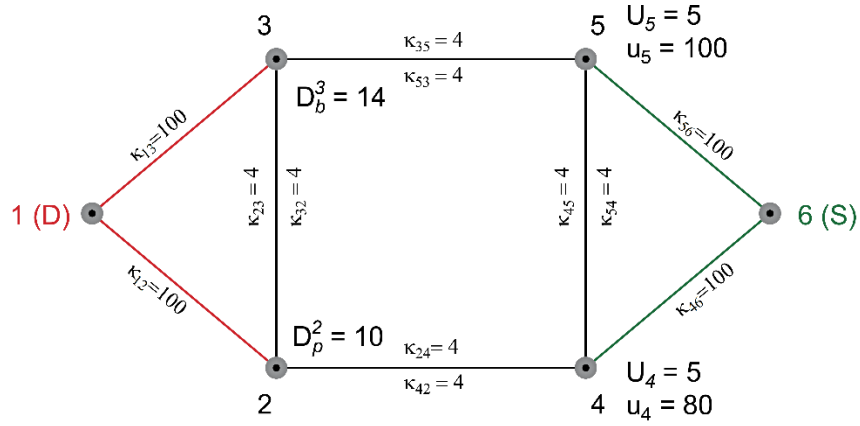


Figure 45. A simple example network designed to examine the proposed formulation

The maximum time step (T) is ten units. The vehicles can depart the shelters every two steps, allowing departures at five time steps, each being two time-unit ($n_t=2$). A parking capacity of 5 cars is considered for both shelters. The shelters' bed capacities at nodes 4 and 5 are set to 80 and 100, respectively. The maximum link capacity is 4 for all links. The travel time matrix for each link at each time step is shown in table 10. Also, the parameters used in this specific example are presented in table 11.

Table 10. The travel time matrix for each link at each time step

Link Travel Times		Time Steps									
		t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10
Links	1-2	0	0	0	0	0	0	0	0	0	0
	3-2	1	2	3	1	1	1	1	1	1	1
	4-2	3	2	4	1	1	1	1	1	1	1
	1-3	0	0	0	0	0	0	0	0	0	0
	2-3	3	3	2	1	1	1	1	1	1	1
	5-3	3	3	2	1	1	1	1	1	1	1
	2-4	2	2	3	1	1	1	1	1	1	1
	5-4	3	2	2	1	1	1	1	1	1	1
	3-5	3	3	4	1	1	1	1	1	1	1
	4-5	3	3	3	1	1	1	1	1	1	1
	4-6	0	0	0	0	0	0	0	0	0	0
	5-6	0	0	0	0	0	0	0	0	0	0

Table 11. The parameters used for the example network evacuation problem

perimeters	Value
D_i^p	Illustrated in figure 29
D_i^b	Illustrated in figure 29
u_i	Illustrated in figure 29
U_i	Illustrated in figure 29
T	10 (smallest time unit)
n_t	2
λ	t = [1, 3, 5, 7]
ω_j	0.4 for node 2 and 0.6 for node 3
$\bar{\omega}_v$	0.9 for buses and 0.8 for personal vehicles
$\bar{\delta}_{Dj}^{n_t\lambda}$	0
δ_{ij}^t	See table 7
ω_i^s	5 (just to make the model work)
$\bar{C}_v$	10 for buses and 4 for personal vehicles
Γ_j	A high value in this simple example
R_{eq}	1.2
κ_{ij}	Illustrated in figure 29

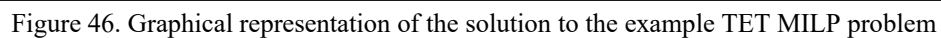
At each time step, the link travel times should be updated iteratively to adjust for volume. This requires a proper definition of the volume-travel time functions. For this study's purpose, the problem is solved for the first iteration with predefined travel times, and the effect of updating the travel time will be investigated in the future. The solution to this problem is summarized in table 12. The entering flow into each link at each time step is represented in figure 46.

The solution's reasonableness is investigated for the first four consecutive time steps. Node 2 is the personal vehicle origin, and node 3 is the bus pick-up point. At time step 1 (t=1), four personal vehicles and six buses start their travel at node 1. Since the travel times in auxiliary links are equal to zero, the originating journeys immediately appear at the subsequent links. For example, the four personal vehicles that start their journey at node 1 immediately appear at link 2-3, and after three consecutive time steps (travel time for link 2-3 at time t=1), enter link 3-5, and after one time step, it arrives at shelter 5.

Table 12. The solution to the example TET MILP problem

Personal Vehicles		Links											
		1-2	3-2	4-2	1-3	2-3	5-3	2-4	5-4	3-5	4-5	4-6	5-6
Time-step	t=1	4	0	0	0	4	0	0	0	0	0	0	0
	t=2	0	0	0	0	0	0	0	0	0	0	0	0
	t=3	0	0	0	0	0	0	0	0	0	0	0	0
	t=4	0	0	0	0	0	0	0	0	4	0	0	0
	t=5	6	0	0	0	2	0	4	1	0	0	0	3
	t=6	0	0	0	0	0	0	0	0	2	0	5	0
	t=7	0	0	0	0	0	0	0	0	0	0	0	2
	t=8	0	0	0	0	0	0	0	0	0	0	0	0
	t=9	0	0	0	0	0	0	0	0	0	0	0	0
	t=10	0	0	0	0	0	0	0	0	0	0	0	0
Buses		Links											
		1-2	3-2	4-2	1-3	2-3	5-3	2-4	5-4	3-5	4-5	4-6	5-6
Time-step	t=1	0	3	0	6	0	0	0	0	3	0	0	0
	t=2	0	0	0	0	0	0	3	0	0	0	0	0
	t=3	0	3	0	3	0	0	0	0	0	0	0	0
	t=4	0	0	0	0	0	0	0	0	0	0	3	3
	t=5	0	0	0	3	0	0	0	0	3	0	0	0
	t=6	0	0	0	0	0	0	3	0	0	0	0	3
	t=7	0	0	0	2	0	0	0	0	2	0	3	0
	t=8	0	0	0	0	0	0	0	0	0	0	0	2
	t=9	0	0	0	0	0	0	0	0	0	0	0	0
	t=10	0	0	0	0	0	0	0	0	0	0	0	0

The first question is why these four public vehicles do not use link 2-4 to arrive at shelter 4 at the end of time step 2 (travel time for link 2-4 at time $t=1$). To justify this evidence, note that the weight of evacuation for bus pick-up is 0.6 while that of node 2, which is the start point for personal vehicles, is 0.4 (See table 11). Therefore, the evacuation of buses at node 3 has priority over personal vehicles' evacuation at node 2. Six buses start their journey at time step 1 at node 3. Three of the buses use link 3-2 with travel time equal to 1. Since each link's maximum capacity is 4, if the four personal vehicles choose to travel on link 2-4, these three buses could not enter link 2-4 at the beginning of time step 2. Also, each shelter's capacity is limited, and fulfillment of the shelter capacity constraints depends on the dynamics of the travels over all time steps.



93

Chapter 7. A Heuristic Solution to the TET MILP Evacuation Model

7.1. TET MILP limitations and the proposed heuristic solution's strengths

The proposed TET MILP model does not include wait-on-link properties. This limitation imposes the vehicles to leave the second node of the link as soon as they use the allocated travel time for that link. In addition, the model uses predefined travel times. This travel time could be set to travel time associated with the at-capacity operation of the links, but this unnecessarily underestimates the capacity of the links, especially at the first stages of the evacuation operation. The capacity of the parallel routes is underrated as well. A possible solution to this problem is to solve the proposed MILP iteratively to consider travel time updates. However, it is highly probable that the travel times hover between the extreme values and prevent the model from converging. This problem can be overcome by gradually increasing the effect of travel times from previous iterations so that a considerable change in travel time does not influence the volumes on parallel routes. These solutions cannot guarantee the convergence of the model. On top of all the mentioned issues, the TET MILP problem cannot be solved for any real-case problem that considers evacuation at the state or city level unless a substantial link reduction and network analysis have been performed to achieve a suitable size for the link numbers in the network. The dynamic nature of the problem, which is implemented in equation 34, further escalates this issue. Therefore, a heuristic solution to the problem is developed in this section. This heuristic solution is based on the Genetic Algorithm (GA) and its multi-objective variant, i.e., Non-Dominated Sorting Genetic Algorithm II (NSGA-II). The main difference between the proposed algorithm and the previous GA-based algorithms is that the proposed framework can adapt to the dynamic nature of the problem and flexibility of leave orders. For example, the comprehensive evacuation model proposed by Goerigk et al. (2014) does not consider the dynamic travel time updates and assumes

that the operation is under peak volume pressure. This model would not allow the operators to control the highest capacity they would like to use. Also, their model assumes that the number of vehicles that will leave at each time is known, which is not realistic. We assume that it is impossible to estimate the vehicles' leaving times, and therefore, authorities must make the leave orders. The details of the assumptions will be mentioned later.

7.2. Algorithm design and assumptions

Since the proposed algorithm was first used on the previously mentioned simple toy network example, the heuristic algorithm will be mentioned on the same network. In the proposed algorithm, auxiliary demand links are no longer helpful and are removed from the network arcs. The network, for a simple example, is shown in figure 47.

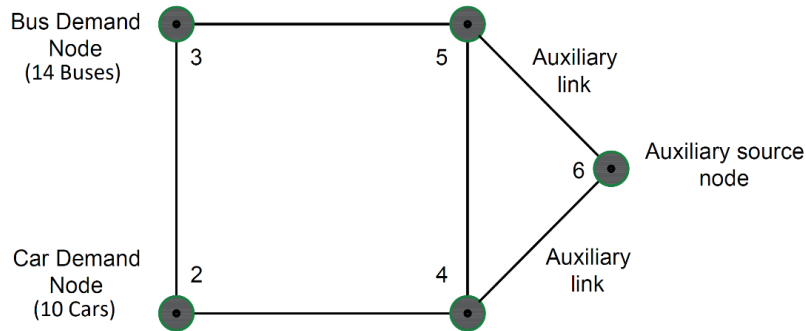


Figure 47. Simple network without an auxiliary source link used in developing the heuristic algorithm

The general idea is to develop an algorithm that heuristically allocates a path to predefined bus and personal-vehicle packets willing to initiate their travel at time zero to evacuate the risk area. Each path is constructed by links that connect the demand node to the auxiliary source node. Therefore, a set of predefined routes must be obtained before initiating the algorithm. We used three mechanisms to generate these paths based on the k simple shortest path algorithm Sedgewick (2001) suggested and later implemented in the NetworkX python library. A simple shortest path does not contain any loop, and each node appears just once. NetworkX package in Python is used

to calculate k shortest paths from each demand node to the auxiliary source node. First, for each shelter, simply k_1 number of shortest paths (weighted by free-flow travel time and distance separately) are obtained from the demand node to the shelter node and preserved in a population set A . Second, starting from the shortest path in A , the routes are deleted iteratively, and the k_2 number of parallel routes to each deleted route is identified. We kept the ramps and necessary connectors within the network. These parallel routes are added to set A . Finally, starting from the shortest path in A , the routes are loaded by vehicles, and new shortest paths (weighted by travel time) are obtained from a partially loaded network and added to set A . These shortest paths are preserved in a python dictionary and are randomly assigned to each packet willing to leave the demand node at time zero.

For the case of the simple example shown in figure 47, there are 10 and 14 personal vehicles and buses, and they need to be evacuated from demand nodes 2 and 3. For the sake of simplicity in this example, we assumed that each packet consists of a single vehicle, regardless of the vehicle type. Therefore, 24 packets must be evacuated from the hazard zones (node2 and node3). The initial population for the genetic algorithm contains n genotypes. Any of the 24 genes within each genotype preserves the information about the path randomly drawn from set A and reserved for the packet. The information for each gene is stored in a python dictionary and listed in a tuple. Figure 48 schematically shows a genotype representing a bus packet that leaves demand node 3 at time 0.

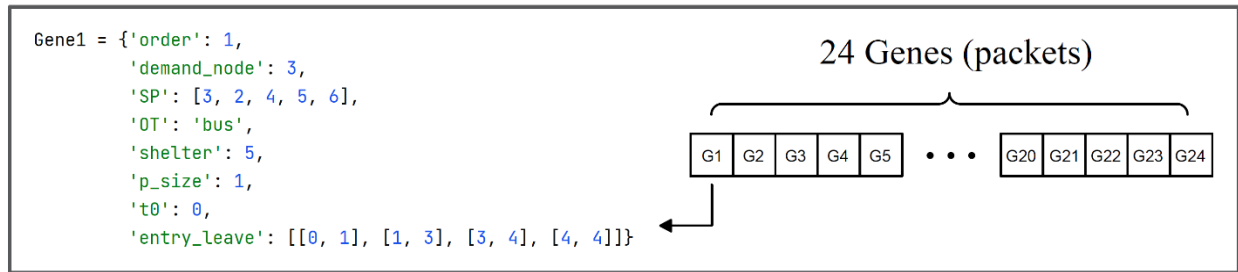


Figure 48. A typical Genotype and the information stored in each gene

The "entry_leave" attribute for each gene represents the pairs of entry time and leave time (in time step units) for each link along the route. Link travel times for each newly generated gene are calculated assuming there are no traffic or capacity restrictions for any link along the route. These travel times are temporary and calculated based on free-flow speeds. After calculating initial free-flow-based travel times, the genotype is sorted from longest to shortest. The packet assignment for the first gene in the sorted genotype is based on the free flow speed. The travel times for the rest of the genes must be updated based on the updated capacity and timetable obtained from the previous gene assignment. For example, once the first packet is assigned to the first gene, the capacity and travel time are updated. Two tables must be updated after packet assignment for each gene: capacity and travel time. Figure 49 illustrates these tables for the toy network example shown in figure 47.

from	to	C1	C2	C3	C4	C5	C6	C7	C8	...
5	6	100	100	100	100	100	100	100	100	...
4	6	100	100	100	100	100	100	100	100	...
3	2	4	4	4	4	4	4	4	4	...
3	5	4	4	4	4	4	4	4	4	...
2	4	4	4	4	4	4	4	4	4	...
4	5	4	4	4	4	4	4	4	4	...
2	3	4	4	4	4	4	4	4	4	...
5	3	4	4	4	4	4	4	4	4	...
4	2	4	4	4	4	4	4	4	4	...
5	4	4	4	4	4	4	4	4	4	...
from	to	T1	T2	T3	T4	T5	T6	T7	T8	...
5	6	0	0	0	0	0	0	0	0	...
4	6	0	0	0	0	0	0	0	0	...
3	2	1	2	3	1	1	1	1	1	...
3	5	3	3	4	1	1	1	1	1	...
2	4	2	2	3	1	1	1	1	1	...
4	5	3	3	3	1	1	1	1	1	...
2	3	3	3	2	1	1	1	1	1	...
5	3	3	3	2	1	1	1	1	1	...
4	2	3	2	4	1	1	1	1	1	...
5	4	3	2	2	1	1	1	1	1	...

Figure 49. Capacity and timetables for the simple network

The process of updating the capacity table and timetable consists of multiple steps shown in figure 50. The goal is to identify the entry-time and leave-time points for each link that emerges sequentially in a path assigned to a gene and update the capacity accordingly. Pairs of subsequent links are constructed, and each pair is considered separately for the assignment of a packet. The links are considered in the order that they appear in the path. For each pair of links, the algorithm scans the travel time in each time step and iteratively looks for a range of time steps in the capacity table so that their capacity does not restrict the packets from staying in the links. For example, in figure 50, the first available time steps for a packet of size 4 for the link (i, j) are time steps 4 and 5. Therefore, in the first step, the entry to the first link is temporarily set to 4. In the second step, the algorithm scans the second link in the pair for the available time steps beyond time step 5, which are time steps 8, 9, and 10. Time-step 8 is temporarily reserved as an entry time-point for the packet, which is the leave time-point for the link (i, j) as well (step 3).

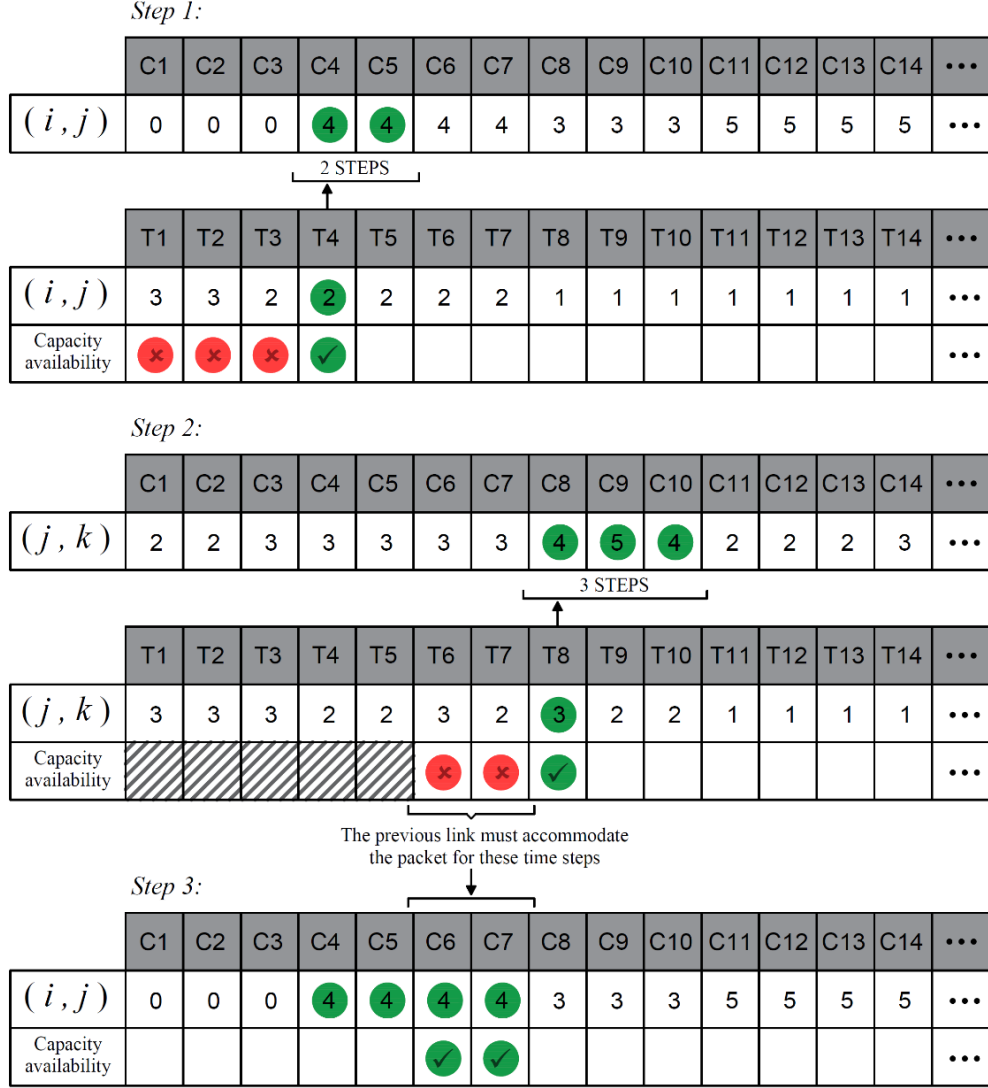


Figure 50. Steps to calculate the entry points for a link pair in a gene path (steps 1 and 2) and leave point for the first link in the link pair (step 3)

Finally, the algorithm checks whether the link (i, j) supports the entry point for the link (j, k) . Since the packet cannot enter the link (j, k) at time-steps 6 and 7, it must wait at the previous link (i, j) . This is the final sanity check for the path to see if the previous link's capacity allows a packet of size 4 to wait in the link. This criterion is satisfied in the example shown in figure 50. If this criterion were not met, the entire search process would start from one time-step after the previously found entry-point for the link (i, j) (time-step 5). This approach is repeated for all pairs of the links that appear along the assigned path. The entry time step for the auxiliary link in the

path is reported as the travel time for the gene. The volume on the links at each timestep is calculated from the remaining capacity of the links. The associated travel times are estimated and stored within the updated timetables. One can calculate the travel times from the BPR function or any function that correlates the volume to the travel time. The other genes in the genotype are iteratively updated based on the updated information from the previous gene assignment. The entire process is repeated for all the genotypes within the initial population.

The Genetic algorithm uses two primary procedures to evolve the initial population: Crossover and Mutation. Within the Crossover Step, a set of parents are randomly selected from the initial population to share a portion of their genes to create child genes. Our algorithm cuts the genes from two random indices in the genotype and mixes the genes from parents to generate six children out of the parent pairs. Figure 51 schematically shows this approach.

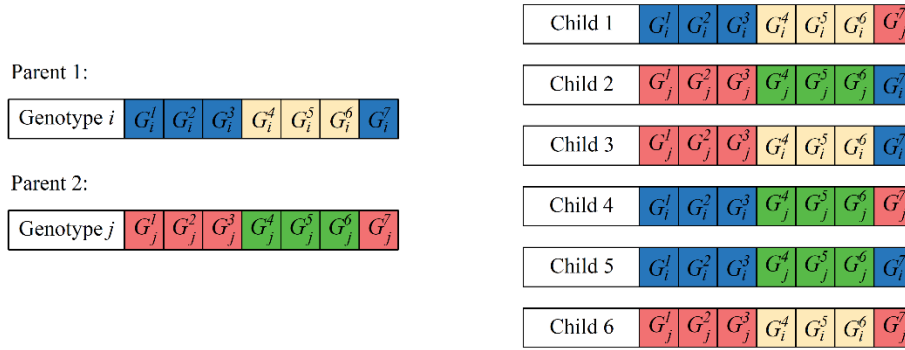


Figure 51. Crossover and obtaining six children out of two parents

The crossover procedure reassigns a new path to a predefined portion of the genes in the genotype. Note that once a new path is assigned to a gene, the travel time and capacity tables must be updated for the gene and all genes after it in the sorted genotype. The capacity of auxiliary links is set to the bed capacities of the shelters. Once a bed or a parking lot is assigned to a shelter, the algorithm reduces the auxiliary links' capacity for all time-steps. If the capacity of an auxiliary link becomes negative, the travel time increases to infinity, indicating a bad genotype. The negative

assignment is only allowed for the auxiliary links. A better approach to account for shelter capacities based on parking capacity is discussed in the following.

Assuming that the shelter capacity follows the parking capacity around the shelter, one can design a multi-objective optimization algorithm to simultaneously minimize the total evacuation time and overutilization of the shelters' parking lots. Overutilization is modeled using the Gini index concept. Figure 52 schematically demonstrates the Gini index concept.

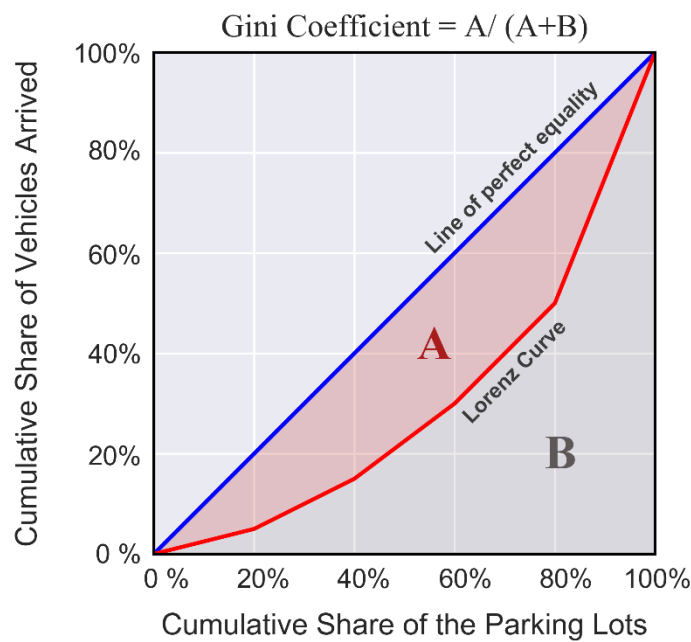


Figure 52. Conceptual demonstration of the Gini index concept

The perfect equality happens when none of the shelters overuse its capacity. Overuse of a shelter corresponds to attracting other shelters' share, leading to lower travel time. The curve that depicts the cumulative distribution of benefits against the cumulative distribution of resources is called Lorenz Curve. The smaller the area between the perfect equality and Lorenz curves, the lower the Gini index and inequality in the parking's share from arrived vehicles. Equation 43 expresses this concept in a mathematical form.

$$OU = \frac{\sum_i (\max[V_{ai}, C_i] - C_i)}{\sum_i C_i} \quad (43)$$

Multi-objective heuristic algorithms can generate a Pareto-Front for the problems. This allows the decision-makers to choose a solution from a list of the solutions. The minimization of the travel time can lead to overutilization of the shelters, and the reverse is true as well. If one objective dominates the other, the decision-maker can find the best trade-off between travel time and overutilization. In this study, we used Non-Dominated Sorting Genetic Algorithm II (NSGA-II) to solve the problem for the multiple solutions as a Pareto-Front. The readers can find the details of the NSGA-II in Deb et al., 2002. NSGA-II uses two mechanisms in identifying the Pareto fronts. Figure 53 shows these mechanisms for an example of minimization of a quadratic function with two objectives.

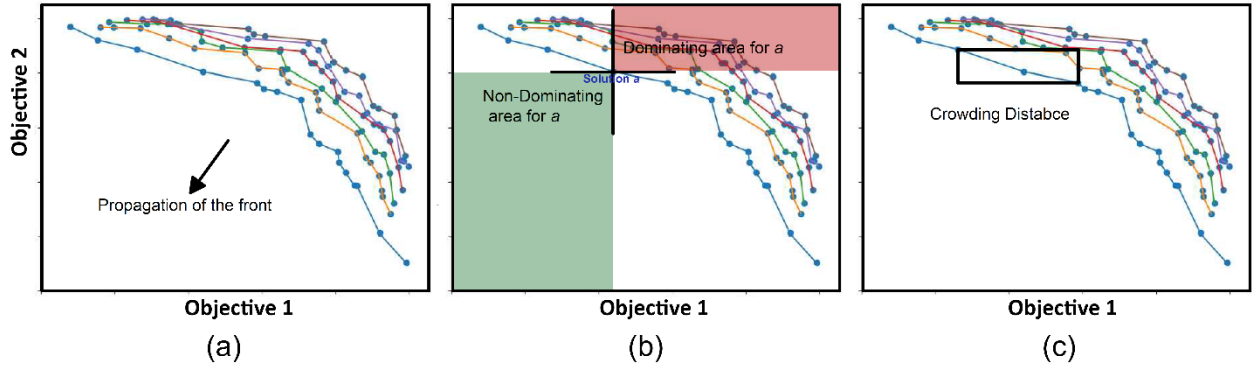


Figure 53. NSGA-II sorting Mechanisms for an example of minimization of a quadratic function with two objectives, (a) Pareto-Front propagation, (b) Non-dominated sorting, and (c) crowding distance

The first mechanism is the non-dominated sorting (figure 53-b). A solution x dominates all the solutions whose objective functions are evaluated to be worse than x . The solution dominating all other solutions belongs to the first Pareto-Front and is ranked 1. The second mechanism is the Crowding Distance (CD). Assuming that O_j^i is the objective function i for the solution j , the CD is calculated as follows:

$$CD_x = \sum_i \left| \frac{O_{x+1}^i - O_{x-1}^i}{O_{last\ in\ the\ front}^i - O_{first\ in\ the\ front}^i} \right| \quad (44)$$

A solution x with a higher CD is most suitable for the Pareto-Front and is more desirable to be kept in the selected solutions. Calculating the rank and CD for the population, the entire population is sorted based on the CD from highest to lowest first and then is sorted again based on the solutions' rank. The sorting order is critical as the rank is more important than the CD. Figure 54 illustrates the flowchart for the entire algorithm.

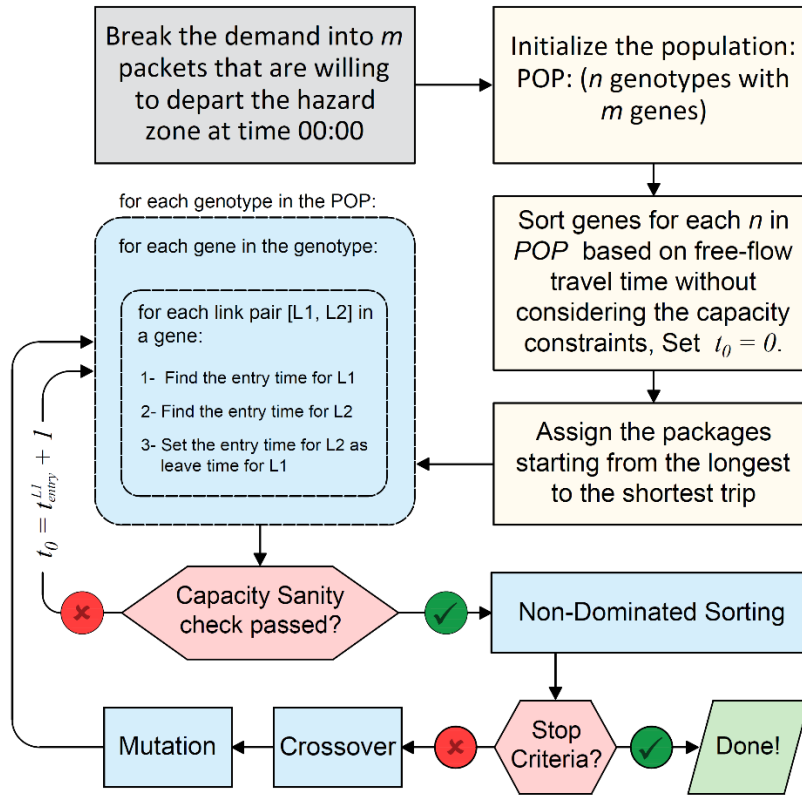


Figure 54. heuristic NSGA-II algorithm designed for the evacuation operation

7.3. Results of the heuristic algorithm

The heuristic solution procedure is first applied to the simple toy network, which is easier to explain and explore the results. In the first attempt, we did not consider the travel time update

to compare the algorithm's results with the exact solution from TET MILP. Figure 55 shows the algorithm results for very few iterations. We experienced that the algorithm quickly converges in the first iteration due to the low number of links in the network. Therefore, we used only five genotypes within the initial population to mimic the algorithm's behavior for extensive networks. The worst performance of the algorithm is reported in figure 55 (in terms of iteration numbers).

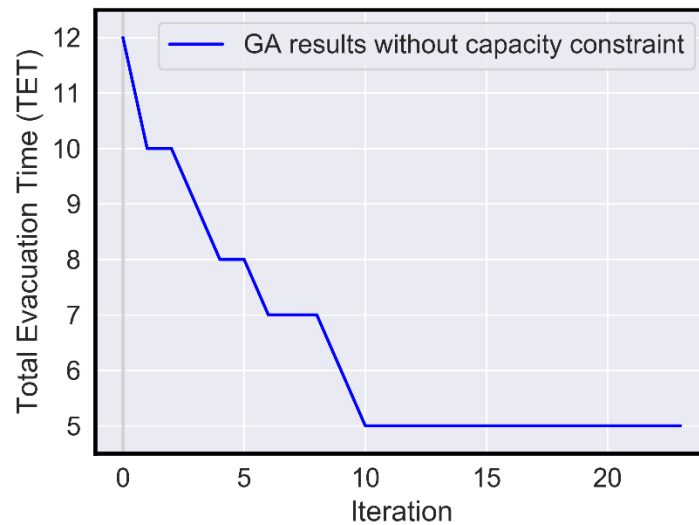


Figure 55. Uni-objective GA algorithm results for a simple example for very few iterations

The total travel time is lower than the travel time from the TET MILP. This observation is due to implementing the waiting-in-nodes in the heuristic algorithm, allowing shorter paths with the expense of waiting in link times. The second scenario is to implement the travel timetable updates. It is assumed that the travel time is a ceiling function of the volume on the links for the simple network. Figure 56 (a) shows the relationship between volume and travel time for the links. The travel time associated with free-flow speed is set to 1 for all links. Once the volume on the links increases by gene packet assignment, the travel time is updated and influences the next gene assignment. Figure 56 (b) shows the result of the algorithm with iterations.

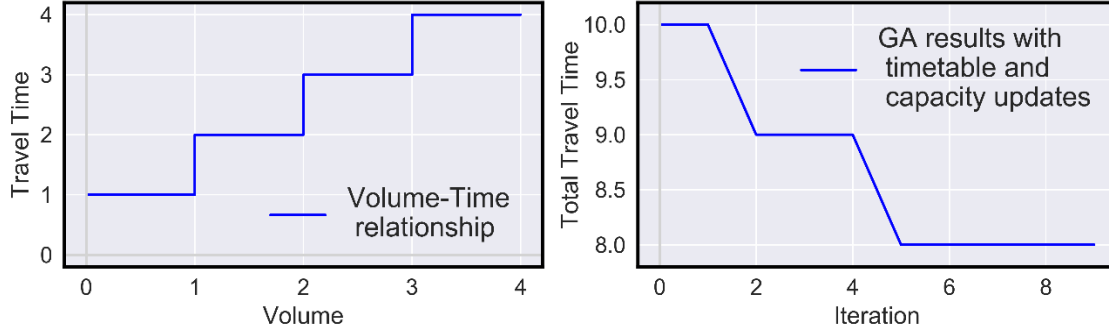


Figure 56. (a) Impedance calculation for the toy network example, and (b) results of GA on toy network implementing time and capacity updates

Table 13 shows the last updated timetable obtained from the best solution. Most of the links are loaded to their maximum capacity, which shows the model's success in simulating the actual evacuation operation.

Table 13. Final updated timetable obtained from the best solution for GA applied to the toy network example

from	to	T1	T2	T3	T4	T5	T6	T7	T8	T9
5	6	0	0	0	0	0	0	0	0	0
4	6	0	0	0	0	0	0	0	0	0
3	2	4	4	4	4	4	3	2	1	1
3	5	3	4	4	4	4	4	3	2	1
2	4	4	4	4	3	3	2	2	2	1
4	5	1	2	3	4	3	3	1	1	1
2	3	4	3	2	1	1	1	1	1	1
5	3	1	1	1	1	1	1	1	1	1
4	2	1	1	1	1	1	1	1	1	1
5	4	1	2	4	2	1	1	1	1	1

The NSGA-II algorithm is applied to this example to evaluate the algorithm's performance in minimizing travel time and overutilization of the shelters. This analysis is mainly performed because from the previous run, the optimum result with zero overutilization was 8. We expect to achieve an optimum value equal to or below this value within a few iterations, which should be the first Pareto front. Figure 57 shows the Pareto-Fronts obtained for seven iterations.

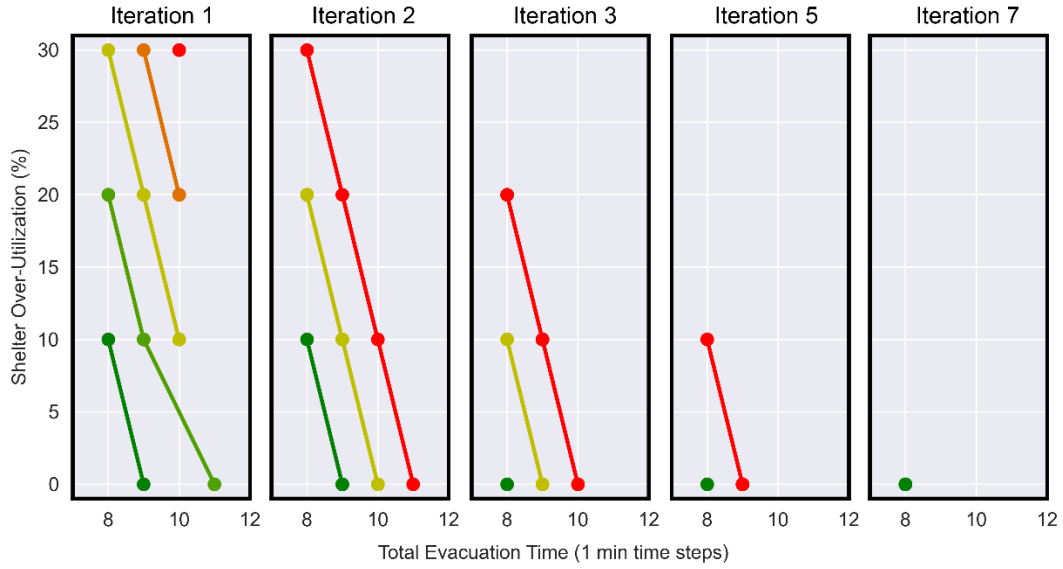


Figure 57. Pareto-Fronts obtained for seven iterations for Multi-variate NSGA-II applied to toy network implementing time and capacity updates

As figure 57 shows, the algorithm was able to identify the optimum solution at iteration 3. Since the shelter overutilization converged to zero at a travel time that is global optimum, all solutions in Pareto-Fronts converged to a single solution at iteration 7. The algorithm shows a healthy performance, and therefore it can be applied to a real case scenario where two census tracts located on the shoreline in Anne Arundel County, Maryland, must be evacuated. We assumed an overnight evacuation operation where the external travels could be ignored. Figure 58 shows the network and census tract boundaries for Anne Arundel County, the demand, shelters, auxiliary sink nodes, and auxiliary links.

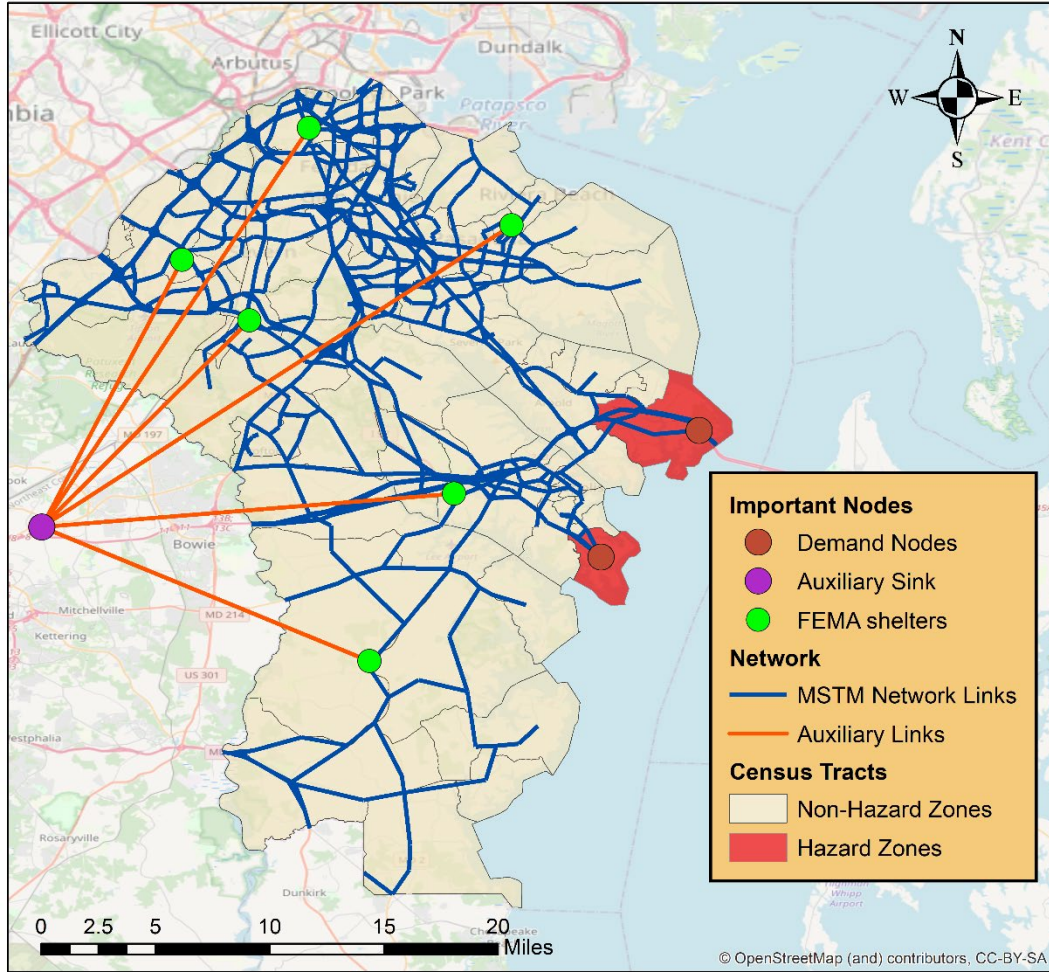


Figure 58. Network for Anne Arundel County (from the MSTM model) and demand, shelter, and auxiliary sink nodes

The main body of the road network is transferred from the Maryland Statewide Transportation Model (MSTM). The fraction of the MSTM network within Anne Arundel County boundaries is cropped. The node numbers and link properties are saved in a file, and the network is reconstructed in Python using the Networkx package. The shelters in Anne Arundel County are extracted from the Federal Emergency Management Agency website. The closest nodes to FEMA shelters are connected to shelters. The shelters are then connected to an auxiliary sink node outside Anne Arundel county. The distance and travel time for the auxiliary links are set to zero. The demand nodes (nodes 1 and 2) are assumed to be the same for bus and car vehicles and are located at the intersection of zone connector edges for each census tract. We iteratively deleted five of the

six auxiliary links and found the shortest paths from each demand node to the auxiliary sink node to obtain the shortest path set. We assumed that 3000 people needed to be evacuated from the two census tracts, and 7% and 5% of the population for nodes 1 and 2 require a bus for evacuation, respectively. The calculated packet size and packet number are presented in table 14.

Table 14. Packet size and packet number calculated based on car-less people percentage and vehicle type

Node 2 (5 % car-less)					
Vehicle Type	Total evacuees	Evacuees by required vehicle type	Vehicle Count	Packet size	Packet number
Bus	3000	150	3	1	3
Personal Car		2850	712	5	143
Node 1 (7 % car-less)					
Vehicle Type	Total evacuees	Evacuees by required vehicle type	Vehicle Count	Packet size	Packet number
Bus	3000	210	5	1	5
Personal Car		2790	697	5	140

The Bureau of Public Roads BPR function (Ortuzar and Willumsen, 1964) calculates the impedance for the links based on the free flow speed and link capacities. Equation 45 represents the impedance for a link a in the network.

$$t_a = t_{ffs} \times \left[1 + \alpha \left(\frac{V_a}{C_a} \right)^\beta \right] \quad (45)$$

α and β are hyperparameters estimated for and specific to each link type in the network. These parameters are directly used from the MSTM documentation, and they are presented in table 15 for each facility type in the network. Equation 45 and the parameters in Table 15 are used to update each link travel time after each gene's packet assignment.

Table 15. MSTM Model BPR Function Parameters

Link Type	α	β
Interstate	0.7	8
Freeway	0.7	8
Expressway	0.7	8
Major Arterial	0.55	6
Minor Arterial	0.55	6
Collector	0.17	4
Ramp	0.7	8
Local	0.17	4
Ramp	0.7	8

The network consists of 2623 links (considering the six auxiliary links and all zone connectors). The developed algorithm's speed depends on the number of shortest paths derived and the mechanisms utilized to obtain these shortest paths. In our run, we derived 100 possible routes for each mechanism mentioned in section 6.4.1. For the initial population, we obtained 300 genotypes and sorted them using the Non-dominated sorting algorithm. We get the top 60 solutions as a start point for the iterative crossover and mutation steps. The crossover in each iteration was randomly applied to 70% of the population, and the mutation was applied to 40% of the genotypes and performed on 20% of the randomly selected genes. We used parallel processing with a server that benefits from 60 cores. The crossover and mutation percentages are chosen so that precisely 60 genotypes are generated from these mechanisms, and all cores of the computing machine can be utilized. The result of running the algorithm with a population of 60 genotypes for 50 iterations is presented in figure 59.

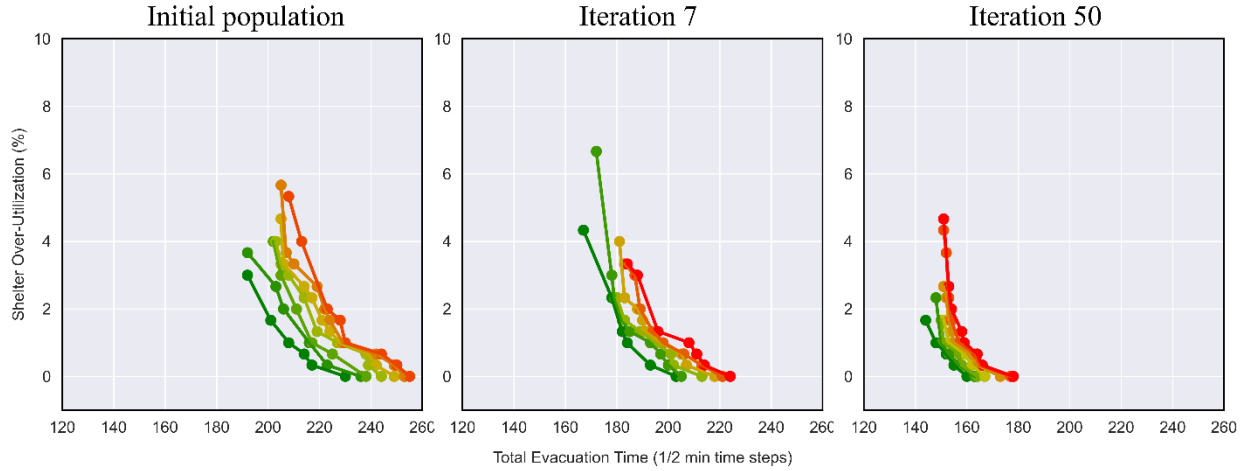


Figure 59. Results of running the algorithm with a population of 60 genotypes for 50 iterations

As shown in figure 59, the first Pareto-Front is converging to a single line. Due to limitations in the usage of the server, we stopped the solution at iteration 50. It is evident in figure 59 (50th iteration) that with 2% overutilization of the shelters, the total evacuation time was reduced by 20 time-steps (10 minutes). In this example problem, we assumed that the capacity of the shelters is equal to 250 vehicles. This capacity limit allows 0% overutilization for some solutions.

Chapter 8. Conclusion and Future Study Pathways

8.1. Conclusions and highlights

The evolution of transportation models through the activity-based model pathway has escalated the problem of micro-area level data needs. Activity models use person-level information and aggregate the travel forecasting to achieve a more realistic overview of travel situations for future years and probable scenarios. The usage of population-level details is not limited to the microscopic travel demand models. The development of accurate and reliable evacuation models is amongst the important problems that extensively rely on multivariate data at its micro area level. Sociodemographic data in its highest resolution exist from national surveys but are typically not available to the public researchers, hindering the models' development. This research suggested a complex yet easy-to-implement machine learning approach that allows generating a synthetic population while preserving the people's privacy.

The sociodemographic data, by its nature, includes categorical and cardinal attributes, and many variable records in the data columns are discrete. Most statistical generative models consider the continuity of the variables as a base in their fundamental theory. Therefore, a robust technic is required to generate synthetic data of a discrete nature. To name an available technique and its restrictive properties, the applicability of the Copula generative model and its limitations in sample generation for multi-level information have been discussed. Then, Generative Adversarial Networks and a couple of their structural variants have been utilized to create synthetic sociodemographic data used in transportation problems. These variants employed Wasserstein's Earth Mover loss function to achieve stable models and prevent mode collapse issues seen when using the cross-entropy loss function. The outperformance of Wasserstein GANs in producing stable results and defeating the mode collapse problem has been shown by multiple examples. This

study showed that using complex models such as convolutional layers for the generative model could increase the stability of the model in converging to the minimum loss function. However, one should note that these models may take longer to train due to the heavy computational time required for training convolution parameters. Also, generating new samples from convolutional neural networks is quite cumbersome if the target population is large.

The most important achievement of this study is the successful application of conditional WGANs to generate data for subareas of a larger geographical area in a standalone model. The suggested approach allows the authorities to develop a synthetic population for micro regions and make it available for public use. It should be noted that the model training for GANs requires skilled staff with experience in data science, computer science, and machine learning. Nowadays, the need and urgency for data scientists, data engineers, and machine learning engineers in agencies are felt. Therefore a substantial budget is allocated for hiring talents to design and train data-driven models. In addition, once a model is trained, it can be integrated with geographical user interfaces (GUI) so that the engineers can easily harness the potential of the black box trained models.

On top of that, application programming interfaces (API) can be utilized for the cloud-based utilization of synthetic information. The method can be used as an online tool to generate real-time synthetic information for any micro area. This approach can provide augmented data for any micro-area of interest and guarantee people's privacy. Beyond that, these data can be used to develop estimation models to predict other parameters. Since most generative approaches (e.g., Copulas) are only valid for continuous variables, the utilization of GANs will improve the results of such models.

The primary goal of this study was to provide population-level synthetic data feed for evacuation models. This dissertation has then been headed toward developing a new evacuation model that considers carless people in very detailed resolution to show an application of such population-level synthetic information. First, a comprehensive non-linear optimization model has been formulated for the bimodal evacuation of the flood zones. The highly nonlinear nature of the problem has been highlighted, and a new MILP formulation that accounts for the capacity change has been presented. The validity of the model has been examined using a small toy network. The toy network has been designed in a minimalistic fashion to track the results and compare the outcomes of different models. Since the proposed exact deterministic models (including our model) are not usually time-efficient or computationally unsolvable for real problems, a heuristic Genetic algorithm-based solution has been developed to extend the applicability of the proposed solution to real-scenario evacuation problems. The GA algorithm has been upgraded to include Non-dominated sorting to simultaneously minimize the overutilization of the shelters and the total evacuation time. This algorithm can iteratively update the travel time and capacity of the links to consider the dynamic nature of the evacuation problems. Another significant contribution of this model is the model's capability to use evacuation orders for the cars willing to depart the hazard zone at time zero. Most research that applied GA-based solutions to evacuation problems lacks this crucial property. The lack of this property hinders the applicability of the models, as the peoples' departure time cannot be estimated beforehand. The proposed model successfully generated a Pareto-Front that covers an acceptable range of utilization status of the shelters and total evacuation time for a network with about 2600 links. The problem aimed to evacuate 6000 cars from two census tracts in Anne Arundel County to 6 shelters away from the hazard zones overnight. Utilizing a Non-dominated Sorting Algorithm allows the decision-makers to decide on

the best trade-off between travel time and overutilization of the shelters, which is measured using the Gini coefficient concept. Applying the algorithm to Anne-Arundel county showed that allowing 2% overutilization of the parking lots would result in 40 minutes of time-saving for evacuation operation of the two census tracts.

Although the heuristic technique proposed in this study can generate a series of reliable solutions for decision-makers, these models are not suitable for real-time strategies. The high computational time associated with the shortest path allocation and capacity constraints and their sanity checks obscures the models' ability to provide real-time solutions and suggestions. In addition, to achieve a reasonable run time, one may require to employ a server with multiple cores and high parallel computing resources.

8.2. Future research suggestions

The Generative Adversarial Neural Networks and their applications and research opportunities are endless. GANs are used in video processing, image processing, data augmentation, feature selection techniques, and imbalance data improvements. For example, the observations with minority labels can be augmented using a GANs model to enhance the application of imbalanced data in supervised learning tools. This study only generated a synthetic population for the data with nine attributes and showed the models' excellent performance. As an expansion of this study, we suggest using GANs to generate a synthetic population for high-dimensional data. As a motivation, remember that GANs is used to create pictures of non-existing human in their most delicate details and resolution. Therefore, capturing the interdependencies among high dimensional data attributes should be achievable with GANs models.

Many research avenues can be followed to improve the optimization models and the heuristic solution. One can add a contraflow strategy, allowing vehicle flows to use all network

links to their total capacity in one direction. The contraflow approach is a fruitful strategy for disaster cases where an all-around statewide evacuation of an area is needed, and no vehicles can pass the risk zone. This is a common characteristic of flood strikes as the affected areas are close to the shorelines. The lanes of bidirectional links can be used in one direction if a lower evacuation time can be achieved. The problem can also benefit from a degree of freedom from direction assignments to each link lane. Two different approaches are proposed for future research to implement the contraflow strategy.

The first approach is to redesign the network to represent each lane of a link as a standalone arc in the graph. Figure 60 illustrates the approach for an example link. Assume that the link from node i to j consists of n number of lanes. These lanes can be modeled in the network as n arcs with two auxiliary connector arcs linking them to node i and node j . The travel times at connectors are equal to 0 and do not impose any impedance to the objective function. The solution to the problem determines the flow direction for each arc.

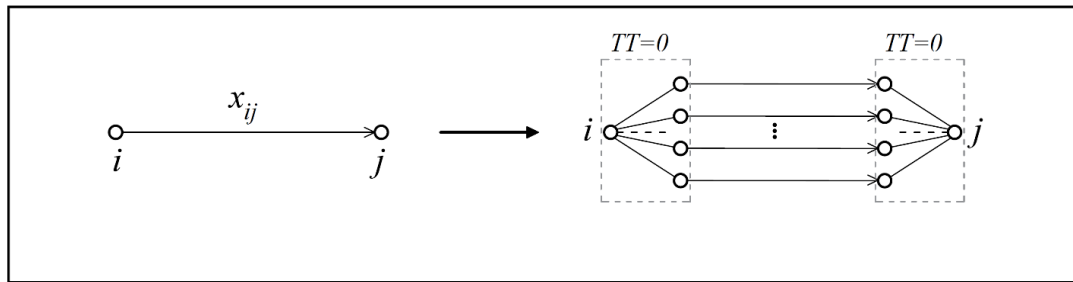


Figure 60. Network redesign to model contraflow

This approach's advantage is that additional constraints for lane numbers are not necessary for implementing the contraflow strategy, which reduces the number of decision variables. This approach's disadvantage is generating a high number of links in the redesigned network that increases the problem's size.

The second approach adds contraflow constraints to the problem that identifies each lane's direction for each link as a new decision variable. These constraints have been described in equations 46-49.

$$\sum_{FT \in \{e, g\}} (l_{FT}^{i,j} + l_{FT}^{j,i}) \leq \mathcal{L}^{ij} \quad \forall (i, j) \in A \quad (46)$$

$$l_{FT}^{i,j} \leq \mathcal{L}^{ij} Y_{FT}^{i,j} \quad \forall FT \in \{e, g\}, (i, j) \in A \quad (47)$$

$$\sum_{v \in \{p, b\}} R_{eq}^v x_{v,i,j}^t \leq \kappa_{ij}^t (l_e^{i,j} + l_g^{i,j}) \quad \forall (i, j) \in A \quad (48)$$

$$x_{p,i,j}^t \leq \kappa_{ij}^t (l_g^{i,j}) \quad \forall t, (i, j) \in A \quad (49)$$

Each lane is categorized into two classes of emergency and general-purpose lanes in this approach. Each link that connects node i to node j has $\mathcal{L}^{ij}$ number of lanes. Equation 46 assigns $l_{FT}^{i,j}$ number of lanes to each link in the direction of i to j for each facility type. If for any reason, the assignment of a positive integer lane number for any facility types or directions is prohibited, equation 47 deprives the link for that specific lane by setting the parameter $Y_{FT}^{i,j}$ to 0. Equation 48 makes sure that the emergency and general-purpose lanes are not overloaded beyond the lane capacity for each link. This formulation assumes that all lanes for each link have the same capacity, κ_{ij}^t , at each time-step. The maximum personal car flow that can use a link is restricted to general-purpose lanes' capacity using equation 49.

This research's focus and core element were to develop a synthetic population generator framework applicable in evacuation model development and planning. The suggested evacuation model was not compared with other existing evacuation models. This was partially due to different assumptions made and strategies employed during the development of different evacuation models. Still, some studies can be used to compare their results with the results of the proposed

study. Therefore, another roadmap to continue this research is designing multiple scenarios and comparing the results using different models. Also, we suggest the researchers apply the generative copula model and GANs model results to the proposed evacuation model and compare the outcome. The comparison can be reinforced by using the two synthetic data in more elaborated transportation models that directly use sociodemographic data. This approach can further reveal the models' strengths and weaknesses.

To sum up, the research paths mentioned above are only a few suggestions that can extend the achievements of this study, and the scope of synthetic population generation and evacuation research requires considerable efforts to improve the existing models.

References

- "2006 report to the Texas House of Representatives,"
<http://www.hurricanescience.org/history/storms/2000s/rita/>
- Abdelgawad H., Abdulhai B., Wahba M. (2010) "Multiobjective Optimization for Multimodal Evacuation," *Transportation Research Record*, 2196(1), 21-33. DOI: <https://doi.org/10.3141%2F2196-03>
- Abdelgawad H., Abdulhai B., Wahba M., (2010) "Optimizing Mass Transit Utilization in Emergency Evacuation of Congested Urban Areas," *Proceedings, the 89th Annual Meeting of The Transportation Research Board*, Washington DC, 2010.
- Amideo A., Scaparra M.P. (2017) "A Scenario Planning Approach for Shelter Location and Evacuation Routing," *Optimization and Decision Science: Methodologies and Applications. Springer Proceedings in Mathematics & Statistics*, 217. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-67308-0_57
- An S., Cui N., Li X., Ouyang Y., (2013) "Location Planning for Transit-Based Evacuation Under the Risk of Service Disruptions," *Transportation Research Part B: Methodological*, 54 1–16. DOI: <https://doi.org/10.1016/j.trb.2013.03.002>
- Apel, H., Thielen, A. H., Merz, B., and Blöschl, G., (2004) "Flood Risk Assessment and Associated Uncertainty," *Nat. Hazards Earth Syst. Sci.*, 4, 295–308, DOI: <https://doi.org/10.5194/nhess-4-295-2004>
- Arentze, T. A., & Timmermans, H. J. P. (2004) "A Learning-Based Transportation-Oriented Simulation System," *Transportation Research Part B: Methodological*, 38(7), 613-633. DOI: <https://doi.org/10.1016/j.trb.2002.10.001>
- Badu-Marfo, G., Farooq, B., and Paterson, Z., (2020) "Composite Travel Generative Adversarial Networks for Tabular and Sequential Population Synthesis," arXiv:2004.06838
- Bailey, D., Swiacki, S., Byrnes, A., Buckley, J., King, D., Piper, V., (2007) "Transportation Equity in Emergencies: A Review of the Practices of State Departments of Transportation, Metropolitan Planning Organizations, and Transit Agencies In 20 Metropolitan Areas (Final Report FTA-PA-26-8001-2007)," Philadelphia: Milligan & Company, LLC, and Mundle & Associates
- Beckman, R. J., Baggerly K. A., and McKay M. D., (1996) "Creating Synthetic Baseline Populations," *Transportation Research Part A: Policy and Practice*, 30(6), 415–429. DOI: [https://doi.org/10.1016/0965-8564\(96\)00004-3](https://doi.org/10.1016/0965-8564(96)00004-3)

- Bhat, C.R., and N. Eluru (2009) "A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling," *Transportation Research Part B: Methodological*, 43(7), 749-765. DOI: <https://doi.org/10.1016/j.trb.2009.02.001>
- Bian R. and Wilmot CG. (2017) "Measuring the Vulnerability of Disadvantaged Populations During Hurricane Evacuation," *Nat Hazards*, 85, 691–707. DOI: <https://doi.org/10.1007/s11069-016-2598-0>
- Bian R. and Wilmot CG. (2018) "An Analysis on Transit Pick-Up Points for Vulnerable People During Hurricane Evacuation: A Case Study of New Orleans," *International Journal of Disaster Risk Reduction*, 31, 1143-1151. DOI: <https://doi.org/10.1016/j.ijdr.2017.07.005>
- Bierlaire M., (2018) "Pandasbiogeme: A Short Introduction. Technical Report," *TRANSP-OR* 181219, Transport and Mobility Laboratory, ENAC, EPFL, Lausanne, Switzerland.
- Bish, D.R., (2011) "Planning for a Bus-Based Evacuation," *OR Spectrum* 33(3) 629–654. DOI: <https://doi.org/10.1007/s00291-011-0256-1>
- Boesch, D.F., Boicourt, W.C., Cullather, R.I., Ezer, T., Galloway, G. E. Jr., Johnson, Z. P., et al., (2018) "Sea-level rise: Projections for Maryland 2018," (p. 27). Cambridge, MD: University of Maryland Center for Environmental Science.
- Borden K.A., Schmidtlein M.C., Emrich C.T., et al. (2007) "Vulnerability of US Cities to Environmental Hazards," *Journal of Homeland Security and Emergency Management*, 1-21. DOI: <https://doi.org/10.2202/1547-7355.1279>
- Born K., S. Yasmin, D. You, N. Eluru, C.R. Bhat, and R.M. Pendyala (2014) "Joint Model of Weekend Discretionary Activity Participation and Episode Duration," *Transportation Research Record*, 2413(1), 34-44. DOI: <https://doi.org/10.3141/2413-04>
- BPS, (2005) "Kecamatan Dalam Angka – Subdistricts in Numbers," *Satistical bureau (BPS)* Kota Padang, Padang.
- Bradley M., Bowman J. L., and Griesenbeck B., (2010) "SACSIM: An Applied Activity-Based Model System with Fine-Level Spatial and Temporal Resolution," *Journal of Choice Modelling*, 3(1), 5–31. DOI: [https://doi.org/10.1016/S1755-5345\(13\)70027-7](https://doi.org/10.1016/S1755-5345(13)70027-7)
- Bretschneider S., (2012) "Mathematical Models for Evacuation Planning in Urban Areas," *Springer-Heidelberg*, New York, Dordrecht, London. DOI: <http://dx.doi.org/10.1007/978-3-642-28759-6>

- Brown D., Knabb R., Rhome J., (2006) "Tropical Cyclone Report: Hurricane Rita 18-26 September 2005" *National Hurricane Center*. 17 Mar. Web: https://www.nhc.noaa.gov/data/tcr/AL182005_Rita.pdf accessed: April 2021.
- Bullock J., Haddow G., (2006) "Introduction to Homeland Security," Butterworth- Heinemann, Published Date: 30th March 2006.
- Bumgarner J.B., (2008) "Emergency Management: A Reference Handbook," *ABC-CLIO*. Santa Barbara, CA.
- C.o.t.R.o.P.T.i.E.E., (2008) "Transportation Research Board of the National Academies, The Role of Transit in Emergency Evacuation," *Special Report* (294).
- Cavusoglu, O., Sisiopiku, V.P., and Juri, N.R., (2013) "Role of Transit in Carless Evacuation Planning," *Natural Hazards Review*, 14(3), 191–199. DOI: [https://doi.org/10.1061/\(ASCE\)NH.1527-6996.0000106](https://doi.org/10.1061/(ASCE)NH.1527-6996.0000106)
- Chakraborty J., Tobin G.A., Montz B.E., (2005) "Population Evacuation: Assessing Spatial Variability in Geophysical Risk and Social Vulnerability to Natural Hazards," *Natural Hazards Review*, 6(1), 23–33. DOI: [https://doi.org/10.1061/\(ASCE\)1527-6988\(2005\)6:1\(23\)](https://doi.org/10.1061/(ASCE)1527-6988(2005)6:1(23))
- Xie C., Lin D., Waller S. T., (2010) "A Dynamic Evacuation Network Optimization Problem with Lane Reversal and Crossing Elimination Strategies," *Transportation Research Part E: Logistics and Transportation Review*, 46(3), 295-316. DOI: <https://doi.org/10.1016/j.tre.2009.11.004>
- Choupani A., and Mamdoohi A., (2016) "Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research," *Transportation Research Procedia*, 17, 223-233. DOI: <https://doi.org/10.1016/j.trpro.2016.11.078>
- CNN (2001) "Live at Daybreak: Escaping a Hurricane," <http://edition.cnn.com/TRANSCRIPTS/>.
- Cuellar L., Kubicek D., Hengartner N. and Hansson A. (2009) "Emergency Relocation: Population Response Model to Disasters," *IEEE Conference on Technologies for Homeland Security*, Boston, MA, USA, 628-635. DOI: <https://doi.org/10.1109/THS.2009.5168096>
- Cutter S.L., Boruff B.J., Shirley W.L., (2003) "Social Vulnerability to Environmental Hazards," *Social Science Quarterly*, 84(2), 242–261. DOI: <https://doi.org/10.1111/1540-6237.8402002>

- Cutter S.L., Finch C., (2008) "Temporal and Spatial Changes in Social Vulnerability to Natural Hazards," *Proceedings of the National Academy of Sciences*, Feb 2008, 105(7), 2301-2306. DOI: <https://doi.org/10.1073/pnas.0710375105>
- Cutter S.L, Mitchell JT, Scott MS (2000) "Revealing the Vulnerability of People and Places: A Case Study of Georgetown County, South Carolina," *Annals of the Association of American Geographers*, 90, 713–737. DOI: <https://doi.org/10.1111/0004-5608.00219>
- Daganzo C. F., (1994) "The Cell Transmission Model: A Simple Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory," *Transportation Research Part B: Methodological*, 28(4), 269-287. DOI: [https://doi.org/10.1016/0191-2615\(94\)90002-7](https://doi.org/10.1016/0191-2615(94)90002-7)
- Deming, W. E. and Stephan F. F. (1940) "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known," *The Annals of Mathematical Statistics*, 11(4), 427–444. DOI: <https://doi.org/10.1214/aoms/1177731829>
- Dong X., Ben-Akiva M. E., Bowman J. L., Walker J. L. (2006) "Moving from Trip-Based to Activity-Based Measures of Accessibility," *Transportation Research Part A: Policy and Practice*, 40(2), 163-180. DOI: <https://doi.org/10.1016/j.tra.2005.05.002>
- Emrich C.T., Cutter S.L., (2011) "Social Vulnerability to Climate-Sensitive Hazards in the Southern United States," *Weather, Climate, and Society*, 3:193–208. DOI: <https://doi.org/10.1175/2011WCAS1092.1>
- Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. (2013) "Simulation Based Population Synthesis," *Transportation Research Part B: Methodological*, 58, 243–263. DOI: <https://doi.org/10.1016/j.trb.2013.09.012>
- FEMA, (1996) "Guide for All-hazard Emergency Operations Planning" *Technical report*, Federal Emergency Management Agency.
- Fermanian, J.D., (2005) "Goodness-of-Fit Tests for Copulas," *Journal of Multivariate Analysis*, 95(1), 119–152. DOI: <https://doi.org/10.1016/j.jmva.2004.07.004>
- Fienberg, S.E., (1970) "An Iterative Procedure for Estimation in Contingency Tables," *The Annals of Mathematical Statistics*, 41(3), 907-917. DOI: <https://doi.org/10.1214/aoms/1177696968>
- Geller R.J., (1997) "Earthquake Prediction: A Critical Review," *Geophysical Journal International*, 131(3), 425–450, DOI: <https://doi.org/10.1111/j.1365-246X.1997.tb06588.x>

- Genest C., and Nešlehová J. (2007) "A Primer on Copulas for Count Data," *ASTIN Bulletin: The Journal of the IAA*, 37(2), 475–515. DOI: <https://doi.org/10.1017/S0515036100014963>
- Genest C., and Rémillard B., (2008) "Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Institut Henri Poincaré, 44(6), 1096–1127. DOI: <https://doi.org/10.1214/07-AIHP148>
- Genest, C., Rémillard, B., Beaudoin, D. (2009) "Goodness-of-Fit Tests for Copulas: A Review and a Power Study," *Insurance: Mathematics and Economics*, 44(2), 199–213. DOI: <https://doi.org/10.1016/j.insmatheco.2007.10.005>
- Goerigk M., Deghdak K., Heßler P., (2014) "A Comprehensive Evacuation Planning Model and Genetic Solution Algorithm," *Transportation Research Part E: Logistics and Transportation Review*. 71 82–97. DOI: <https://doi.org/10.1016/j.tre.2014.08.007>
- Goerigk M., Deghdak K., TKindt V., (2015) "A Two-Stage Robustness Approach to Evacuation Planning with Buses," *Transportation Research Part B: Methodological*, 78, 66–82. DOI: <https://doi.org/10.1016/j.trb.2015.04.008>
- Goerigk M., Grün B., (2014) "A Robust Bus Evacuation Model with Delayed Scenario Information," *OR Spectrum*, 36(4), 923–948. DOI: <https://doi.org/10.1007/s00291-014-0365-8>
- Goerigk M., Grün B., Heßler P., (2013) "Branch and Bound Algorithms for the Bus Evacuation Problem," *Computers & Operations Research*, 40(12) 3010–3020. DOI: <https://doi.org/10.1016/j.cor.2013.07.006>
- Goerigk M., Grün B., Heßler P., (2014) "Combining Bus Evacuation with Location Decisions: A Branch-And-Price Approach," *Transportation Research Procedia*, 2, 783–791. DOI: <https://doi.org/10.1016/j.trpro.2014.09.088>
- Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014) "Generative Adversarial Nets" *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2, 2672–2680. DOI: <https://dl.acm.org/doi/10.5555/2969033.2969125>
- Günthermann L., Simpson I., Roggen D., (2020) "Smartphone Location Identification and Transport Mode Recognition Using an Ensemble of Generative Adversarial Networks," *UbiComp-ISWC 20: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable*, 311–316. DOI: <https://doi.org/10.1145/3410530.3414353>

- Guo J.Y., and Bhat C., (2007) "Population Synthesis for Microsimulating Travel Behavior," *Transportation Research Record: Journal of the Transportation Research Board*, 2014(1), 92– 101. DOI: <https://doi.org/10.3141%2F2014-12>
- Han L.D., Yuan F., Chin S.-M., Hwang H., (2006) "Global Optimization of Emergency Evacuation Assignments," *INFORMS Journal on Applied Analytics*, 36(6), 502–513. DOI: <https://doi.org/10.1287/inte.1060.0251>
- Zhang H., Xu T., Li H., Zhang Sh., Huang X., Wang X., and Metaxas D. N., (2016) "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," *IEEE International Conference on Computer Vision (ICCV)*, 5908-5916, DOI: <https://doi.org/10.1109/ICCV.2017.629>
- Ho J., and Ermon S., (2016) "Generative Adversarial Imitation Learning," *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4565–4573. DOI: <https://dl.acm.org/doi/10.5555/3157382.3157608>
- HRO, (2006) "Evacuation Planning in Texas: Before and After Hurricane Rita," *House Research Organization's Interim News*, Number 79-2. February 14.
- IFRC. (2011) "The Red Cross Red Crescent Approach to Disaster and Crisis Management," position paper, Geneva.
- Ireland C.T., Kullback S. (1968) "Contingency Tables with Given Marginals," *Biometrika*, 55(1), 179–188. DOI: <https://doi.org/10.2307/2334462>
- Isola P., Zhu J.Y., Zhou T., Efros A., (2017) "Image-to-Image Translation with Conditional Adversarial Nets," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1125-1134. <https://arxiv.org/abs/1611.07004v3>
- James J.Q.Y., (2020) "Semi-Supervised Deep Ensemble Learning for Travel Mode Identification," *Transportation Research Part C: Emerging Technologies*, 112, 120-135. DOI: <https://doi.org/10.1016/j.trc.2020.01.003>
- Javanmardi M., Auld J., and Mohammadian K., (2012) "Integration of TRANSIMS with the ADAPTS Activity-Based Model," *4th Transportation Research Board Conference on Innovations in Travel Modeling (ITM)*, Tampa, Florida, April 29 – May 2, 2012.
- Jeong B., Lee W., Kim D.-S., and Shin H. (2016) "Copula-Based Approach to Synthetic Population Generation" *PLoS One*, 11(8). DOI: <https://doi.org/10.1371/journal.pone.0159496>

- Jin Y., Zhang J., Li M., Tian Y., Zhu H., and Fang Z., (2017) "Towards the Automatic Anime Characters Creation with Generative Adversarial Networks," *Computer Vision and Pattern Recognition*, <https://arxiv.org/abs/1708.05509v1>
- Jones J.L., (2004) "Mapping a Flood... Before It Happens," U.S. Geological Survey Fact Sheet 2004-3060, <http://pubs.usgs.gov/fs/2004/3060/pdf/fs20043060.pdf>, (accessed 22 Jan 2021)
- Kalafatas G., and Peeta S., (2006) "Network design for evacuation planning," *11th IFAC Symposium on Control in Transportation Systems*, Delft, The Netherlands, August 29-30-31, 2006
- Kao, S.C., Kim H., Liu C., Cui X., and Bhaduri B. (2012) "Dependence-Preserving Approach to Synthesizing Household Characteristics," *Transportation Research Record: Journal of the Transportation Research Board*, 2302(1), 192–200. DOI: <https://doi.org/10.3141%2F2302-21>
- Karras T., Aila T., Laine S., and Lehtinen J., (2017) "Progressive Growing of Gans for Improved Quality, Stability, and Variation," *Neural and Evolutionary Computing*, <https://arxiv.org/abs/1710.10196v2>
- Kimms A., Maassen K.C., (2011) "Optimization and Simulation of Traffic Flows in the Case of Evacuating Urban Areas," *OR Spectrum*, 33(3), 571–593. DOI: <https://doi.org/10.1007/s00291-011-0264-1>
- Kimms A., Seekircher K., (2015) "Network Design to Anticipate Selfish Evacuation Routing," *EURO Journal on Computational Optimization*. 4, 1–28. DOI: <https://doi.org/10.1007/s13675-015-0057-4>
- Konduri K.C., You D., Garikapati V.M., and Pendyala R.M., (2016) "Enhanced Synthesis Population Generator that Accommodates Control Variables at Multiple Geographic Resolutions," *Transportation Research Record, Journal of the Transportation Research Board*, 2563(1), 40-50. DOI: <https://doi.org/10.3141%2F2563-08>
- Kuefler A., Morton J., Wheeler T., and Kochenderfer M., (2017) "Imitating driver behavior with generative adversarial networks," *IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, 204-211. DOI: <https://doi.org/10.1109/IVS.2017.7995721>
- Kulshrestha A., Lou Y., Yin Y., (2014) "Pick-Up Locations and Bus Allocation for Transit-Based Evacuation Planning with Demand Uncertainty," *Journal of Advanced Transportation*. 48 (7) 721–733. DOI: <https://doi.org/10.1002/atr.1221>
- Lämmel G., Grether D., and Nagel K. (2010-2) "The Representation and Implementation of Time-Dependent Inundation in Large-Scale Microscopic Evacuation Simulations,"

- Transportation Research Part C: Emerging Technologies*, 18(1) 84-98. DOI: <https://doi.org/10.1016/j.trc.2009.04.020>
- Lämmel G., Rieser M., and Nagel K. (2010-1) "Large Scale Microscopic Evacuation Simulation," In: *Klingsch W., Rogsch C., Schadschneider A., Schreckenberg M. (eds) Pedestrian and Evacuation Dynamics 2008*. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-04504-2_48
- Lighthill M. J. and Whitham J. B. (1955) "On Kinematic Waves. I: Flow Movement in Long River; II: A Theory of Traffic Flow in Long Crowded Roads," *Proceeding of Royal Society A*, 229, 281-345. DOIs: <https://doi.org/10.1098/rspa.1955.0088> & <https://doi.org/10.1098/rspa.1955.0089>
- Lin Y., Dai X., Li L., and Wang F., (2019) "Pattern Sensitive Prediction of Traffic Flow Based on Generative Adversarial Framework," in *IEEE Transactions on Intelligent Transportation Systems*, 20(6), 2395-2400, DOI: <https://doi.org/10.1109/TITS.2018.2857224>
- Liu Y., Tremblay J., Cirillo C., (2014) "An Integrated Model for Discrete and Continuous Decisions with Application to Vehicle Ownership, Type and Usage Choices," *Transportation Research Part A: Policy and Practice*, 69, 315-328. <https://doi.org/10.1016/j.tra.2014.09.001>
- Lv Y., Chen Y., Li L., and Wang F., (2018) "Generative Adversarial Networks for Parallel Transportation Systems," in *IEEE Intelligent Transportation Systems Magazine*, 10(3), 4-10, DOI: <https://doi.org/10.1109/MITS.2018.2842249>
- Mamuji A.A., and Etkin D., (2019). "Disaster Risk Analysis Part 2: The Systemic Underestimation of Risk," *Journal of Homeland Security and Emergency Management*, 16(1). DOI: <https://doi.org/10.1515/jhsem-2017-0006>
- MSTM Model Documentation Version 1.0, Report, prepared by The National Center for Smart Growth and Parsons Brinckerhoff for Maryland State Highway Administration, October 2013.
- Masozera M., Bailey M., and Kerchner C., (2007). "Distribution of Impacts of Natural Disasters Across Income Groups: A Case Study of New Orleans," *Ecological Economics*, 63(2-3), 299–306. DOI: <https://doi.org/10.1016/j.ecolecon.2006.06.013>
- McLoughlin D., (1985) "A framework for integrated emergency management," *Public Administration Review*, 45, 165–172. DOI: <https://doi.org/10.2307/3135011>

- Mosteller F., (1968) "Association and Estimation in Contingency Tables," *Journal of the American Statistical Association*, 63(321), 1-28. DOI: <https://doi.org/10.2307/2283825>
- Naghawi H., and Wolshon B., (2015) "Operation of Multimodal Transport System During Mass Evacuations," *Canadian Journal of Civil Engineering*, 42(2), 81-88. <https://doi.org/10.1139/cjce-2014-0177>
- Naghawi H., and Wolshon B., (2010) "Transit-Based Emergency Evacuation Simulation Modeling," *Journal of Transportation Safety & Security*, 2(2), 184-201. DOI: <https://doi.org/10.1080/19439962.2010.488316>
- O'Kelly M., and Horner M., (2003) "Aggregate Accessibility to Population at the County Level: U.S. 1940–2000," *Journal of Geographical Systems*, 5(1), 5-23. DOI: <https://doi.org/10.1007/s101090300101>
- Openshaw S., and Rao. L., (1995) "Algorithms for Reengineering 1991 Census Geography," *Environment and Planning A: Economy and Space*, 27(3), 425–446. DOI: <https://doi.org/10.1068%2Fa270425>
- Pinjari A.R., Eluru N., Copperman R.B., Sener I.N., Guo J.Y., Srinivasan S., and Bhat C.R., (2006) "Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP Models, Framework, Software Architecture, and Application Results," *Technical report*, Midwest Regional University Transportation Center, University of Wisconsin, USA
- Pritchard D.R., and Miller E.J., (2012) "Advances in Population Synthesis: Fitting Many Attributes Per Agent and Fitting to Household and Person Margins Simultaneously," *Transportation*, 39(3), 685–704. DOI: <https://doi.org/10.1007/s11116-011-9367-4>
- Pukelsheim F., (2014) "Biproportional matrix scaling and the iterative proportional fitting procedure," *Annals of Operations Research*, 215(1), 269–283. DOI: <https://doi.org/10.1007/s10479-013-1468-3>
- Radford A., Metz L., and Chintala S., (2015) "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," <https://arxiv.org/abs/1511.06434v2>
- Rao, J.N.K., and Molina I., (2015) "Small-Area Estimation," Wiley, 2nd edition. DOI: <https://doi.org/10.1002/9781118735855>
- Richards P.I., (1956) "Shock Waves on the Highway," *Operation Research*, 4(1), 42-51. DOI: <https://doi.org/10.1287/opre.4.1.42>

- Rui S., Shiwei H., Zhang L., (2009) "Optimum Transit Operations During the Emergency Evacuations," *Journal of Transportation Systems Engineering and Information Technology*, 9(6), 154–160. DOI: [https://doi.org/10.1016/S1570-6672\(08\)60096-3](https://doi.org/10.1016/S1570-6672(08)60096-3)
- Salvini P., and Miller E.J., (2005) "ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems," *Networks and Spatial Economics*, 5(2), 217-234. DOI: <https://doi.org/10.1007/s11067-005-2630-5>
- Schafer J. L., (1997) "Analysis of Incomplete Multivariate Data," *Chapman and Hall/CRC Press*, USA. DOI: <https://doi.org/10.1201/9780367803025>
- Sklar A., (1959) "Fonctions de Répartition à n Dimensions et Leurs Marges," *Publications de l'Institut Statistique de l'Université de Paris*, 8, 229-231. DOI: <https://doi.org/10.4236/oalib.1106100>
- Smith S.K., and McCarty C., (2009) "Fleeing the Storm(s): An Examination of Evacuation Behavior During Florida's 2004 Hurricane Season," *Demography*, 46(1), 127–145. DOI: <https://doi.org/10.1353/dem.0.0048>
- Stern E., (1989) "Evacuation Intentions of Parents in an Urban Radiological Emergency," *Urban Studies*, 26(2), 191–198. DOI: <https://doi.org/10.1080%2F00420988920080161>
- Sun L., Erath A., Cai M., (2018) "A Hierarchical Mixture Modeling Framework for Population Synthesis," *Transportation Research Part B: Methodological*, 114, 199-212. DOI: <https://doi.org/10.1016/j.trb.2018.06.002>
- Theis L., Oord A.v.d., and Bethge M., (2015) "A Note on the Evaluation of Generative Models," <https://arxiv.org/abs/1511.01844v3>.
- Thiede B.C., and Brown D.L., (2013) "Hurricane Katrina: Who Stayed and Why?," *Population Research and Policy Review*, 32(6), 803–824. DOI: <https://doi.org/10.1007/s11113-013-9302-9>
- Trivedi P., Zimmer D., (2017) "A Note on Identification of Bivariate Copulas for Discrete Count Data" *Econometrics*, 5(1), 10. DOI: <https://doi.org/10.3390/econometrics5010010>
- Tüydeş H., and Ziliaskopoulos A., (2005) "Network Redesign to Optimize Evacuation Contra-Flow," *Proceeding of the 84th TRB Annual Meeting*, Washington, D.C.
- Tüydeş H., Ziliaskopoulos A., (2014) "Modeling Demand Management Strategies for Evacuations," *Annals of Operations Research*, 217(1), 491–512. DOI: <http://dx.doi.org/10.1007/s10479-014-1533-6>

- U.S. Census Bureau (2016) "American Community Survey, 2016 Public Use Microdata Area," available at <https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2016/>
- UNISDR (United Nations International Strategy for Disaster Reduction) (2016) UNISDR Terminology, 2009. Accessed November 2. <https://www.unisdr.org/we/inform/terminology>.
- Voas D., and Williamson P., (2000) "An Evaluation Of The Combinatorial Optimization Approach To The Creation Of Synthetic Microdata," *International Journal of Population Geography*, 6(5), 349–366. DOI: [https://doi.org/10.1002/1099-1220\(200009/10\)6:5%3C349::AID-IJPG196%3E3.0.CO;2-5](https://doi.org/10.1002/1099-1220(200009/10)6:5%3C349::AID-IJPG196%3E3.0.CO;2-5)
- Yin W., Murray-Tuite P., Ukkusuri S.V., Gladwin H. (2014) "An Agent-Based Modeling System for Travel Demand Simulation for Hurricane Evacuation," *Transportation Research Part C: Emerging Technologies*, 42(5), 44-59. DOI: <https://doi.org/10.1016/j.trc.2014.02.015>
- Wang C., Yarnal B., (2012) "The Vulnerability of the Elderly to Hurricane Hazards in Sarasota, Florida," *Natural Hazards*, 63, 349–373. DOI: <https://doi.org/10.1007/s11069-012-0151-3>
- Wang J., Wang H., Zhang W., Ip W., Furuta K., (2013) "Evacuation Planning Based on the Contraflow Technique with Consideration of Evacuation Priorities and Traffic Setup Time," *IEEE Transactions on Intelligent Transportation Systems*, 14(1), 480–485. DOI: <https://doi.org/10.1109/TITS.2012.2204402>
- Wendell, C. (2006) "Emergency Evacuation Report Card 2006; 25 Urban Areas Could Face Greater Challenges than New Orleans Experienced After Hurricane Katrina." *American Highway Users Alliance*, http://www.highways.org/wp-content/uploads/2006/10/Evacuation_Study-FINAL.pdf, accessed on Sep 16, 2020.
- Williamson P., Birkin M., and Rees P.H., (1998) "The Estimation of Population Microdata by Using Data From Small Area Statistics and Samples of Anonymized Records," *Environment and Planning A: Economy and Space*, 30(5), 785–816. DOI: <https://doi.org/10.1068/a300785>
- Xu L., and Veeramachaneni K., (2018) "Synthesizing Tabular Data using Generative Adversarial Networks," <https://arxiv.org/abs/1811.11264v1>.
- Yamada T., (1996) "A Network Flow Approach to a City Emergency Evacuation Planning," *International Journal of Systems Science*, 27(10) 931–936. DOI: <https://doi.org/10.1080/00207729608929296>

- Yang X., Ban X., Mitchell J., (2018) "Modeling Multimodal Transportation Network Emergency Evacuation Considering Evacuees' Cooperative Behavior," *Transportation Research Part A: Policy and Practice*, 114, Part B, 380-397. DOI: <https://doi.org/10.1016/j.tra.2018.01.037>
- Yazdizadeh A., Patterson Z., Farooq B., (2019) "Semi-Supervised GANs to Infer Travel Modes in GPS Trajectories," <https://arxiv.org/abs/1902.10768v1>
- Ye X., Konduri K., Pendyala R.M., Sana B., and Waddell P., (2009) "A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations," *88th Annual Meeting of the Transportation Research Board*, Washington, DC.
- Zakharov E., Shysheya A., Burkov E., and Lempitsky V. S., (2019) "FewShot Adversarial Learning Of Realistic Neural Talking Head Models," <https://arxiv.org/abs/1905.08233v2>
- Zhao X., Ren G., Huang Z.f., (2015) "Optimizing One-Way Traffic Network Reconfiguration and Lane-Based Non-Diversion Routing for Evacuation," *Journal of Advanced Transportation*, 50(4), 589-607. DOI: <https://doi.org/10.1002/atr.1362>.
- Zheng H., (2014) "Optimization of Bus Routing Strategies for Evacuation," *Journal of Advanced Transportation*. 48 (7), 734–749. DOI: <https://doi.org/10.1002/atr.1224>
- Zhou Y., Li N., Wu W. et al. (2014) "Local Spatial and Temporal Factors Influencing Population and Societal Vulnerability to Natural Disasters," *Risk Anal*, (34), 614–639. DOI: <https://doi.org/10.1111/risa.12193>
- Zhu J-Y, Park T., Isola P., and Efros A.A., (2017) "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *IEEE International Conference on Computer Vision (ICCV)*. <https://arxiv.org/abs/1703.10593v7>
- Ziemke D., Joubert J., & Nagel, K., (2018) "Accessibility in a Post-Apartheid City: Comparison of Two Approaches for Accessibility Computations," *Networks and Spatial Economics*, 18(2), 241-271. DOI: <https://doi.org/10.1007/s11067-017-9360-3>