ABSTRACT

Title of dissertation: COPULA BASED POPULATION SYNTHESIS AND BIG DATA DRIVEN PERFORMANCE MEASUREMENT

Kartik Kaushik, Doctor of Philosophy, 2019

Dissertation directed by: Professor Cinzia Cirillo
Department of Civil Engineering

Transportation agencies all over the country are facing fiscal shortages due to the increasing costs of management and maintenance of facilities. The political reluctance to increase gas taxes, the primary source of revenue for many government transportation agencies, along with the improving fuel efficiency of automobiles sold to consumers, only exacerbate the financial dire straits. The adoption of electric vehicles threatens to completely stop the inflow of money into federal, state and regional agencies. Consequently, expansion of the network and infrastructure is slowly being replaced by a more proactive approach to managing the use of existing facilities. The required insights to manage the network more efficiently is also partly due to a massive increase in the type and volume of available data. These data are paving the way for network-wide Intelligent Transportation Systems (ITS), which promises to maximize utilization of current facilities. The waves of revolutions overtaking the usual business affairs of transportation agencies have prompted the development and application of various analytical tools, models and and procedures to transportation. Contributions to this growth of analysis techniques are documented in this dissertation.

There are two main domains of transportation: demand and supply, which need to be simultaneously managed to effectively push towards optimal use of resources, facilities, and to minimize negative impacts like time wasted in delays, environmental

pollution, and greenhouse gas emissions. The two domains are quite distinct and require specialized solutions to the problems. This dissertation documents the developed techniques in two sections, addressing the two domains of demand and supply. In the first section, a copula based approach is demonstrated to produce a reliable and accurate synthetic population which is essential to estimate the demand correctly. The second section deals with big data analytics using simple models and fast algorithms to produce results in real-time. The techniques developed target short-term traffic forecasting, linking of multiple disparate datasets to power niche analytics, and quickly computing accurate measures of highway network performance to inform decisions made by facility operators in real-time.

The analyses presented in this dissertation target many core aspects of transportation science, and enable the shared goal of providing safe, efficient and equitable service to travelers. Synthetic population in transportation is used primarily to estimate transportation demand from Activity Based Modeling (ABM) framework containing well-fitted behavioral and choice models. It allows accurate verification of the impacts of policies on the travel behavior of people, enabling confident implementation of policies, like setting transit fares or tolls, designed for the common benefit of many. Further accurate demand models allow for resilient and resourceful planning of new or repurposing existing infrastructure and assets. On the other hand, short-term traffic speed predictions and speed based reliable performance measures are key in providing advanced ITS, like real-time route guidance, traveler awareness, and others, geared towards minimizing time, energy and resource wastage, and maximizing user satisfaction. Merging of datasets allow transfer of data such as traffic volumes and speeds between them, allowing computation of the global and network-wide impacts and externalities of transportation, like greenhouse gas emissions, time, energy and resources consumed and wasted in traffic jams, etc.

COPULA BASED POPULATION SYNTHESIS AND BIG DATA DRIVEN PERFORMANCE MEASUREMENT

By

Kartik Kaushik

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
2019.

Advisory Committee:

Professor Cinzia Cirillo, Chair and Advisor
Professor Eric V. Slud
Professor Parthasarathi Lahiri, Dean's Representative
Dr. Sevgi Erdogan
Dr. Stanley E. Young

Dedication

To my beloved sister Syamini Kaushik and
my beautiful wife Preeti Lakhole.
Two women I cherish the most.

Contents

List of Tables

List of Figures

List of Abbreviations

AADT	Annual Average Daily Traffic
ABM	Activity Based Model
ACM	Association for Computing Machinery
ACS	American Community Survey
ADAPTS	Agent-based Dynamic Activity Planning and Travel Scheduling
AITS	Advanced Intelligent Transportation System
ALBATROSS	A Learning Based Transportation Oriented Simulation System
ARIMA	Auto Regressive Integrated Moving Average
ASTIN	Actuarial Studies in Non-Life Insurance
ATAS	Advanced Traveller Assistance System
ATIS	Advanced Traveller Information System
BIC	Bayesian Information Criterion
BTI	Buffer Time Index
BTM	Bluetooth Traffic Monitoring
BTR	Basic Trust Region
CATT	Center for Advanced Transportation Technology
CDF	Cumulative Distribution Function
CEMDAP	Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns
CHART	Coordinated Highway Action Response Team
CRAN	Comprehensive R Archive Network
DC	District of Columbia
DOT	Department of Transportation
EPA	Environmental Protection Agency
ETH	Eidgenössische Technische Hochschule (Swiss Federal Institute of Technology)
EUSFLAT	European Society for Fuzzy Logic and Technology
FHWA	Federal Highway Administration
FIPS	Federal Information Processing Standards
FRC	Functional Road Classification
GHG	Greenhouse Gases
GPS	Global Positioning System
GoF	Goodness of Fit
HPMS	Highway Performance Monitoring System

IAA	International Actuarial Association
IEEE	Institute of Electrical and Electronics Engineers
IET	Institution of Engineering and Technology
ILUTE	Integrated Land Use, Transportation, Environment
INFORMS	Institute for Operations Research and the Management Sciences
IPF	Iterative Proportional Fitting
IPU	Iterative Proportional Updating
IQR	Inter Quartile Range
IRS	Internal Revenue Service
ISO	Internation Standards Organization
ITS	Intelligent Transportation System
IVT	Institut für Verkehrsplanung und Transportsysteme (Institute for Transport Planning and Systems)
JSTOR	Journal Storage
LLC	Limited Liability Company
MAC	Media Access Control
MAP-21	Moving Ahead for Progress in the 21st Century Act
MCMC	Monte Carlo Markov Chain
MD	Maryland
MPO	Metropolitan Planning Organization
NA	Not Applicable/Available
NALCA	the North American Location Code Alliance
NATMCCA	The North American TMC Code Alliance
NHS	National Highway System
NHTS	National Household Travel Survey
NPMRDS	National Performance Management Research Dataset
NPRM	Notice of Proposed Rule Making
NREL	National Renewable Energy Lab
ORNL	Oak Ridge National Laboratory
OSM	Open State Map
OUP	Oxford University Press
PDF	Portable Document Format
PM	Performance Measurement
PTI	Planning Time Index
PUMA	Public Use Micro Area
PUMS	Public Use Micro Sample
RMSE	Root Mean Squared Error

RRMSPE	Relative Root Mean Squared Prediction Error
RV (rv)	Random Variable
SAE	Small Area Estimation
SHA	State Highway Administration
STRAHNET	Strategic Highway Network
T.T.I.	Texas Transport Institute
TAZ	Traffic Analysis Zone
TIM	Traffic Incident Management
TISA	Traveler Information Services Association
TMC	Traffic Message Channel
TMS	Traffic Message Sign
TOPS Lab	Traffic Operations and Safety Laboratory
TPEG	Transport Protocol Experts Group
TRANSIMS	TRansportation ANalysis SIMulation System
TTI	Travel Time Index
U.K.	United Kingdom
U.S.	United States (of America)
UMD	University of Maryland
VPP	Vehicle Probe Project
WGS	World Geodetic System
WSDOT	Washington State Department of Transportation
cdf	cumulative density function
cfd	cumulative frequency distribution
cmf	cumulative mass function
pmf	probability mass function

Chapter 1: Introduction

There are two main aspects of managing performance of transportation facilities: supply side and demand side. Supply side management deals with allocating required resources to meet the demand. It also deals with controlling access, or imposing other restrictions, including tolling, taxing, and so on, to ensure the available resources are better utilized. These insights depend heavily on real-time data and analytics, and benefit from accurate short-term predictions, as management can be made proactive rather than reactive. While supply side management is dependent mainly on drafting the right control policies and the political will to implement them, demand side management is rather more difficult, as it not only involves inferring the behavior of individuals, but also influencing individual behavior towards system optimality. Recently, tremendous progress has been made in understanding and modeling human choices and behavior by studying and implementing a new framework of Activity Based Models (ABM). ABM takes the view that travel demand is more endemic to human nature, that it is product of the activities of individuals cognizant of their social groups. Estimations of demand from ABM are based on simulations and require a reliable synthetic population.

The synthetic population is created using copulas, which requires individual observations of a sample of people. Copulas are multivariate distribution functions that represent just the underlying dependence in multivariate random variables, decoupled from the univariate marginal distribution of the sample data. Copulas are the perfect tool to capture the dependence between various characteristics of people,

such as age, gender, race, education, income and others, because they are not biased by the marginal distributions. Therefore, a copula that captures such multivariate dependence in some area can be used to generate synthetic population with the same dependence structure in the target area, so long as it is reasonable to expect the same dependence structure at the target area. While it may be empirically testable, it can be seen that this assumption holds in most cases, as people of certain socioeconomics and demographics lead similar lifestyles.

Forecasting of traffic patterns, even for short-term durations, is a very complex problem as there are many internal and external variables that affect the roadway conditions. However, by accumulating sufficient data, some patterns can be gleaned, which can be used to increase the accuracy of the predictions. An innovative adaptation of the small area estimation method is used for predicting traffic speeds. The method considers the future unobserved time points as small areas, and borrows strength from historic observations of the speeds. Similarly, accumulating sufficient data from the recent past over the same time window, say 24-hours, allows filtering of outliers and direct estimation of the traffic conditions, for any time interval of interest. The congestion measures thus developed can be computed quickly, in real-time, or to study certain effects retrospectively. The techniques to simultaneously filter outliers and compute performance measures have been adapted by the Federal Highway Administration (FHWA) for drafting reporting rules under the Moving Ahead for Progress in the 21st Century Act (MAP-21), which directly impacts federal funding provided for maintenance and management of facilities.

Data used in transportation exist in very disparate datasets, which are reported to and are available in different spatial and temporal domains. A lot of studies and uses for the data are hampered due to the discrepancies. Consequently, techniques to merge different datasets together across spatial and temporal differences are in high demand. Two innovative conflation methods that merge different datasets together

are demonstrated in this dissertation. One method filters incidents to include just those that occur in the spatial and temporal vicinity of traffic bottlenecks on the roadway. This allows attributing incidents to bottlenecks, simplifying computation of the true, system-wide cost of an incident, in terms of wasted time, energy, resources, and emissions, apart from the direct costs due to life, limb and property damages. The second conflation algorithm joins two roadway networks in the same spatial plane, but with different geometry representations, enabling transfer of information, such as traffic volumes, from one map layer to another. Volume information especially allows estimating the impacts due to transportation, and computing costs and resource consumption over all users. An extension to this method exposes issues with the National Performance Management Research Data Set (NPMRDS), a federal dataset of traffic speeds and volumes on roadways, made available to transportation agencies to aid in MAP-21 reporting.

1.1 Problem Statement

Multiple problems are tackled in this dissertation, with the underlying theme of improving the data analytics powering the insights into transportation decisions made for providing safe, efficient, speedy and equitable travel for all. The issues resolved by the developed methods include aspects of both the demand for travel and supply of infrastructure and facilities to meet that demand. The availability of a dataset containing full socioeconomic and demographic records of everyone in the region of interest is very expensive to obtain and maintain. However, it is crucial to have such a dataset to be able to simulate the choices of every individual in the region of interest. These choices ultimately precipitate as demand for travel and specific transportation modes. Understanding these choices allows preparing the right solution to meet the demand, and to incentivize choices that allow for better optimality. Therefore, a model that can accurately produce the population at the individual level in the area

of interest of utmost importance.

Management of facilities in real-time requires fast insights, usually obtained from big data observed all over the network. Therefore, the methods and algorithms need to be accurate, yet compute quickly to produce the results in real-time. Therefore, the method developed to compute performance measures simultaneously filters outliers, and can be deployed in a rolling manner to produce results for any number of segments in the roadway network. Similarly, the short-term forecasting models are fitted and selected before hand for each day of the week. Both methods use historic observations from a few weeks to even years prior to the time point where results are required.

Often, in transportation science, the required data for producing certain important results are separated over multiple datasets. Merging of the data are essential to produce the required results, which for the core of reports and studies of long-term metrics on transportation. Such reports are now mandatory under the MAP-21 rules. Therefore, algorithms that accurately merge data together across spatial and temporal domains are highly sought after. The algorithms developed not only merge incidents, roadway speeds and volume data together across a period of a year, but also uncovered merge problems in FHWA provided NPMRDS dataset.

1.2 Research Objectives

The main goal of the researches documented in this dissertation was to produce solutions to problems, and strengthen weak links in the state-of-art of transportation science. The solutions also add to human knowledge, as they are all innovative, and novel in the area where they are applied. Providing multidisciplinary solutions was also an objective of the conducted research, and all developed methods are strongly founded in multidisciplinary practices, and use techniques from domains of probability, statistics, economics, transportation, and others. Lastly, developed methods are simple and straightforward, with a low barrier to entry for real-world applications.

The use of copulas for population synthesis was motivated by the need to capture the dependencies among the characteristics of people, and increase number of characteristics simultaneously modeled. Simultaneously, the copula approach combines data from different datasets, which also enables additional analysis on the combined data. Computing performance measures, after filtering outliers, had a high demand in the industry after the NPMRDS was released. Similarly, the computing assets and technical expertise of professionals working for government transportation agencies influenced the choice of the forecasting method. Therefore, the developed methods were designed to be simple, fast and easy to implement on standard hardware. Lastly, the conflation of multiple geocoded data allows producing results that are not possible with a single dataset, vastly simplifying analysis methods required to produce the required insights.

The underlying objectives central to the developed methods described in this dissertation are to use simple and efficient yet effective techniques to achieve the combined goal of reliable and accurate estimation of required variables. Many of the supply side methods are targeted at practitioners, while the synthetic population is meant for use by advanced modelers implementing ABM.

1.3 Contributions

Transportation is undergoing massive shifts to become more efficient, equitable and ubiquitous, and much of this change is powered by data, analytics and policies driven by the results of those analytics. Most of the data available today never existed a decade ago. Further, the breakneck pace of technological revolution is constantly opening new modes, options and challenges in transportation domains. Compounded by the advancements in computation power, and improved, yet cheap, access to a high number of processors, many models that have only been theoretically possible have been implemented. Methods documented in this dissertation have been produced to

address the challenges encountered over the past six years. Consequently, there is no single problem definition to the work documented in this dissertation.

The tackled issues can also be classified into the demand and supply domains. Population synthesis to census tracts that act as building blocks when estimating origin and destination matrices, mode shares matrices, and other demand related aspects has been nearly impossible before the development of copula based techniques. Similarly, the method to simultaneously filter outliers and estimate performance measures for segments forming road networks was speedy and effective for use not only in real-time applications, but also by FHWA in drafting reporting rules under MAP-21. Further, the synthetic time series framework for forecasting traffic speeds is extremely fast and versatile to make accurate predictions for the whole roadway network every minute, in real-time. Lastly, the conflation algorithm not only enables transfer of volume data for estimation of the total impacts of transportation, but also uncovers issues with the NPMRDS, a dataset crucial for reporting under MAP-21, and securing federal funds.

A summary of the contributions of the developed methods documented in this dissertation are:

1. A novel method using copulas to synthesize population representative of a geographical area that overcomes limitations in the current state-of-art. The obtained synthetic population can be used with ABM, and other microsimulation models to provide accurate estimates of demand on links in the network.

2. Beginnings of solutions to problems inherent in using copulas with discrete, and categorical data, and validation of the goodness of fit of copulas.

3. A flexible and extensible framework founded on Small Area Estimation literature for producing short-term forecasts of various transportation related variables. Any model can be used in this framework to predict the required data. The model specification is also unrestricted, so any number of auxiliary variables can

be used.

4. Algorithm to compute performance measures from NPMRDS, despite the noise and outliers in the dataset. This algorithm has been adapted into the national performance management rules by the FHWA under the directions in MAP-21.

5. Spatial and temporal join of congestion and incident datasets so that a causative relation can be established. The findings are used by Maryland State Highway Administration to direct patrol distribution on Maryland roads.

6. Conflation of the speed and volume data so that estimates of collective impacts can be obtained. Collective impacts refer to the externalities of transportation like cost of delays, energy and resource consumption, emission of greenhouse gases and other pollutants. The conflation is being used at National Renewable Energy Laboratory to produce estimates of externalities.

1.4 Dissertation Organization

This dissertation is organized in two main sections corresponding to the demand and supply management domains of transportation. Each section has an introduction, a literature review, description of the data used, a discussion about the methods applied, results and conclusions. The demand domain is covered in the first section, while the analytics required for supply management is documented in the second section. The overall conclusions and future work make up the last two chapters.

SECTION I
Population Synthesis

Chapter 2: Introduction to Population Synthesis

Managing demand is increasingly viewed as an effective solution to controlling the negative effects of transportation, chiefly congestion, environmental impacts, wasted energy, time and resources. Construction of new facilities in response to increasing demand, once viewed as the only solution, is no longer feasible due to the escalating costs of construction, and subsequent maintenance of the facilities. Revenue streams have not kept pace with the rising costs, chiefly because gasoline taxes have not been raised in decades. Furthermore, increasing numbers of drivers, travel distances, vehicles and fuel efficiencies have contributed to exacerbate the wear and tear of roadways without contributing to the revenues. The idea that building more capacity only fuels more demand till the network is congested again has become an adage.

Understanding demand is now the prime focus, however, it requires careful study of human lifestyles and decision making. A large set of logistic regression models have been developed over the past two decades that deal with modeling various choices made by people. Choices modeled not only include lifestyle decisions like selected activity, time spent doing the activity, location of residence, school, or work, vehicle ownership, and others, but also transportation variables like time of departure for the activity, mode of travel, etc. Activity Based Model (ABM) framework combines these individual choice models together, and thus allows for modeling joint decisions by all member of household, while taking into consideration the individual activity, work and other constraints like dropping or picking kids from school, buying groceries etc.

The models comprising ABM can be fitted using well developed and studied

algorithms on sample datasets of people. However, to make predictions about travel behavior in the region of interest using ABM, it is critical to have a dataset that accurately represents the entire population of that region. Such dataset would be used for simulating the choices of every individual. The most accurate would be a survey detailed of each individual in the area of interest to populate the dataset. However, such a data collection effort is often perceived as too intrusive in private lives, quickly becomes prohibitively expensive, is difficult to maintain over time, and gets outdated in a few years as technology, policies, land use patterns, and lifestyles change. Therefore, a solution is sought with modeling to create a synthetic population dataset that accurately mimics the true population of the area. The proposed method uses people as the primary unit, and attributes household level variables to individuals. However, it can also be expanded to other units that can be surveyed, such as families and households.

The synthetic population is produced using copulas, which are multivariate probability functions that model the dependence within multivariate random variables, independent of the univariate marginal distributions of each variate. Copulas fitted to survey data capture the dependence among various characteristics of people, and can be used to generate data with the same structure, producing the synthetic population. It allows transferring the dependence from one region to another, where the marginals might be different, but the dependence is expected to be the same. The dependence is usually transferable, as people with in a certain age bracket, having completed a certain level of education, and working in a specific industry have similar scales of income. The marginals from one location to another might be different, for example, a neighborhood of predominantly blue collar workers, compared to a neighborhood of affluent white collar employees, or the differences in pay scales, or household sizes from one city to another.

Current state-of-art methods in literature to synthesize population use an iterative

procedure to estimate the contribution of basic demographic characteristics of people to the whole population. It uses a method of ratios to compute the relative abundance for each record in a sample such that marginal totals are satisfied. However, given that it is a numerical procedure that requires a full enumeration of all samples in a matrix form, with marginal totals, it does not scale will with additional characteristics or even attributes in each characteristic. Consequently, it is limited to very few dimensions. Copulas, on the other hand, are regularly used with hundreds of dimensions, and therefore, have the potential to model not just basic demographic information, but also socioeconomic, and even lifestyle and choice variables that directly impact transportation demand.

Once an ABM is fit, the synthetic population is used to generate predictions about behavior and choices for areas of interest. The developed method can directly synthesize population to census tracts, which is impossible to do with the methods explored in current literature because the number of samples in any one census tract are too few. Having predictions about the choices, including likely mode of transport, for all simulated individuals in a census tract, an origin-destination matrix can be built to index the number of trips bound from one region in a given time window, to destination regions. These trips could then be used to estimate the demand load on infrastructure pertinent to various modes, plan for future expansions of the network by mode, and determine fares, tolls and charges based on consumer profiles and willingness to pay.

The ABM along with the synthetic population, when correct, provide simple means to verify the impacts of various policy decisions. Like the fuel price elasticity can be computed to understand the impacts of changing the fuel tax. Similarly, the impact of increasing frequencies or coverage of public transport on the people affected can be quickly estimated by changing a few basic control variables like cost and time of traveling. Expected demand increases can be used to decide the tolling on dynamic

tolled facilities to ensure the level of service is maintained on the facility, while the revenues are maximized.

In this half of the dissertation, the literature review is presented in chapter 3, followed by a description of the datasets used in chapter 4. Chapter 5 discusses the developed method using synthetic data for illustration. Additionally, various mathematical issues related to discrete data, which are methodologically unsolved, are raised, and workarounds are explored. The results are presented in chapter 6 for a randomly selected PUMA in Maryland and two randomly selected census tracts within that PUMA as an example of the prowess of the developed method. Beginnings of workarounds to solve the issues encountered with the use of discrete data in copulas, and testing the similarity between the multivariate dataset produced by the constructed copulas with the original multivariate dataset are presented in chapter 7. Finally, the conclusions are discussed in chapter 8.

2.1 Overview of Developed Method

Copulas are used to capture the dependence in basic characteristics reported by respondents in a survey. The goal is to use the constructed copula to generate a synthetic population that accurately reproduces the relation between various attributes of the real population. The developed method accomplishes this objective by jointly modeling household and individual data. The synthetic population so produced can be used for a large number of microsimulation models that require the population and its detailed characteristics in the study area. A class of such models are the ABM, which are increasingly gaining popularity for understanding and modeling travel behavior and choices of individuals in family units [3, 8, 43, 51, 92, 103, 134]. Moreover, the method can also produce synthetic population in regions with limited number of observations in the original dataset, that is, in small areas [41, 48, 54, 74, 97, 98, 99].

This study uses 1-year Public Use Micro Sample (PUMS) data coded to Public

Use Micro Area (PUMA), released by the American Community Survey (ACS) for the year 2016 [10]. The copulas are constructed using the PUMS data, which contains the characteristics of sampled respondents. The constructed copulas are then used to generate the synthetic population in census tracts. The generated data are inverse probability integral transformed using the aggregated totals provided by the 2010 Decennial U.S. census [9]. The complete procedure illustrates population synthesis, with preserved dependence among characteristics, within small areas.

2.2 Theory of Copulas

A d-dimensional copula C is a multivariate distribution function on $[0,1]^d$ having all marginals uniformly distributed on $[0,1]$ [55, 57, 81]. Copulas are a fairly modern mathematical probability tool which are increasingly being used in applied areas — mainly in the field of actuarial sciences [23] — as they allow modeling of jointly distributed random variables. We refer to Genest and Favre [33] and Okhrin, Ristig, and Xu [84] for a gentle introduction and to the books by Joe [55, 57] and Nelsen [81] for readers seeking concrete mathematical proofs, theorems and derivations related to copulas. A fundamental result, due to Sklar [109], states that any multivariate distribution can be represented by means of a copula.

Theorem. *Let H be a multivariate distribution function with margins $F_1, \dots, F_d$, then there exists a copula C such that*

$$H(x_1, \dots, x_d) = C\left\{F_1(x_1), \dots, F_d(x_d)\right\}, \quad x_1, \dots, x_d \in \overline{\mathbb{R}}. \tag{2.1}$$

If F_j are continuous for $j = 1, \dots, d$ then C is unique. Otherwise C is uniquely determined on the Cartesian product of the range of the marginals $F_1\left(\overline{\mathbb{R}}\right) \times \dots \times F_d\left(\overline{\mathbb{R}}\right)$. Conversely, if C is a copula and $F_1, \dots, F_d$ are univariate distribution functions, then function H defined above is a multivariate distribution function with margins $F_1, \dots, F_d$.

We denote a parametrized copula as C_θ, where θ is the parameter vector (possibly of dimension one). A copula family, indexed by the parameter vector, spans a subset of the range of dependencies between two bounds given by the Fréchet–Hoeffding theorem,

$$W(u_1, \dots, u_d) \leq C_\theta(u_1, \dots, u_d) \leq M(u_1, \dots, u_d), \tag{2.2}$$

$$\begin{aligned} W(u_1, \dots, u_d) &\coloneqq \max\left\{1 - d + \sum_{j=1}^{d} u_j, 0\right\}, \\ M(u_1, \dots, u_d) &\coloneqq \min\left\{u_1, \dots, u_d\right\}. \end{aligned} \tag{2.3}$$

M is the upper bound and represents the comonotone case. The independence copula, $\Pi(u_1, \dots, u_d)$, occurs somewhere in between W and M. Independence copula, also known as product copula, has the distribution function

$$\prod_{j=1}^{d} u_j. \tag{2.4}$$

Note that the upper bound M is always a copula and is sharp, that is, attainable by the function. However, the lower bound W is only a copula in two dimensions. In higher dimensions, W is point-wise sharp, that is, for some $\mathbf{u} = (u_1, \dots, u_d)$, there exists a copula $\breve{C}_\theta$ such that $\breve{C}_\theta(\mathbf{u}) = W(\mathbf{u})$ [55, 57, 81]. Further, a copula family does not necessarily span the whole range between the upper and lower Fréchet–Hoeffding bounds [see table 4.1 in 81, for limiting and special cases of dependencies of many Archimedean copulas]. The identity (2.1), along with the inequality (2.2), presents how copulas allow measurement of the dependence in the data by decoupling the effects of marginal processes on the data from the dependence structure [55, 57, 81].

Given some data with intrinsic dependence, which must exist between the bounds of (2.2), multiple straightforward techniques are available to estimate the parameter of a copula family under various scenarios [34, 56, 67]. Among them, the maximum

likelihood estimation method has been shown to be unbiased, and efficient [34, 36, 66, 69, 70]. However, the challenge is to validate the correct copula family, as Sklar's representation theorem only ensures the existence of a copula, but does not provide any clues about the possible families, or its construction. Moreover, the copula is not necessarily unique if the marginals are not continuous.

The uniqueness issue is relevant to this study because of the use of survey data, which are usually discretized with a predetermined and fixed number of responses, or are categories, or are aggregated into bins. For example, continuous variables like age, income, distance to school or work, area of house and land, and so on are usually reported in bins. Further, the survey also records responses which are discrete by default, like number of household or family members, number of rooms or cars, and so on. Some variables are inherently categorical, like race, gender, level of education, or industry of employment, etc. Discrete variables can sometimes, but not always, be approximated by continuous variables. Categorical variables cannot be derived from a continuous distribution [57].

Consider a dataset $\mathcal{X}$ formed of n independent and identically distributed copies of d-dimensional vector $\mathbf{x}$: $\mathcal{X} = \{\mathbf{x}_i\},\ i \in \{1, \dots, n\}$. We denote the j^{th} component of $\mathbf{x}_k$ by x_{kj}. If any x_{kj} is discrete, it may cause ties in the dataset. In other words, the data tend to repeat: $x_{ij} = x_{kj}$, for some $i \neq k$, and some $j \in \{1, \dots, d\}$. Surveys acknowledge this fact, and even take a step further by providing weights with the data for each unique record. The sum of the weights are usually designed to equal the total population in the area surveyed. Weights are inherently built into surveys as they are conducted on a sample of the population [10, 74]. In the simplest use case, the weights are intended as a frequency measure to determine the number of times each record should be repeated to get a dataset approximating the full population of the area.

While, by definition, a copula is defined over the entire hypercube, $[0, 1]^d$, in order to fit a copula on discrete data, we only have to consider it over the domain formed

by the Cartesian product of the marginals [see, for example 71], an approach pursued in the method documented in this dissertation as well. There have also been some attempts at using Bayesian inference techniques to construct copulas with discrete marginals, but they have not been widely adopted by researchers [112, 113]. Maximum likelihood estimation method can be successfully used to construct copulas with discrete marginals, as the copula family is specified before estimating the parameter [36]. The grunt work then remains to decide whether the chosen copula family is indeed appropriate to model the data at hand. A tie adapted procedure proposed by Kojadinovic [66] is used for testing the fit of the copulas strictly within the domain formed by the Cartesian product of the marginals.

2.2.1 Empirical Marginals

In order to construct copulas, specification of the marginals is essential, as they form the core of the copulas. Two main avenues can be used to approximate a marginal distribution. The first one fixes the distribution of the marginals, subject to some parameters, before estimating the copula parameters and is useful when the processes generating the marginals are known. The second approach relies on the empirical distribution function,

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left(x_{ij} \leq t \right), \quad j \in \{1, \dots, d\}, \tag{2.5}$$

where $\mathbf{1}$ is the indicator function. This two-step approach is favored in the literature as it is immune to misspecification of the marginals. This approach further ensures that the copula is constructed directly from the data [34, 56, 64, 67, 70].

The empirical distribution function (2.5) is tantamount to the ranks of the observations, as $r_{ij} = n\hat{F}_j(x_{ij})$ is the rank of x_{ij} among all $x_{1j}, \dots, x_{nj},\ j \in \{1, \dots, d\}$.

Probability integral transformed observations using (2.5) can be obtained as

$$\hat{\mathbf{u}}_i = \frac{1}{n+1}\left(r_{i1}, \ldots, r_{id}\right), \quad i \in \{1, \ldots, n\}, \tag{2.6}$$

where $\hat{\mathbf{u}}_i := (\hat{u}_{i1}, \ldots, \hat{u}_{id})$ are also called pseudo-observations. The asymptotically unbiased scaling factor $n/(n+1)$ ensures that the pseudo-observations are contained within $(0, 1)$, preventing application of the inverse distributions to 0 or 1, as it could produce infinities for unbounded distributions.

Obtaining the marginal distributions using (2.6) accords some favorable advantages:

1. Since ranks are a maximally-invariant statistic of data, the estimated marginals are invariant under scale-transformations, or monotone increasing transformations of the data, and are produced from the data without loss of information [33].

2. Use of ranks prevents biases in inferences about the copulas [35, 66].

3. Rank-based measures allow for misspecification of the marginals [36]. Since explicit specification of the marginals is not possible for ACS data, this property is indispensable.

4. All of the accepted, reliable, and accurate goodness-of-fit tests in literature use rank-based estimates of marginals [40].

When the marginals are continuous, the empirical distribution function produces ordinal ranks, and (2.6) is a consistent and asymptotically unbiased estimator of the marginal distribution function. However, (2.5) produces maximal ranks when ties are present in the data. Maximal ranks are inherently extreme value distributed, and thus (2.6) is no longer unbiased. Therefore, average ranks are used in (2.6), instead of (2.5) [as also advised in 66]. Average ranks are essentially the average value of the ordinal ranks assigned to each tied value.

A simple example is presented to illustrate average ranking with ties. Let $\{a,\ b,\ c,\ d\}$ be four numbers such that $a \leq b \leq c \leq d$, and suppose $b = c$. The ordinal ranks of the numbers are $1,\ 2,\ 3,\ 4$, while the corresponding average ranks are $1,\ 2.5,\ 2.5,\ 4$. On the other hand, (2.5) would produce maximal ranks, which are $1, 3, 3, 4$. Average ranks ensures that tied values follow the central tendency of the ordinal ranks. These properties make it the most suitable for constructing copulas using maximum likelihood, and for testing the goodness of fit [66].

Chapter 3: A Review of Population Synthesizers

Population synthesis enables microsimulation of people and their choices. Additionally, synthesized population is also used to mask the true identities of the people, while retaining general characteristics about the population. The population synthesized using the presented copula method can be used for either, or any other application like obtaining estimates for small areas, since the outcome is generic population. This section presents a review of recent literature in population synthesis related to the transportation engineering domain, although many other domains have developed specific methods to address their needs. The set of variables used in this dissertation have been set up specifically to answer research question pertinent to transportation, such as demand modeling.

Through the long history of literature on population synthesis, there have been very few attempts at using copulas for generating population [17, 78]. Most prevalent methods have been based on numerical approximation techniques, and specifically the Iterative Proportional Fitting (IPF). IPF is due to Deming and Stephan [22], and was first used in transportation for synthesizing population needed for microsimulation in the TRANSIMS project [5]. IPF essentially reweighs counts from a sample, till the combined weight matches the constraints imposed by the aggregate statistics of the area of interest [see 17, for a brief explanation of the IPF procedure]. IPF, however, suffers from some limitations that have been a constant source of more research to improve population synthesis [17, 78]. As examples:

1. IPF convergence may take an indeterminate amount of time and is not guaran-

teed, especially as the number of characteristics increase. Therefore, combined estimation of household and personal characteristics is a challenge [95]. An updated procedure called Iterative Proportional Updating (IPU) proposed by Ye et al. [134] handles this issue to some extent.

2. IPF will not update a cell with a starting value of zero, i.e., no observations for some rare characteristic in the sample dataset, regardless of the value of the marginal [93].

3. If a marginal total is zero, it is more problematic as it converts all values in the IPF table for that marginal to zeros. This eventually crashes the convergence of the IPF algorithm [134]. A simple fix is to use a small non-zero number for the marginal, however that may introduce biases in the final result [43].

4. IPF matrices can occupy a large space in memory, increasing exponentially with the dimensions [93].

5. Also, the sparsity of the IPF tableau increases exponentially with each new characteristic added to tableau, feeding the zero cell problem. Therefore, IPF is severely limited to just a few tightly coupled characteristics [93].

6. The final IPF tableau cells may contain non-integer numbers, and rounding may induce significant biases [93], although sampling the final population from the weights using Monte Carlo Markov Chains has the ability to somewhat alleviate this limitation [78].

7. In its native form, IPF cannot work across geopolitical zones, which is a serious limitation because data from different surveys are seldom collected over the same zone [78, 134].

There have been developments that overcome the issues mentioned above, at least partially. Most notably, the works by Arentze and Timmermans [3], Bradley, Bowman,

and Griesenbeck [8], Guo and Bhat [43], Javanmardi, Auld, and Mohammadian [51], Pinjari et al. [92], Pritchard and Miller [93], Salvini and Miller [103], and Ye et al. [134], have significantly contributed to the current practices of population synthesis. These methods mainly focus on simultaneous or two-step reweighing of both household and individual characteristics across multiple geopolitical levels. Additionally, some address the size of IPF tableau in memory [93, 103], the convergence problem of IPF [134], and the rounding biases introduced in the final tableau [43].

A clean solution for the zero-cell problem, unfortunately, does not exist when using IPF, making it difficult to fit to small geographies like census tracts [43]. Ye et al. [134] offer a solution that borrows information from the larger area. However, the method requires some fine-tuning to ensure that the demographic groups are not over represented. Similarly, they also overcome the zero-marginal problem encountered in the updated IPF — called Iterative Proportional Updating (IPU) — by assigning a small positive value to zero-marginal categories.

The IPU is an improvement to IPF method to handle combined household and population marginal controls at the marginals. It can also work with marginals from different geographic levels, as long as they have been coaxed beforehand into the geographic level where the population is being synthesized. However, IPU still suffers from all the other shortcomings of IPF, including biases introduced by rounding errors of the final weights. Moreover, IPU also sacrifices the convergence guarantees of IPF, and therefore, Ye et al. [134] advocate a heuristic to solve for the optimal final weights assigned to each observation.

Another crippling feature of IPF is the requirement of sample data, which is not readily available in much of the world. Sample free techniques to synthesize population have embraced other methods. As example, Barthelemy and Toint [4] proposed a complicated hierarchical method dependent on heuristic sampling to fit household and individual characteristics. Schafer [104], among various techniques, presented a

Bayesian technique to infer the final values of the initial IPF tableau, which potentially deals with the requirement of samples, provided the prior distributions are correctly selected. With flat priors, this technique may circumvent the requirement of sample observations.

Statistically driven techniques are quite rare in the literature history of population synthesis for transportation. Copulas, for instance, have only been used in two studies previously. Kao et al. [59] fitted only multivariate normal distributions, and modified the covariance matrix considerably before the results could be validated. A purely computational technique, using empirical copula, was proposed by Jeong et al. [53], mainly intended to validate dependence captured by the IPF method. There have been few other papers that use optimization techniques such as combinatorial optimization [86, 131], simulated annealing [123], and numerical sampling methods [25]. However, these methods have not gained as much traction as IPF.

The Monte Carlo Markov Chain (MCMC) method proposed by Farooq et al. [25] served as small motivation for this study. The authors attempt to circumvent the difficulties involved in explicitly finding and sampling from the joint distribution among people's characteristics. They successfully use MCMC with a Gibbs sampler to produce synthetic population with significantly higher accuracy than competing methods, including IPF. However, the method is quite complex, and may not be easily extendable to different geographies. This study aims to directly find the joint distribution, but in a form where sampling from it would be easy and feasible. Additionally, the proposed method is transferable across geographies, and therefore has a wider application. A richer review of population synthesis techniques for microsimulation can be found in works by Choupani and Mamdoohi [17], Ma [75], Müller and Axhausen [78], and Pritchard [94].

Chapter 4: Description of Data Used for Population Synthesis

There are two main datasets used for population synthesis: the American Community Survey (ACS) and the decennial U.S. Census data.

4.1 American Community Survey

The sample data used are from the 2016 release of the 1-year Public Use Micro Sample (PUMS) data, by ACS, for the state of Maryland. ACS collects detailed data on 1 % of the Census population in every state, each year. These detailed data, after cleaning, filtering, anonymization, and weighting are made available to the public. The data are geocoded to Public Use Micro Area (PUMA), which enclose no less than 100 000, and no more than 200 000 people [10]. The PUMA usually contains complete census tracts within, yet can span over multiple counties in a state. Maryland has 44 PUMAs, and 59 408 observations in total, which represent the about 6 million residents. In this dissertation, the copulas were constructed using observations from each PUMA independently of others. The constructed copulas were then used to generate population for the census tracts within the PUMA. The results are presented only for PUMA 1107, chosen at random, for the sake of brevity.

The ACS forms the backbone of many studies involving sample data, as it provides detailed information on many characteristics of people. Some information reported in each record by ACS is also captured in the census data, as the ACS is the present

manifestation of the long-form census, now discontinued. The primary limitation in ACS data collection is the large number of missing item response. This occurs when respondents do not answer one or more questions, or answer them incorrectly, or select the equivalent of "I do not wish to answer". Such missing item responses are imputed by borrowing information from other relevant data available in the region by experts at the Census Bureau. An indicative flag is provided with the characteristics to distinguish reported and imputed values. The priority of such imputation is to conform ACS totals to the regional values, rather than interrelation among attributes. Consequently, the dependence structure in the data is altered. However, in this study, imputed and reported values are treated equivalently, because they are the best publicly available representation of the population.

The ACS data is available in two sets, at the household level, which forms the survey unit, and for all persons within the selected household. The datasets provide a key to join both household and person data together, creating a complete dataset. The person dataset contains information about the person, like age, gender, race and so on, while the household file contains information about all residents in a household, like number of people, size and type of the house, total family income, presence of children, and others. The joined data are used in this study to construct copulas. This ensures the dependence of the entire household is captured, and then recreated. Usually, decisions made by people are in consensus with other members of the household, therefore, it is imperative that persons be recreated into complete household units for use in microsimulation models, especially choice based models.

Following on the research conducted by Arentze and Timmermans [3], Bradley, Bowman, and Griesenbeck [8], Guo and Bhat [43], Javanmardi, Auld, and Mohammadian [51], Pinjari et al. [92], Salvini and Miller [103], and Ye et al. [134] in population synthesis, the characteristics currently used from the ACS in this study are provided below.

1. Number of people in a household (household characteristic),
2. Type of household (household characteristic),
3. Total household income (household characteristic),
4. Presence and age of children (household characteristic),
5. Number of workers in the household (household characteristic),
6. Age of people in household (person characteristic),
7. Gender of people in household (person characteristic),
8. Race of people in household (person characteristic), and
9. Employment status of the people in household (person characteristic).

The dataset, therefore, has 9 dimensions, and only copulas that can be generalized to 9 dimensions are considered. At present such copulas are Clayton, Frank, Joe and Gumbel families from the Archimedean family of copulas [81]. Although elliptical copulas, like the normal and t copulas are also feasible with 9 dimensions, their goodness of fit test takes an unacceptably large amount of time, due mainly to the large parameter space, and degrees of freedom [49]. Moreover, as Kao et al. [59] demonstrate, the data are unlikely to be well represented by normal or t distributions. In this study, the characteristics are taken together in a single dataset and the person weights are used to duplicate each record to form the full population dataset for each PUMA, as advised in the ACS documentation [10].

4.2 U.S. Decennial Census

The census is a decennial survey, that aims to enumerate and survey every individual, as required by the constitution. However, it only captures basic information about

people, and the data release is in aggregate form. Census tracts are widely used in transportation as the finest geographic regions for which census aggregates are used. The aggregated totals are available by attributes of each characteristic collected, most of which overlap with the ACS data. Therefore census aggregates form the marginals of the characteristic at the census tract level [9]. While public consumption census data are reported for geographic regions smaller than census tracts, this study synthesizes population at the census tract level.

The census data in the form of totals are publicly available for the census tracts, which are the target areas for the synthetic population. The data are the totals of collected characteristics about people, such as number of people of a given age, or race, or gender, etc. in a census tract. These data provide the frequencies of people in each given attribute of a characteristic, and can be easily converted into the cumulative mass function (cmf). These cmf for each characteristic forms the marginal distribution at the census tract level, which is the target area in this study. The marginal distributions are used to inverse integral transform the data produced by the copulas. The census data used here are obtained from the 2010 Census Summary File 1 dataset.

Chapter 5: Developed Methodology for Population Synthesis

The methodology and algorithm proposed in this dissertation are illustrated using simulated data which are created so as to mimic ACS PUMS data. Using simulated data provides two main advantages:

1. a control benchmark to assess the performance of the proposed method, and
2. an unbiased presentation of the method.

Although the proposed methodology and algorithm are applied to data from all 44 PUMA in the ACS PUMS dataset for Maryland, results obtained with PUMS data from only one randomly selected PUMA are presented in this dissertation. To ensure simulated data accurately and robustly mimics PUMS data, they are produced to have the same shape as data from the selected PUMA. Shape here refers to the number of rows and columns in the ACS PUMS data, after applying the provided person weights as a frequency multiplier, that is, the sum of all weights. However, the dependency and marginal distributions are pre-specified to act as controls and help verify the method.

5.1 Data Simulation

The simulated data are generated using a Gumbel copula, randomly chosen among other Archimedean copulas, and is defined by the distribution function

$$\exp\left\{-\left[\sum_{j=1}^{d}(-\ln u_j)^{\theta}\right]^{\frac{1}{\theta}}\right\}, \tag{5.1}$$

where

θ is the copula parameter, and

$u_j \in (0,1),\ j = 1,\ldots,d$ [see table 4.1 in 81, for distribution functions of all one-parameter Archimedean families].

The admissible parameter space of the Gumbel copula is $\theta \in [1,\infty)$. With $\theta = 1$ the Gumbel copula corresponds to the independence copula, which is inappropriate for social-science joint categorical, ordinal, and numeric data.

Simulated data are produced iid from the Gumbel copula by setting the parameter $\theta = 1.500$. The mixture models due Marshall and Olkin [77] can then be used to generate observations having dependence characterised by the selected member. Since 9 dimensions reported in ACS are used in this study, the simulated data produced are also in 9 dimensions. Further, the number of observations in the simulated data are 118 583, the same as available from the PUMA demonstrated in the next chapter. First few observations of these simulated data, rounded to three significant digits, are shown in table 5.1, where the column headers $V1,\ldots,V9$ refer to the 9 dimensions as Variable1 to Variable9.

Since these simulated data inherently have continuous marginals, simple binning like,

$$f(u) = v_i \text{ if } u \in (a_i, a_{i+1}], \tag{5.2}$$

where $\cup_i (a_i, a_{i+1}] = (0,1]$, is used to introduce ties and thus discretize the data.

Table 5.1: Observations of Simulated Data from Gumbel Copula

V1	V2	V3	V4	V5	V6	V7	V8	V9
0.000	0.225	0.279	0.062	0.248	0.001	0.043	0.060	0.538
0.763	0.929	0.605	0.828	0.750	0.584	0.926	0.656	0.663
0.601	0.377	0.252	0.592	0.742	0.236	0.600	0.157	0.783
0.882	0.967	0.773	0.803	0.909	0.918	0.876	0.814	0.955
0.966	0.901	0.923	0.937	0.989	0.851	0.883	0.888	0.794
0.943	0.642	0.897	0.565	0.629	0.863	0.790	0.725	0.439
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Choosing $a_i = 1/k$, one can create k evenly spaced bins. Here we set $k = 10$, and the value $v_i = i,\ i \in \{1, \ldots, 10\}$ is the label of the bin, chosen such that $v_i < v_j$ if $i < j$. The simulated data thus has pre-specified dependence and marginal structure, yet conforms to the shape of the real data. It is not expected that the features of this simulated data are exactly like the survey data, but are similar in discreteness, number of rows and columns, embedded dependence, and marginal distributions. The binned data by applying (5.2), for the same observations shown in table 5.1, are presented in table 5.2. These observations are similar to the data available in PUMS. From here onward in this dissertation, the proposed method is illustrated using simulated data presented in table 5.2. The same method is directly transferable and applied to PUMS data from any PUMA.

Table 5.2: Observations of Binned Simulated Data from Gumbel Copula

V1	V2	V3	V4	V5	V6	V7	V8	V9
1	3	3	1	3	1	1	1	6
8	10	7	9	8	6	10	7	7
7	4	3	6	8	3	7	2	8
9	10	8	9	10	10	9	9	10
10	10	10	10	10	9	9	9	8
10	7	9	6	7	9	8	8	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

5.2 Probability Integral Transformation using Empirical Marginal Distributions

The first step in constructing the copula is probability integral transformation of the marginal components of the data. The data in table 5.2, and similarly PUMS data from any PUMA, need to be probability integral transformed into the unit hypercube, $[0,1]^d$. This is necessary, as the support of the copula lies in the unit hypercube [34, 67]. The rank-based estimate of the marginal distribution is obtained by applying (2.6) to each component of the multivariate data. The transformed pseudo-observations from the simulated data are presented in table 5.3, for the data presented in table 5.2.

Table 5.3: Pseudo-Observations of Binned Simulated Data from Gumbel Copula

V1	V2	V3	V4	V5	V6	V7	V8	V9
0.050	0.252	0.251	0.049	0.250	0.050	0.050	0.050	0.550
0.750	0.951	0.652	0.851	0.751	0.553	0.950	0.651	0.650
0.651	0.353	0.251	0.550	0.751	0.251	0.652	0.151	0.751
0.850	0.951	0.752	0.851	0.950	0.950	0.851	0.851	0.951
0.950	0.951	0.951	0.950	0.950	0.851	0.851	0.851	0.751
0.950	0.652	0.852	0.550	0.651	0.851	0.752	0.751	0.450
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

5.3 Likelihood Maximization

The likelihood maximization is the second step in constructing copulas with parameter θ. As fully expounded by Genest, Ghoudi, and Rivest [34], the likelihood specification of copulas is fairly straightforward. It is the product of the probability densities or masses evaluated at realizations of iid copies of the random variable. Natural logarithm transformation is applied to the likelihood function to prevent cancellation due to rounding errors, and to simplify computations. The likelihood specification

then becomes a log-likelihood,

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \left[c_\theta(\mathbf{u}_i) \right], \tag{5.3}$$

where

n is the number of observations $\mathbf{u}_i$ in $(0,1)^d$ (obtained by applying (2.6) to dataset $\mathcal{X}$),

θ is the copula parameter, and

c_θ is the multivariate density of the parametric copula for continuous data, or the converted probability masses for discrete data (see Nelsen [81] for density or equivalent mass distributions of various copula families).

Maximizing the log-likelihood yields an unbiased estimate $\hat{\theta}_n$ (we assume the maximum is unique, usually the case with single parameter Archimedean copulas) of the associated parameter θ:

$$\hat{\theta}_n := \arg\max_{\theta} \mathcal{L}(\theta), \tag{5.4}$$

and we denote $C_{\hat{\theta}_n}$ the copula built with the parameter estimator to differentiate it from the generic parametric copula C_θ.

As n can be very large, the computation of (5.3) can be quite expensive. This cost can be significantly reduced by avoiding computations of the same terms several times. Same terms occur as there are identical observations in the dataset. Let w_i denote the number of observations identical to $\mathbf{u}_i$ and for each set of identical observations, elect a representative $\hat{\mathbf{u}}_i$, for instance by choosing $\hat{\mathbf{u}}_i = \mathbf{u}_j$, with $j = \arg\min_k \mathbf{u}_k$ such that $\mathbf{u}_k = \mathbf{u}_i$. We then have a set of n^* unique observations, and (5.3) can be replaced by

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n^*} w_i \log \left[c_\theta(\hat{\mathbf{u}}_i) \right]. \tag{5.5}$$

Note that w sums to n yet has n^* elements, and therefore, (5.5) is normalized by

n. This is essentially similar to, but not the same as, the data and corresponding weights provided in PUMS. In this case, the weights w are computed after producing the pseudo-observations by applying (2.6) so that the ranking is not disturbed. This compression is applied only during log-likelihood estimation, and results in significant computational speed-ups. The datasets compressed during log-likelihood estimation also include the ones generated during the parametric bootstrap.

As the copula family that truly characterises the dependence is unknown, multiple copula families are tried to find a member that best describes the dependence in the data. The estimated parameter essentially selects a member that captures the dependence structure of the given data from the chosen family of copulas. Parameters estimated for Clayton, Frank, Gumbel and Joe families of copulas on the simulated pseudo-observations shown in table 5.3 are presented in table 5.4. It is to be noted that the estimated Gumbel copula parameter almost recovers the original parameter of 1.500.

Table 5.4: Parameters Estimated for Archimedean Copula Families

Copula Family	Estimated Parameter
Clayton	0.761
Frank	3.216
Gumbel	1.543
Joe	1.861

5.4 Goodness of Fit Testing

As discussed with (2.2), parametric copulas can be viewed as families covering part of the monotonicity curve between the Fréchet–Hoeffding bounds, each member of which is identified by the copula parameter. Further, the two-step maximum likelihood estimation procedure of Genest, Ghoudi, and Rivest [34] requires presumption of the copula family. Therefore, it is necessary to evaluate how well the chosen copula family

describes the dependence in the multivariate random variable being modeled.

Let $\mathcal{C}$ denote the chosen family. The null hypothesis is that the parametric copula C_θ that characterizes the dependence is a member of the assumed family as given by

$$H_0 : C_\theta \in \mathcal{C} := \{\mathsf{C}_\theta : \theta \in \mathcal{O} \subseteq \mathbb{R}^p\}, \tag{5.6}$$

where

θ takes any value from the admissible set $\mathcal{O}$, and

p is an integer, such that $p \geq 1$, representing the dimensionality of the parameter.

The admissible space is dependent on the family of copula, which may not always attain the Fréchet–Hoeffding bounds [see table 4.1 in 81, for Archimedean copulas].

Genest, Rémillard, and Beaudoin [40] examine the power of the most common method of testing the null hypothesis using statistics based on the empirical process

$$\mathbb{C}_n = \sqrt{n}\left(C_n - C_{\hat{\theta}_n}\right), \tag{5.7}$$

where

C_n denotes the empirical copula, and

$C_{\hat{\theta}_n}$ is derived under H_0 [39, 40].

The empirical distribution of multivariate random variables was originally given by Deheuvels [21]. An asymptotically equivalent and simpler to compute distribution is

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\mathbf{u}_i \leq \mathbf{u}); \quad \mathbf{u} \in [0,1]^d, \tag{5.8}$$

where the inequalities are evaluated component-wise [40]. This empirical distribution is piece-wise constant and thus not a copula, yet is often called the empirical copula, and this paper will continue that nomenclature. The empirical copula is an unbiased

and consistent estimator of the joint distribution under various conditions, as examined by Genest and Rémillard [39]. Therefore it forms the most objective benchmark to judge the parametric copula [40].

Genest, Rémillard, and Beaudoin [40] show empirically that the Cramér–von Mises statistic, which quantifies separation between two distributions, is one of the most powerful tests for evaluating the similarity of distributions produced by the empirical copula and the estimated parametric copula. Cramér–von Mises statistic expressed as a continuous functional of (5.7) becomes

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\mathbf{u})^2 \mathrm{d}C_n(\mathbf{u}) = \sum_{i=1}^{n} \left(C_n(\mathbf{u}_i) - C_{\hat{\theta}_n}(\mathbf{u}_i) \right)^2 . \tag{5.9}$$

The straightforward way to understand (5.9) is to note that Cramér–von Mises statistic is like a "distance" between the distribution obtained from the parametric copula and the distribution obtained by an unbiased empirical distribution. The null hypothesis is rejected with large values of the statistic (5.9) as $n \to \infty$ [35, 39, 40, 69, 70, 90].

5.4.1 Parametric Bootstrap

Genest and Rémillard [39] show that as $n \to \infty$, under H_0, S_n converges weakly to a limiting distribution and that their tests based on S_n are consistent; that is, H_0 is almost surely rejected if $C_\theta \notin \mathcal{C}$. The limiting distribution of S_n is unknown, but can be approximated using parametric bootstrap as in Genest, Rémillard, and Beaudoin [40]. In their extensive Monte Carlo experiments, both Genest, Rémillard, and Beaudoin [40] and Kojadinovic [66] show that the statistic S_n holds the most power in rejecting H_0. Let $S_n^{(k)}$ be the k^{th} bootstrap replicate of the statistic. An approximate p-value can be computed as

$$\frac{1}{K+1} \left[\sum_{k=1}^{K} \mathbf{1} \left(S_n^{(k)} \geq S_n \right) + 0.5 \right], \tag{5.10}$$

where K is the total number of bootstrap replications, assumed to be large enough. Here, we set $K = 1\,000$. $0.5/(K+1)$ is an asymptotically insignificant adjustment factor used to ensure the p-values are always in the interval $(0, 1)$. A p-value above a statistical level of significance indicates that H_0 cannot be rejected. Therefore the observed data might be described by the parametric copula C_θ, which is generically referred to as 'well-fitting copula' in the succeeding discussion.

Since the data generated from the constructed copula have continuous marginals, but the marginals of the survey data used are discrete, the proofs by Genest and Rémillard [39] are no longer valid, and may produce biased inferences about H_0 [36, 38, 66, 68, 89]. However, Kojadinovic [66] adapted the parametric bootstrap procedure of Genest, Rémillard, and Beaudoin [40] to the presence of ties in the data. The modified algorithm is presented in Procedure 6.1 in the paper by Kojadinovic [66], and is shown to be consistent despite ties in the data. Goodness of fit to the data of any copula presented in this study is assessed using 1 000 parametric bootstrap replications as presented in algorithm 1. Readers are referred to the paper by Kojadinovic [66] for implementation and additional details about the procedure.

The Cramér–von Mises statistic, S_n, and p-values for the simulated data obtained by applying algorithm 1 are presented in table 5.5. The statistic and p-value for the independent copula are also presented as a baseline. The results clearly indicate that the independent copula can be rejected, and has the largest value of the Cramér–von Mises statistic. Moreover, the p-values indicate that both the Clayton and Frank families can be rejected. The Gumbel and Joe families cannot be rejected as they score very high p-values.

Despite knowing that the simulated data were produced using the Gumbel copula, statistical discrimination between copula families for which H_0 is not rejected is tricky as it involves computing a measure of similarity between two multivariate distributions. Training dataset formed from real world data, or simulated data, used to construct the

Algorithm 1 The Parametric Bootstrap

Inputs:
$\mathbf{u}_1, \ldots, \mathbf{u}_n$: n, d-dimensional observations, assumed to be realizations of an (unknown) random vector (obtained by applying (2.6) to realizations of a random vector $\mathbf{x}$)

Initialize:
Assume a parametric copula family $\mathcal{C}$
$K \leftarrow 1\,000$ (or a sufficiently large number)

Construct a copula $C_{\hat{\theta}_n}$, an estimate of $C_\theta \in \mathcal{C}$ using $\hat{\theta}$ estimated by (5.4) over the n observations $\mathbf{u}_1, \ldots, \mathbf{u}_n$

Compute S_n from C_n and $C_{\hat{\theta}_n}$ using (5.9)

for $k := 1$ **to** K **do**

$\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_n^{(k)} \leftarrow n$ observations in d dimensions drawn from $C_{\hat{\theta}_n}$ using the procedure developed by Marshall and Olkin [77]

$\mathbf{w}_1^{(k)}, \ldots, \mathbf{w}_n^{(k)} \leftarrow \mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_n^{(k)}$ modified as per Step 4(b) of Procedure 6.1 proposed by Kojadinovic [66]

Compute $\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_n^{(k)}$ from $\mathbf{w}_1^{(k)}, \ldots, \mathbf{w}_n^{(k)}$ using (2.6)

Construct a copula $C_{\hat{\theta}_n}^{(k)}$, an estimate of $C_\theta \in \mathcal{C}$ using $\hat{\theta}^{(k)}$ estimated by (5.4) over $\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_n^{(k)}$

Compute $S_n^{(k)}$ from $C_n^{(k)}$ and $C_{\hat{\theta}_n}^{(k)}$ using (5.9)

end for

Compute the approximate p-value using (5.10)

Table 5.5: Parameters Estimated for Archimedean Copula Families

Copula Family	Estimated Parameter	Cramér–von Mises Statistic	p-value
Independent	NA	2 421.214	0.001
Clayton	0.761	1 024.078	0.001
Frank	3.216	587.086	0.001
Gumbel	1.543	282.586	0.993
Joe	1.861	274.518	0.999

copula can be considered as observations from some unknown multivariate random distribution. Any validation dataset, not used to construct the copula, would also have observations resulting from this unknown random distribution. Therefore, there would be no statistical difference between the training and validation datasets.

On the other hand, multivariate data generated from a copula by the mixture models due Marshall and Olkin [77] are also ultimately realizations from a defined multivariate random distribution. If the assumed copula family captures the unknown random distribution producing the training and validation observations, data produced from the copula should be statistically indistinguishable from either datasets. Quantifying this requires comparison between multivariate random variables. Comparative statistics based on any empirical process, like (5.7), are not distribution free, requiring tools like the bootstrap to resolve [27, 39, 40]. Finally, recall that from Sklar's representation theorem, in presence of discrete marginals, more than one copula can correctly represent the underlying multivariate distribution, so it is ultimately often illusionary to isolate one copula only. In chapter 7, additional discussion is presented as a starting point in the development of tools to improve the tests conducted to assess the goodness of fit of copulas to provided multivariate data.

5.5 Inverse Probability Integral Transformations of Marginals

The goal of the paper is to produce synthetic population at some geographic area, which could be the same area where the copulas are fit, or for another area where the dependence captured by the copula is transferable. In this dissertation, well-fitting copula to PUMS data from a PUMA is used to generate the population at finer geopolitical division: the census tracts enclosed by the PUMA. The dependence structure of each enclosed census tract is assumed to be the same as that of the PUMA as a whole. This assumption is reasonable because the census bureau bases the ACS sample design on the decennial census and other datasets, while also designing the

next census based on the rolling ACS data collection. Further, the relation between different demographic properties of people should remain similar across space. As an example, consider the relation between age, education, job industry, salary, household size and location, number of owned cars, and family structure. It is easy to perceive the relationship between the characteristics, and similarly deduce that it would be applicable to other locations as well.

Any well-fitting copula can be used to generate random data with the same dependence structure at the target geographic area as the data used to construct it using the mixture models by Marshall and Olkin [77]. These data will be in the space of the support of the copula, i.e., uniform in $[0,1]^d$. However, a synthetic population can be obtained by inverting the marginal draws from the copula. Suppose the individuals in a census tract are realizations of m iid copies of a d-dimensional multivariate random vector $\mathbf{Y} = (Y_1, \dots, Y_d)$. Let $G_1, \dots, G_d$ denote the cumulative mass functions of each of the d characteristics at the census tract level. Further, let $\mathbf{v} = (v_1, \dots, v_d)$ be generated from the constructed copula. Synthesis of the population is then given by

$$\hat{\mathbf{y}} = (G_1^{-1}(v_1), \dots, G_d^{-1}(v_d)), \tag{5.11}$$

where $G_1^{-1}(v_1), \dots, G_d^{-1}(v_d)$ are the pseudo-inverses of the cumulative mass functions $G_1, \dots, G_d$, defined as

$$G_j^{-1}(v) = \inf\{y_j \mid G_j(y_j) \geq v\}; \quad j \in \{1, \dots, d\}, \tag{5.12}$$

Algorithm 2 presents the whole process of the inverse transformation.

Some dimensions from ACS PUMS used in this study are not available in the decennial census marginal data (table 6.4). Therefore, those dimensions have been inverse transformed using marginal cumulative mass functions computed from the ACS PUMS dataset for each PUMA. Further, since a single copula family with scalar

Algorithm 2 Inverse Probability Integral Transform of Data Generated from Well-Fitting Copula

1: **Inputs:**
   $C_{\hat{\theta}_n}$, a well fitting copula constructed with estimated parameter $\hat{\theta}_n$
   $(G_1, \ldots, G_d)$, a vector of cumulative mass functions describing the marginals at the required geopolitical level
   m the required number of records of inverse transformed data
2: **Initialize:**
   $i \leftarrow 1$
3: **while** $i \leq m$ **do**
4: $\mathbf{v}_i \leftarrow d$-dimensional realization drawn from the copula $C_{\hat{\theta}_n}$
5: $\hat{\mathbf{y}}_i \leftarrow (G_1^{-1}(v_{i1}), \ldots, G_d^{-1}(v_{id}))$
6: **if** $\hat{\mathbf{y}}_i$ is acceptable **then**
7: Add $\hat{\mathbf{y}}_i$ to output dataset
8: $i \leftarrow i + 1$
9: **else**
10: Reject $\hat{\mathbf{y}}_i$
11: **end if**
12: **end while**

parameters was used to capture the dependence of the whole dataset, relative differences in strengths of association between various dimensions in the dataset are averaged out. Consequently, some sanity checking of the data produced after inverse transformation is required (see the condition in step 6 in algorithm 2). Quality controls are actually applied at two instances in the process, the first after inverse transforming the data dimensions available in the census data, and the second after inverse transforming the remaining dimensions using ACS PUMS data.

Chapter 6: Analysis of Synthesized Population

The results from the randomly selected case study PUMA are presented in this section. Figure 6.1 shows a map with the location of the PUMA, along with enclosed census tracts, and place names. The PUMA is from the south-eastern side of Washington DC, where the state lines of Washington DC, Maryland and Virginia meet. This PUMA is home to about 120 000 people, and is a combination of low density suburbs, high density housing, many retail zones, and some business centers.

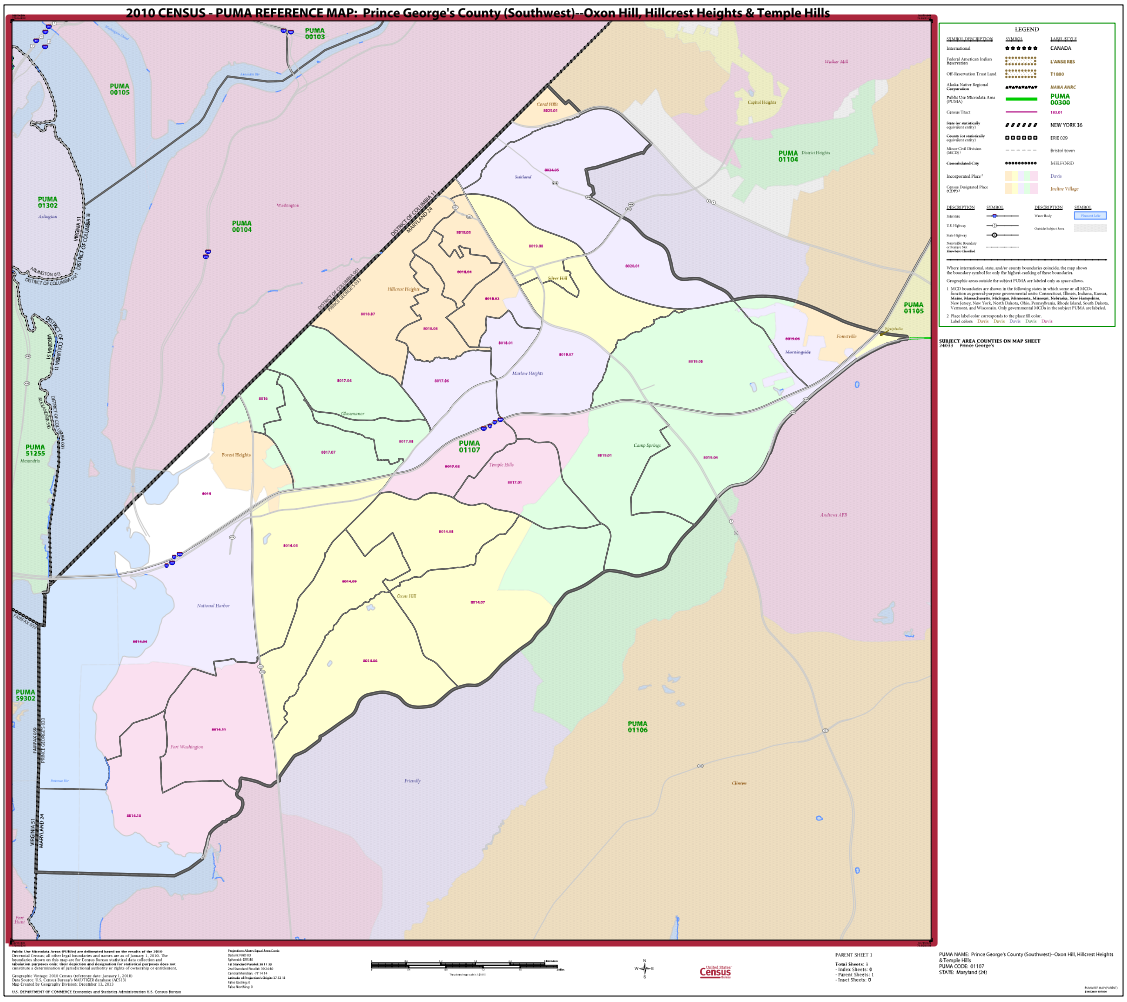


Figure 6.1: Map of PUMA 1107

6.1 Copula Construction

The dataset of PUMA 1107 is formed by joining the household and person sets using the key provided in ACS PUMS data [10]. A total of 118 583 observations were available from this PUMA, after duplicating the records by the provided ACS person weights [as recommended by 10]. A sample of the data are shown in table 6.1. This table is similar to table 5.2 of simulated observations. Pseudo-observations produced using equation (2.6) for data in table 6.1 are shown in table 6.2, which is similar to table 5.3.

Table 6.1: Observations from Training Dataset for PUMA 1107

NP	HHT	HINCP	HUPAC	WIF	AGEP	SEX	ESR	RAC1P
2	3	160510	4	1	53	2	1	2
2	3	160510	4	1	53	2	1	2
2	3	160510	4	1	53	2	1	2
2	3	160510	4	1	53	2	1	2
2	3	160510	4	1	53	2	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 6.2: Pseudo-Observations from Training Dataset for PUMA 1107

NP	HHT	HINCP	HUPAC	WIF	AGEP	SEX	ESR	RAC1P
0.269	0.640	0.865	0.717	0.392	0.708	0.724	0.455	0.496
0.269	0.640	0.865	0.717	0.392	0.708	0.724	0.455	0.496
0.269	0.640	0.865	0.717	0.392	0.708	0.724	0.455	0.496
0.269	0.640	0.865	0.717	0.392	0.708	0.724	0.455	0.496
0.269	0.640	0.865	0.717	0.392	0.708	0.724	0.455	0.496
0.269	0.640	0.865	0.717	0.392	0.708	0.724	0.455	0.496
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The column headers in tables 6.1 and 6.2 are the same as field names in the 2016 ACS PUMS Data Dictionary. The header abbreviations are expanded as follows:

1. NP: Number of person records following the housing record,

2. HHT: Household or family type,

3. HINCP: Household income (past 12-months),

4. HUPAC: Household presence and age of children,

5. WIF: Workers in family during the past 12-months,

6. AGEP: Age of the person,

7. SEX: Sex of the person,

8. ESR: Employment status of the person (recode),

9. RAC1P: Recoded detailed race code.

Four Archimedean copula families, Clayton, Frank, Gumbel and Joe, were fitted to this dataset using the log-likelihood maximization method. Goodness of fit was then tested for each family. Goodness of fit of the independent copula was also evaluated for baseline comparison. Additionally, standard errors for each parameter estimate were computed using the 80 replicate weights provided with PUMS data. The standard errors are given as per the ACS recommended procedure [10] by,

$$SE(\hat{\theta}) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}\left(\hat{\theta}_r - \hat{\theta}\right)^2}, \tag{6.1}$$

where

$\hat{\theta}_r$ is the parameter estimated using the r^{th} replicate weights, and
$\hat{\theta}$ is the parameter estimated using the full sample weights.

Note that any negative replicate weights were truncated at 0. The parameter estimates with standard errors, Cramér–von Mises statistics and p-values of the copula families are presented in table 6.3.

Table 6.3: Constructed Archimedean Copulas for PUMA 1107

Copula	Parameter	Standard Error	CvM Statistic	p-value
Independent	NA	NA	261.520	0.001
Clayton	0.098	0.022	220.655	0.999
Frank	0.113	0.079	255.123	0.001
Gumbel	1.059	0.015	233.572	0.999
Joe	1.083	0.018	243.633	0.999

Table 6.3 clearly indicates that the null hypothesis must be rejected for the independent copula, which enforces that there is some dependence between the 9 dimensions of the PUMS data used in this study. Frank copula is likewise rejected as it cannot faithfully capture the dependence in the data. The other three copulas — Clayton, Gumbel and Joe — describe the data from PUMA 1107 equally well, and any of them would serve the purpose of generating the synthetic population. This is further supported by additional Monte Carlo simulations presented in Appendix A. Joe copula is arbitrarily picked and used to generate synthetic population for census tracts enclosed in the PUMA (see figure 6.1 for details). The estimated parameters are stable over the replicate weights as reflected by the small standard errors, suggesting that the number of observations is large enough for the purpose of copula calibration.

6.2 Population Synthesis

Population was generated for the census tracts enclosed by each PUMA. In this section, results from two randomly selected census tracts are presented. The decennial census data contains aggregate totals at the census tract level for six of the nine characteristics used from the ACS PUMS. Table 6.4 presents a crosswalk between the characteristics from ACS PUMS and the corresponding decennial census matrix from the Census Summary File 1 Dataset [9, 10].

Note that for some matrices, especially P19 and P20, only selected columns were used such that the data description of the categories in the decennial census data

Table 6.4: Crosswalk between ACS PUMS and Decennial Census Matrix

ACS PUMS Characteristic		Census Summary File 1 Matrix
NP	$\longmapsto$	H13
HHT	$\longmapsto$	P19, H13 for vacant properties
HINCP	$\longmapsto$	No matching matrix
HUPAC	$\longmapsto$	P20, H13 for vacant properties
WIF	$\longmapsto$	No matching matrix
AGEP	$\longmapsto$	PCT12
SEX	$\longmapsto$	PCT12
ESR	$\longmapsto$	No matching matrix
RAC1P	$\longmapsto$	P3

dictionary matches the description in the ACS PUMS data dictionary. Additionally, summations were used to club the numbers from different columns in the decennial census matrices to match the ACS PUMS characteristics — for example, columns in P20 for HUPAC, and in PCT12 for AGEP. In cases like RAC1P and P3 (also, NP and H13), not all categories from the ACS PUMS data are available in the census matrix. However, these differences in categories do not cause any problems while inverse transforming data produced from copulas using the marginals from census and ACS PUMS. The final values are always in the domain of the marginal mass functions.

To derive the quantile marginal distributions, the decennial census data for a characteristic (or dimension) from a census tract are converted into a cumulative mass function. The plateaus of the mass distribution define the bins of the quantile function. All data generated by the copula, for the characteristic (i.e., for the given dimension), that fall in a bin of the quantile function are assigned the value of the bin. This is similar to using function (5.2), except the bin edges, a_i and a_{i+1}, are informed by the quantile function. Characteristics without a counterpart in the decennial census data — HINCP, WIF and ESR — are transformed using marginals obtained from the ACS PUMS data for the whole PUMA, thus demonstrating that all marginal distributions need not come from the same source.

Algorithm 2 was used to produce synthetic population for all census tracts in PUMA 1107. To maintain brevity, results from two randomly selected census tracts are presented. These census tracts are 8016 and 8017.01, with considerably different characteristics (see figure 6.1). Tract 8016 is part of the Glassmoor area, and shares a border with Washington DC. It is characterized by high density apartment and condominium style homes. Approximately, census tract 8016 has as many residents as households. On the other hand tract 8017.01 is part of Temple Hill, a suburb of Washington DC and has low density housing, characteristic of suburban sprawl. The average number of people per house in tract 8017.01 is about 2.5. Majority of the people in either tract are African Americans.

As previously discussed, a rigorous out-of-sample validation of the data produced by the copulas is not straightforward. Future work would be devoted to the investigation of additional tests comparing multivariate distributions. It is nevertheless easy to empirically compare the univariate marginal frequencies of the population synthesized using copulas. Figures 6.2a and 6.2b show the distributions of percent errors of the population synthesized from the copulas. The errors are computed for each category in a given dimension (characteristic). They are the percent difference between the predicted frequency of the category by inverse transformed copula data, to the actual reported frequency in decennial census or ACS PUMS data. Mathematically it is

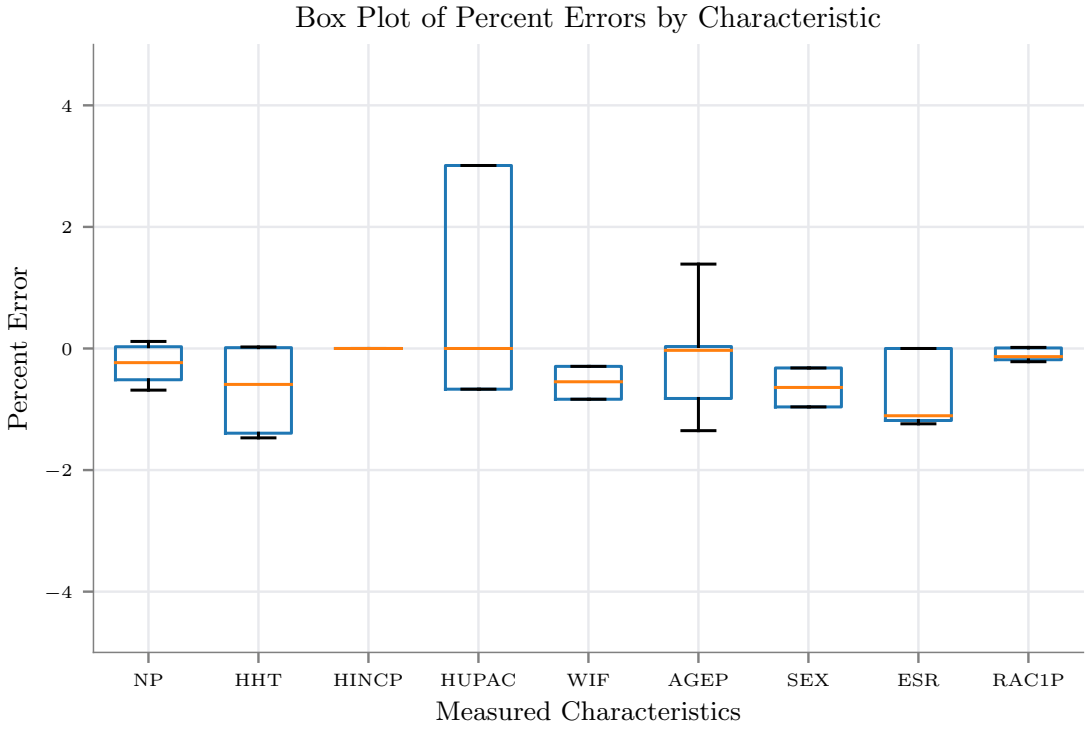
$$\hat{\varepsilon}_j := \frac{\hat{\nu}_j - \nu_j}{\nu_j} 100 \qquad \forall j \in \{1, \dots, d\}, \tag{6.2}$$

where

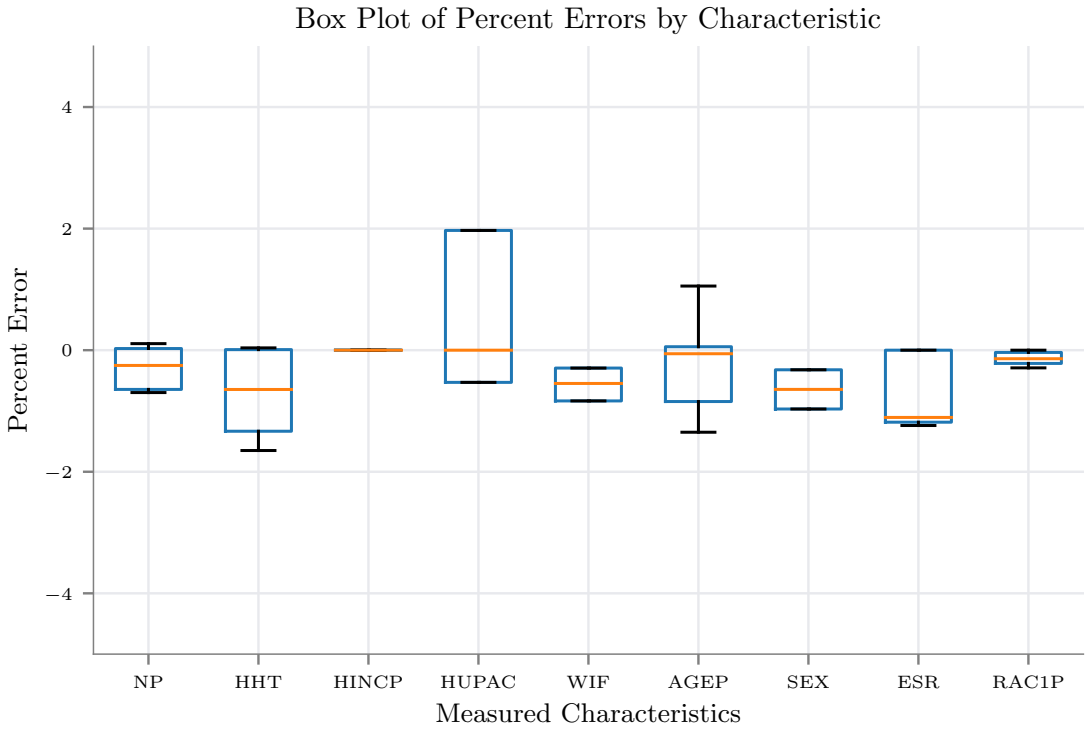
$\hat{\nu}_j$ is the vector of frequencies of each category in characteristic or dimension $j \in \{1, \dots, d\}$, and

ν_j is the vector of true frequencies from decennial census or ACS PUMS data.

Box plots in figure 6.2a and 6.2b are created from the errors for all categories in a



(a) Errors in Census Tract 8016



(b) Errors in Census Tract 8017.01

Figure 6.2: Errors in Synthetic Population Marginals

characteristic. In the box plot, the box represents the interquartile range ($[25\,\%, 75\,\%]$), while the median is shown by the bar within each box. The ends of the whiskers outside the boxes show the confidence interval ($[5\,\%, 95\,\%]$).

6.3 Discussion

The errors shown in figures 6.2a and 6.2b are very small, proving the accuracy of the copulas in reproducing the synthetic population marginals. The dependence between the dimensions is represented by a single parameter in Archimedean copulas, which captures the equivalent of an average dependence. Even though a single parameter offers protection against overfitting, some nuances of the dependence are missed. These missed nuances necessitate sanity checking the data after inverse transformation (as described in step 6 in algorithm 2).

Moreover, the very high and very low p-values also indicates some issues with the goodness of fit testing procedure. There are multiple reasons possible for this, including the sanity checking step which is not included within the parametric bootstrap. Another related reason is the change in the dependence structure by using the weights provided in ACS. However, by not using the weights, we work with a sample of the population which has sampling bias designed into it. Therefore, further exploration is required in this direction to understand the applicability of the well-studied inference processes to select copulas that represent the data at hand.

The missed details of the dependence also suggest exploration of more granular copula models. The presented models can also be expanded to include many more dimensions, and multiple data sources, producing a holistic modeling framework to capture multivariate characteristics and synthesize data that encapsulates those characteristics. Further, the search for reliable statistical techniques to compare discrete multivariate distributions is posed as an open question in this dissertation. This could provide alternate methods to discriminate between copula families or

hierarchies and proceed to out-of-sample validations, keeping in mind that for discrete observations, more than one copula can adequately capture the dependence structure. These four avenues provide interesting objectives for future research.

Chapter 7: Alternative Tests for Copula Goodness of Fit

As identified in chapters 5 and 6, and noted by Genest and Nešlehová [36] and Kojadinovic [66], the current copula goodness of fit tests based on Monte Carlo simulations have room for improvement when used with survey data, although they work reasonably well for simulated data. Therefore, in this chapter (see paper by Kaushik, Cirillo, and Bastin [60] for original work), we explore the tie adapted parametric bootstrap inference procedure for copula goodness of fit suggested by Kojadinovic [66] with survey data. Relation between the frequency and nature of ties and performance of the goodness of fit tests is yet to be elucidated. Theoretical results to that extent might be quite difficult to obtain given the underlying complexity of the models [37, 40, 66]. The consequent recommendation follows that any inferences from these goodness of fit tests should be adequately examined before use.

Fortunately, the survey sample data are available at hand to independently estimate the dependence measures with, say, the pairwise Kendall's rank correlations (τ). Data generated from a copula family using the parameter estimated by (5.4) can be used to produce an independent copy of population over the same area. Rank correlation between these synthetic data can be compared with the correlations found in the original survey data. Any values assumed by the copula outside the space of the Cartesian product of the marginals are immaterial and should not be used for any tests or inferences. A resulting complication for discrete data with many ties is the sharply reduced space over which the copula is valid. For example, consider that the space of any two categorical characteristics like gender and marital status

form very few unique points, regardless of the number of observations in the dataset. Consequently, the inference process of Genest and Rémillard [39], even adapted as Kojadinovic [66] suggests, breaks down. The only solution, at the moment, seems to be to increase the dimensionality of the dataset to enlarge the space forming the support of the copulas. Therefore, we rely on pairwise Kendall's rank correlation coefficient to compare the synthetic dataset with the original.

7.1 Data Preparation

9 dimensions from the ACS PUMS data are used in this study, chosen such that three groups can be formed each containing three dimensions. The dimensions within each group are chosen to have similar relation to each other. The average dependence of a group — obtained by computing the mean of the $d(d-1)/2$ unique pairwise dependence (lower triangle of dependence cross-tabulation) — classify each group into independent, weakly dependent and strongly dependent. The relations between dimensions in each group are estimated by Kendall's rank correlation coefficient, τ. The overall relation of a group is denoted henceforth as $\bar{\tau}$ [67].

Figure 7.1 provides the pairwise Kendall's rank correlation coefficient matrix for the three dimensions in each of the three groups. Note that pseudo-observations were generated for all data using (2.6) before computing the rank correlation coefficients. Race code was found to have countermonotone relation with age and gender. Therefore, the subtraction rule, which inverts the direction of the axis, was applied to the race code dimension to convert countermonotonous dependence to comonotonous (figure 7.1a). Group subcaption of the figures include the average correlation coefficient ($\bar{\tau}$) of the group in parenthesis. The dimension headers are acronyms, and are the same as the ACS data dictionary for each field. The description of each field are as follows:

1. Independent Dimensions:

(a) AGEP: Age of the person (discrete ordinal),

(b) SEX: Sex of the person (discrete categorical),

(c) RAC1P: Recoded detailed race code (discrete categorical).

2. Weakly Dependent Dimensions:

(a) VEH: Vehicles (1 ton or less) available (discrete ordinal),

(b) HINCP: Household income (past 12-months) (discrete ordinal),

(c) WIF: Workers in family during the past 12-months (discrete ordinal).

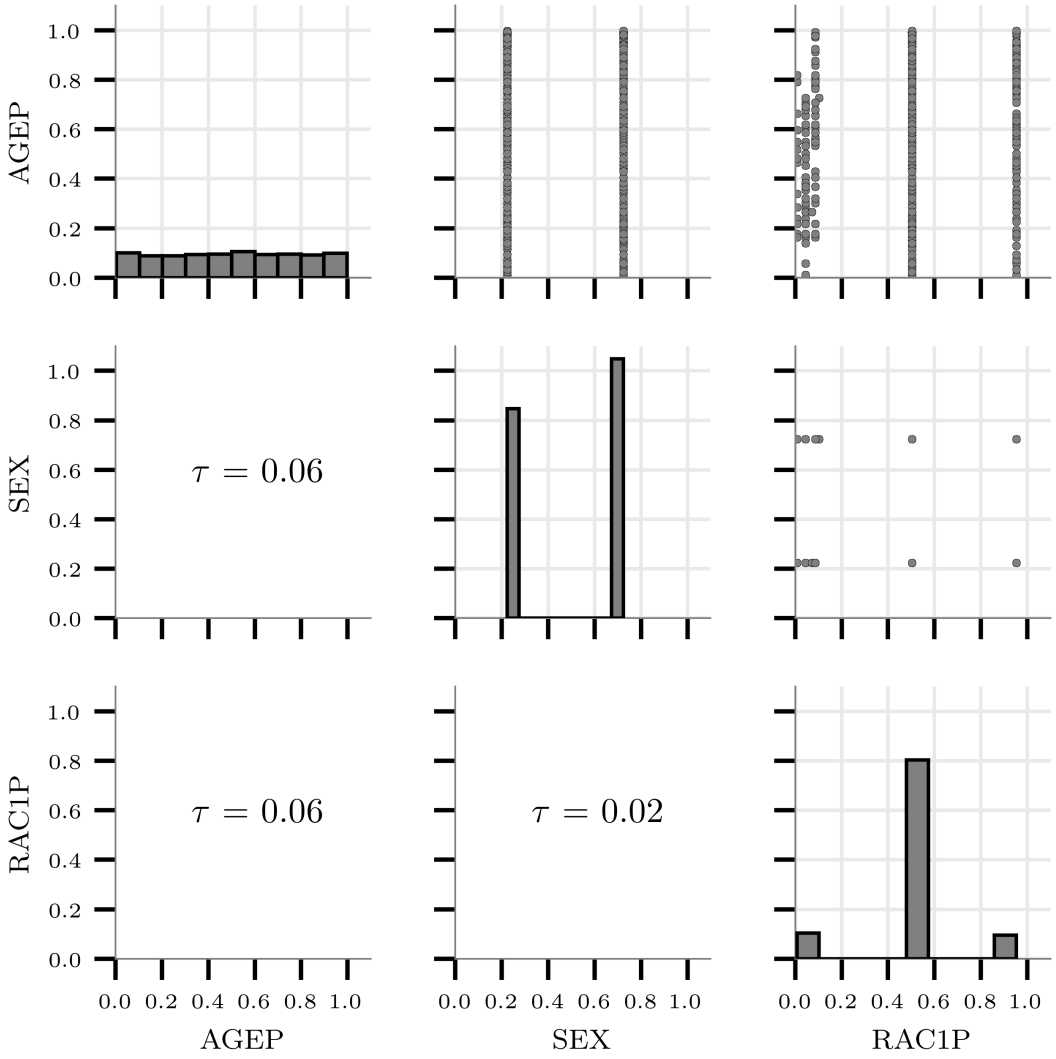
3. Strongly Dependent Dimensions:

(a) JWMNP: Travel time to work (discrete ordinal),

(b) JWRIP: Vehicle occupancy (discrete ordinal),

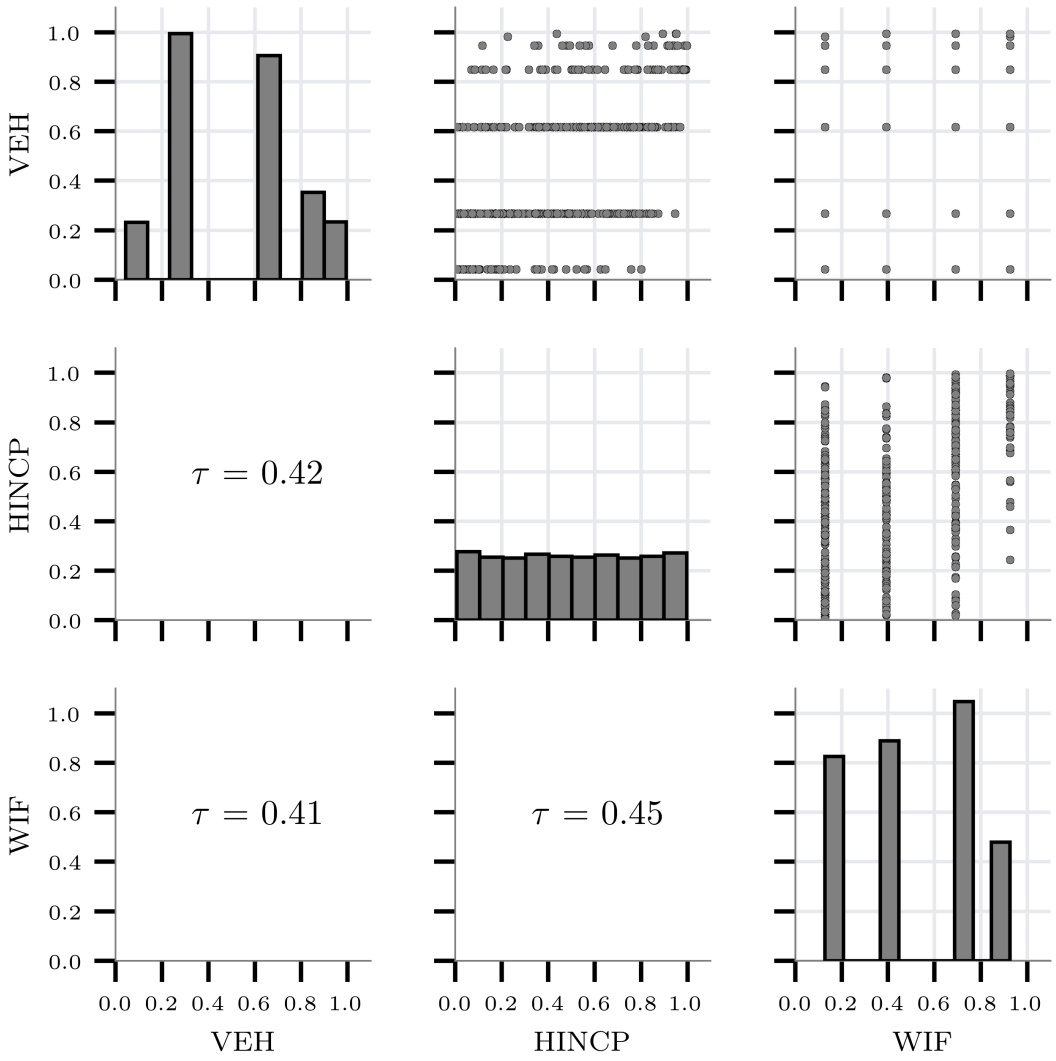
(c) JWTR: Means of transportation to work (discrete categorical).

[1]The pseudo-observations for dimension RAC1P (obtained by (2.6)) have been replaced by applying the subtraction rule: $1 - \hat{\mathbf{U}}_{RAC1P}$, so as to ensure the rank correlations with dimensions AGEP and SEX are positive.



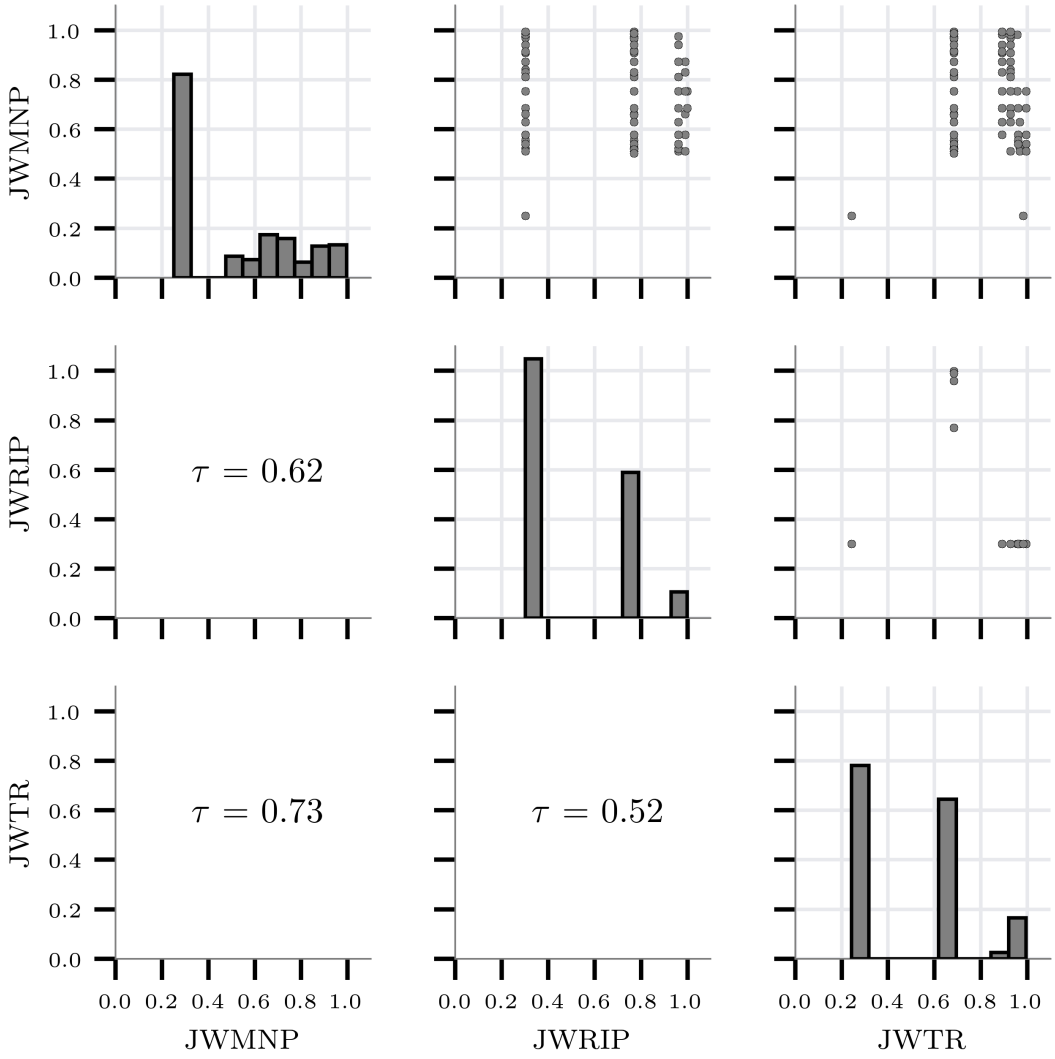
(a) Independent Dimensions ($\bar{\tau} = 0.050$)[1]

Figure 7.1: Kendall's Rank Correlation Among Dimensions in ACS PUMS for PUMA 1107



(b) Weakly Dependent Dimensions ($\bar{\tau} = 0.430$)

Figure 7.1: Kendall's Rank Correlation Among Dimensions in ACS PUMS for PUMA 1107



(c) Strongly Dependent Dimensions ($\bar{\tau} = 0.620$)

Figure 7.1: Kendall's Rank Correlation Among Dimensions in ACS PUMS for PUMA 1107

7.2 Results

The results presented here were originally published in the paper by Kaushik, Cirillo, and Bastin [60].

We have only explored Archimedean copulas with specified distribution functions for 3 or more dimensions. These copulas currently include the Clayton, Frank, Gumbel-Hougaard, and the Joe families. Only the Gumbel-Hougaard family of copulas (henceforth called Gumbel copula) are presented in this paper. Other copula families are also applicable. However, the parameter values assumed by Clayton and Frank families produce infinities for the independent case, while a simple relation between the Joe family parameter and Kendall's rank correlation is not yet defined [68]. The distribution function for the Gumbel Family was presented in (5.1), but is repeated here for convenience:

$$\exp\left\{-\left[\sum_{j=1}^{d}(-\ln u_j)^{\theta}\right]^{\frac{1}{\theta}}\right\}, \tag{7.1}$$

where u_j are realizations from the uniform marginals (see identity (2.1)) [the text by Nelsen 81, contains distribution functions for other copula families]. The relation between Kendall's rank correlation and the Gumbel copula parameter is straightforward and given as

$$\tau = 1 - \frac{1}{\theta}. \tag{7.2}$$

The maximum likelihood estimates of the Gumbel copula parameter obtained from (5.4) for the three groups of dependencies are provided in table 7.1. The Cramér–von Mises statistics for the ACS PUMS dataset from PUMS 1107 computed using (5.9) are also included in the table. Approximate p-values computed using (5.10) from 1 000 replications of the adapted parametric bootstrap procedure due to Kojadinovic [66] are shown in the table 7.1. As an additional reference, Kendall's rank correlation coefficient obtained from the estimated parameter using (7.2) are also presented in

table 7.1. Note that the these rank correlation coefficients are quite different from the mean coefficients included with figure 7.1. Genest and Nešlehová [36] finds that estimating copula parameters by inversion of Kendall's rank correlation coefficient may be biased, and recommends using the maximum likelihood estimator instead. Further, all recommendations included with the modified procedure due to Kojadinovic [66] were followed while computing the results shown.

Table 7.1: Results from Estimating Gumbel Copula to Data Groups

Data Group	$\hat{\theta}$	$\hat{\tau}$	$\hat{S}_n$	p-value
Independent	1.098	0.089	8.710	0.001
Weakly Dependent	1.556	0.357	48.045	0.001
Strongly Dependent	1.908	0.476	1 804.954	0.001

The low p-values shown in table 7.1 indicate that the Gumbel copula is not a good fit to any of the three groups of data. However, that is somewhat a limitation of the inference procedure, and not Gumbel copula. We show this by creating a synthetic dataset representing the population of PUMA 1107 from the estimated Gumbel copula. The same number of realizations as the original expanded dataset of PUMA 1107 were simulated from the three estimated Gumbel copulas using the mixture method proposed by Marshall and Olkin [77]. These three sets of 118 583 observations, in 3 dimensions each, encapsulate the dependence captured by the copula parameter. Each dimension in these generated independent, weakly dependent and strongly dependent datasets are inverse transform sampled using cumulative mass functions estimated from the corresponding dimension of PUMA 1107 to yield a synthetic dataset for the given PUMA.

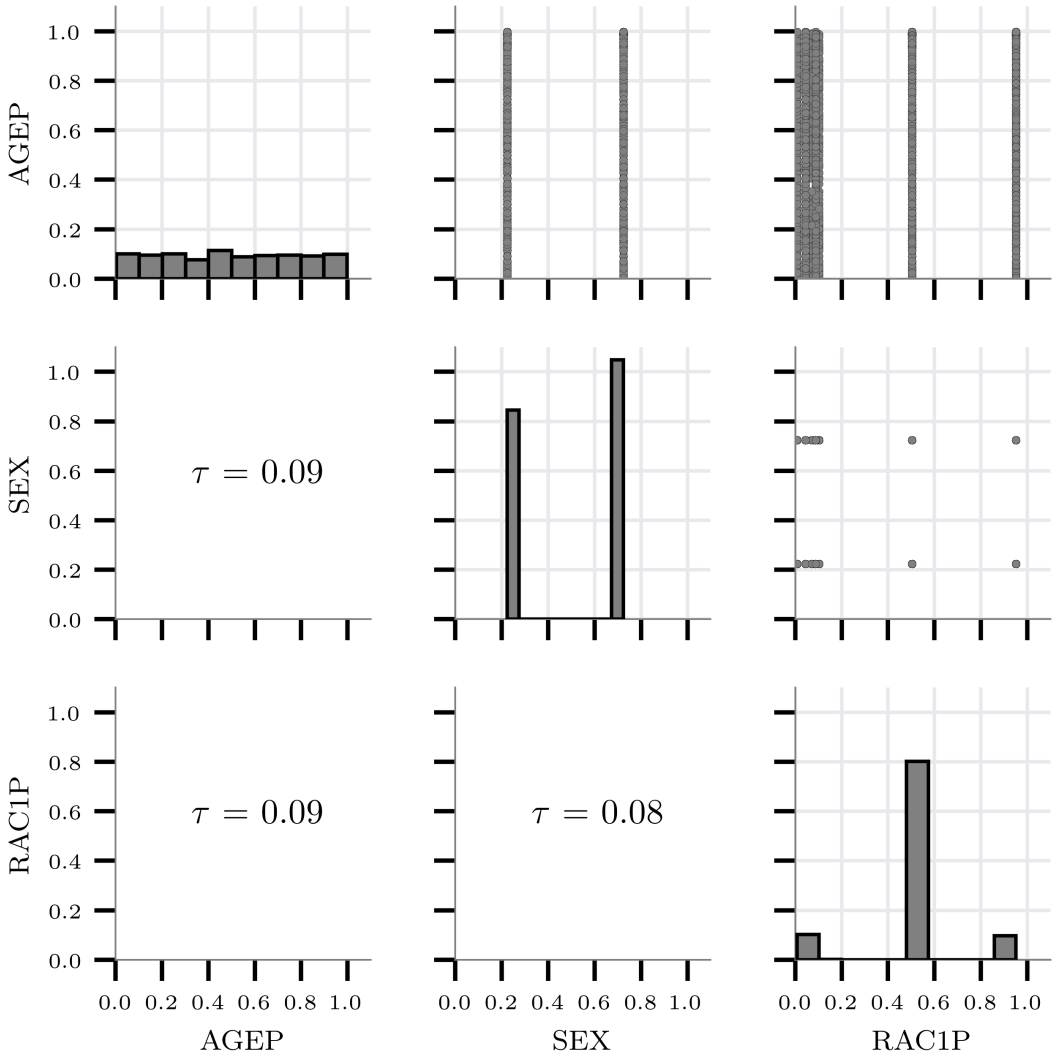
Subtraction rule was applied to the dimension of race code (RAC1P) to convert negative dependence into positive dependence prior to copula estimation. This is a necessary step as the Gumbel copula cannot model negative dependence. Therefore, for this dimension, the rule was reapplied on the values generated from the copula

before inverse transformation. This ensured the simulated synthetic population data were as close to the PUMS data from ACS as allowed by the dependence captured by the copula.

Pairwise Kendall's rank correlation coefficients were estimated for these synthetic population datasets generated using the Gumbel copula for PUMA 1107. (2.6) was used to generate pseudo-observations from the simulated synthetic population data before estimating the correlation coefficients. Further, subtraction rule was applied to the race code dimension to ensure the estimated rank correlation coefficients are positive, as was done while producing figure 7.1a. The results are presented in figure 7.2, using the same format as figure 7.1. Average values of the rank correlation coefficients $\bar{\hat{\tau}}$ are included in parenthesis in the subcaptions of the figure. Again we find that these correlation coefficients are different from the ones estimated using (7.2) from the estimated copula parameters. This result reinforces that method of moments based estimators might produce biased estimates of the copula parameter [36].

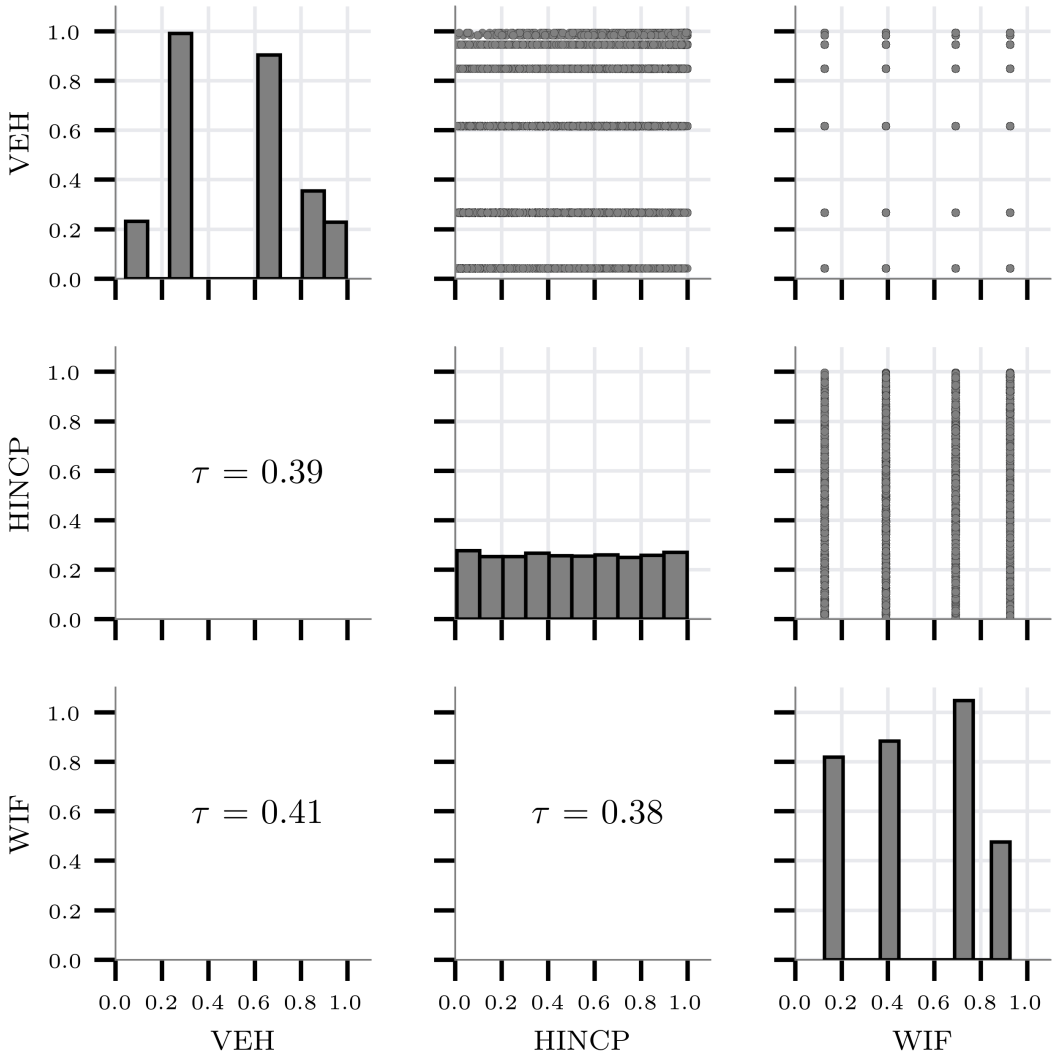
Comparing figure 7.2 with figure 7.1, a few things are immediately clear. The Gumbel copula perfectly captures the weak dependence case, and almost recovers the independence case. In both of those datasets, the estimated p-values were low, and suggested rejection of the Gumbel copula. The same low p-vales were also computed for the strong dependence case, for which the PUMS and synthetic datasets are drastically different. Consequently, we can conclude that the inference procedures with discrete data are not reliable enough to produce the prudent decision in all cases. Further, we show that the modified procedure due to Kojadinovic [66], while adequately reliable with simulated discrete data, fail with survey data. The need for additional and alternate tests are strongly advised before using the results of the inference processes on discrete data.

[2]The realizations for dimension RAC1P (obtained from the Gumbel copula) have been replaced by applying the subtraction rule to values produced using (2.6) so that positive correlation coefficients are produced similar to 7.1a.



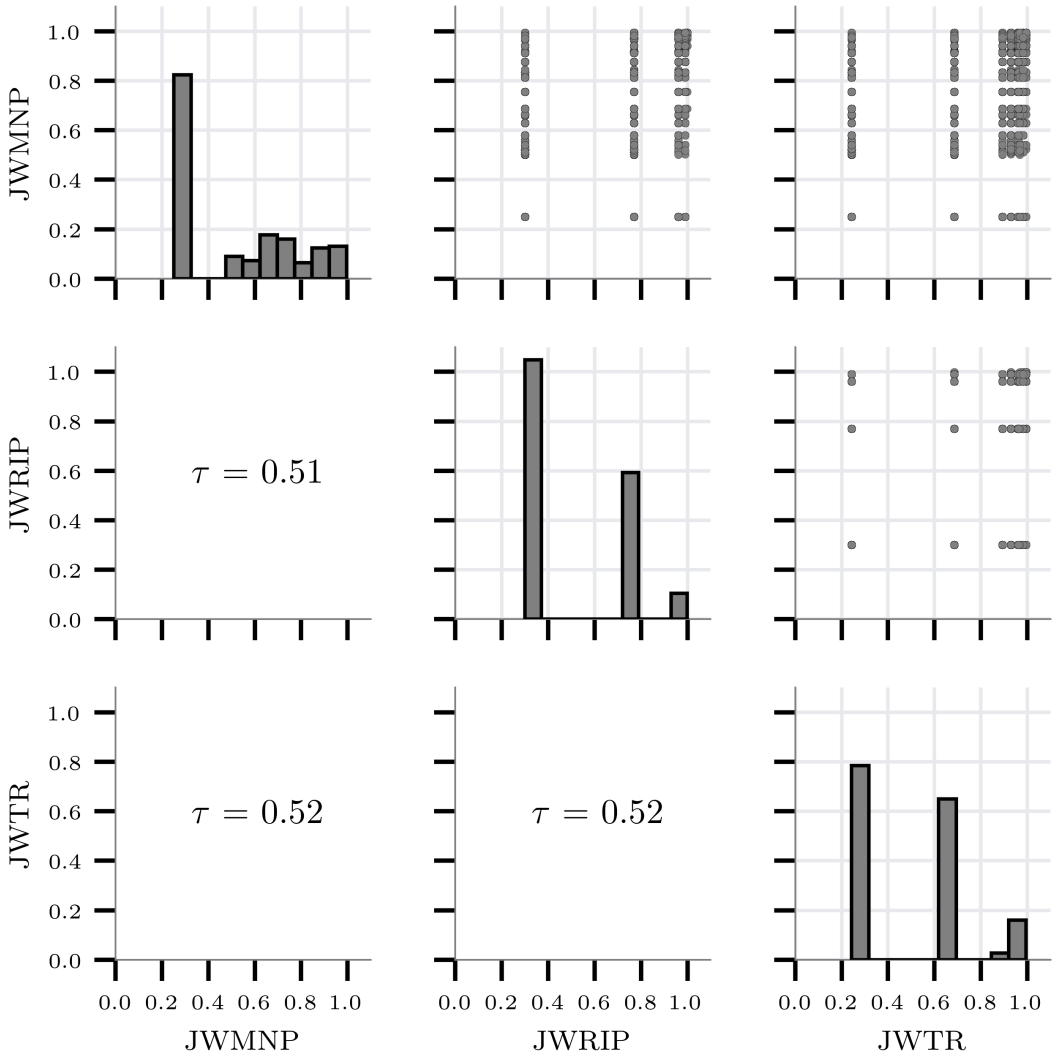
(a) Independent Dimensions ($\bar{\tilde{\tau}} = 0.090$)[2]

Figure 7.2: Kendall's Rank Correlation Among Dimensions in the Synthetic Population Datasets



(b) Weakly Dependent Dimensions ($\bar{\hat{\tau}} = 0.390$)

Figure 7.2: Kendall's Rank Correlation Among Dimensions in the Synthetic Population Datasets



(c) Strongly Dependent Dimensions ($\bar{\hat{\tau}} = 0.520$)

Figure 7.2: Kendall's Rank Correlation Among Dimensions in the Synthetic Population Datasets

The findings by Genest and Rémillard [39] are still valid, as large values of the statistic S_n lead to rejection of the null hypothesis by the inference procedures. The issues are apparent when the statistic values are small, and there is no prudent reason to reject the assumed copula family, like demonstrated with the independent and weakly dependent datasets. However, since the distribution of the statistic depends on the unknown distribution of the copula parameter, which in turn depends on the marginal distribution functions of the multivariate random variable, quantification of the values of the statistic would be difficult [27, 40]. Therefore, data-driven methods of validating copula models are recommended as further research avenues.

Chapter 8: Conclusions from Population Synthesis

In this half of the dissertation, a novel method to generate synthetic population is demonstrated on simulated and real world case study. Using the simulated data, it was shown that the method can recover the dependence, and indicate the correct copula family. In the real world case study, not only the dependence was captured using copulas, but also the captured dependence was used to generate synthetic population at a much finer geopolitical division [see also 98, 99]. Moreover, a linkage between the ACS PUMS data and the decennial Census aggregate data was established by the captured dependence.

The demonstrated method provides contributions to multiple disciplines simultaneously. In microsimulation, it promises to significantly improve the availability of accurate data over all required characteristics. Further contribution to small area estimation is provided by generating population to census tracts, which are widely used in many studies as the smallest unit of measure. The presented method also exposes the applicability of the developed statistical knowledge about copulas to discrete data. Lastly, an alternate form data linkage is provided by the use of the dependence captured from one dataset, and used with another.

Over the current state of the art in population synthesis, the proposed method has numerous advantages.

1. It produces synthetic population for a different and finer geopolitical area.

2. Combined estimation of household and individual characteristics is easily possi-

ble.

3. Additional dimensions can be simply included requiring neither an increase in memory requirements, nor the sparsity of data. More observations, similarly have no negative impacts on the method.

4. The method can work with other datasets, and also potentially in areas without samples, so long as the dependence can be reliably transferred from another area.

5. Parametric specification of the dependence is direct to interpret and debug.

6. Estimation of the parameter is guaranteed and can be programmed to occur quickly, even with very large datasets.

7. Rare observations do not hinder the method, nor require additional preprocessing.

This study illustrates that the use of copulas to generate synthetic multivariate datasets is a promising avenue, deserving more exploration, especially regarding the copula selection and its validation.

The examination of the goodness of fit tests of copulas with survey data provided beginnings of tools with which the tests might be improved, especially when used with discrete data. However, a lot of work still remains. Nonetheless, the novel contribution of the work was the application of the inference methods to detect the suitability of a copula to model a dataset produced by some unknown multivariate random process. A test on real-world data was never conducted in the literature, and consequently the applicability of the methods to datasets outside the simulated realm was unknown. Using simulation, we showed that the Gumbel–Hougaard copula family is able to capture the multivariate dependence among the dimensions of two out of three groups. Alternate methods to compare two multivariate datasets are strongly suggested. Although the case where the Gumbel copula is a bad choice is

correctly identified by the inference procedures, the cases where the copula is an acceptable choice are erroneously falsified. Therefore, the strong recommendation is to use alternate, preferably data-driven, methods to validate the similarities between the original and synthetic datasets, and thus evaluate the applicability of a copula to model the original data.

SECTION II
Performance Measurement

Chapter 9: Introduction to Performance Measurement

While the use of ABM with synthetic population constitutes an elaborate framework where a set of models are used to estimate demand and loads on infrastructure from the individuals and their behavioral patterns, supply side management requires real-time and direct insights into the current operational characteristics of transportation infrastructure. Such insights can be used to pinpoint areas of low performance, and allow taking immediate measures to reduce the negative impacts. Consequently, performance measurement has to happen in real-time, and provide reliable results that can be used to make important decisions. The data that power these inferences have been increasing in volume and velocity for almost 15-years, and are now considered terabyte scale big data. Most performance measurement algorithms are simple so that computation does not take too long with big data, and results are available quickly.

One of the primary real-time sources of data for performance measurement is data collected from GPS devices with mobile telemetry. These data are available over the whole network, represented as small road segments, and at frequencies exceeding an observation per minute. Consequently, these data are quite large, and analysis may take a long time to complete. However, since the insights are required in real-time, the analysis methods, tools and algorithms developed are designed to be simple, yet reliable and straightforward to provide results as quickly as possible, with minimal interpretation of the model itself. Another advantage of using simple models for real-time inferences are to suit the skill level of practitioners in government agencies responsible for coordinating responses to deteriorating performance. Two

such models, one for forecasting and another for computing congestion measures are presented in this dissertation. The model that computes performance, while simple, is reliable enough to be adopted with slight modifications by the Federal Highway Administration (FHWA) in the formulation of performance measurement rules under Moving Ahead for Progress in the 21st Century Act (MAP-21) [26, 61]. The forecasting approach was likewise conferred the Best Paper from Americas award at the 24th Intelligent Transportation Systems World Congress for the simplicity and accuracy of the predicted traffic speeds [18].

The framework developed for producing short-term forecasts adapts ideas at the core of Small Area Estimation (SAE) from the spatial domain to the temporal domain [18]. Here, short-term refers to predictions up to 30-minutes into the future. The chief contribution of the developed framework is this tectonic shift in domain which has never been attempted before. In this framework, future time points are treated as small areas in time, that is, there are no observations from those time points. The key idea of borrowing strength from other larger areas is then used to make estimates about the future time points. Larger areas, in this context, is the past, where many observations are available. The developed method uses these past observations in a simple model that predicts the the future. The performance measurement methods, likewise, build on the strength of the past observations to filter outliers, and produce reliable estimates. The method involves overlaying data from a fixed window of time, say one day, from some similar days. For example, all data from weekdays for a couple of weeks can be overlayed within the 24-hour window to create a dataset with many observations for the measurement interval, which could be a minute or longer. This increases the density of data available in the measurement interval, increasing confidence when rejecting outliers, imputing missing data, or computing the performance of that measurement interval with a baseline interval [18].

Outside real-time requirements, performance analysis is used to measure and report

the overall impacts of transportation on human lives, environment, energy and resource utilization, time consumption, various costs of operation and service, and many other such facets. Such reports are also mandatory under MAP-21, and are directly linked by law to the availability of federal funding. However, these results cannot be obtained solely with the use of big data measured over the network, because the actual number, type and fuel efficiency of vehicles is unknown. Other datasets, like the Highway Performance Monitoring System (HPMS) might record some information like volume of vehicles. Descriptive statistics like totals and averages by make and fuel efficiencies can be compiled from administrative datasets, like vehicle registration information by type and vintage, and fuel efficiency and tailpipe emissions data from the Environmental Protection Agency (EPA). However, these data come from disparate sources available in different spatial and temporal domains creating an illusion of data scarcity. The result of data incoherence is an illusion of data scarcity [88]. Consequently, linkage of data across spatial and temporal domains is key to unlocking various analysis possible with the available data.

Two methods developed for linking spatial and temporal datasets are documented in this dissertation. The first conceived method links the GPS based traffic big data with incident information to find the impact of incidents on congestion and the resulting effects on people and government agencies. The second method conflates two spatial map layers, one from HPMS and another from a private corporation that provides real-time speeds on the network [63]. This spatial join allows transfer of Average Annual Daily Traffic (AADT), a proxy of roadway volume, to augment the speeds dataset, which can then be used to compute systemwide measures like, vehicle miles traveled, total emissions, cost and time spent traveling, fuel, energy and other resource consumption, etc. Note that AADT must be disaggregated into required intervals of time [63]. The paper resulting from the second conflation method won the Outstanding Paper Award at the 97^{th} Annual Meeting of the Transportation Research

Board.

9.1 Need for Short-Term Traffic Predictions

Short-term speed predictions are hugely in demand in the industry in recent times. The main benefit to predicted speeds is the ability to adapt quicker to changing conditions, and provide additional driver assistance. The goal of such Advanced Traveler Assistance Systems (ATAS) are to reduce travel times, and the negative effect of congestion, like wasted resources, time and traveler wellbeing. In the modern world of performance management, the new imperative is to extract as much value from existing infrastructure as possible. Therefore, the past 35 or so years have seen an increasing adoption of Intelligent Transportation Systems (ITS) which includes not only Advanced Traveler Information Systems (ATIS) — such as congestion and travel time information, dynamic message signs, next bus service, and so on — but also Traffic Management Systems (TMS) — like traffic incident management, real-time performance and congestion updates — with the goal of achieving system optimal network flow and utilizations [16, 135]. Much of this has been made possible by significant progress and revolutions in data collection and computational power and algorithms [32].

These systems currently rely mainly on real-time data. Travelers, however, will arrive at points downstream along their path only after a period of time. Consequently, a report of current conditions at downstream points may be of little value, especially because the conditions can change rapidly by the time the traveler arrives at those points. For example, a congestion caused by an accident may dissolve by the time the traveler reaches the area currently congested. However, current conditions may prompt the traveler to alter the path to destination, potentially increasing travel time, and achieving the opposite of what was intended. The ideal situation is to have traffic conditions available 15 to 30 minutes into the future [1, 65, 110, 121, 124] to take the

necessary decisions.

The limitation of not having predictions is well-known, and has greater repercussions than just increased travel time for a few travelers. Travel time and congestion predictions will help all domains of ITS, including ATIS and TMS, as predictive conditions allow actively applied measures instead of reactive measures. Reactive measures are when action is taken after the realization of an event, such as diverting traffic only when congestion worsens beyond a threshold. However, by predicting the progress of congestion, control measures may be set up earlier mitigating the worsening of congestion, which is perceived as a positive benefit by society in general [121]. Further, reducing congestion also makes travel more sustainable and less damaging to the environment and quality of life. Most of modern work in predicting traffic is focused in this area precisely because of this strong demand for better systems, and because travel time is one of the most easily understood of all performance measure on a roadway [83, 138].

9.1.1 Motivation

Predictive models require to be fit, i.e., their coefficients need to be estimated, before predictions can be made. This is true even for non-parametric models and neural networks where the fitting algorithm or neurons need to be trained to make predictions as real-time data flow in. Most recent literature dealing with online models take a combined approach where, even if a model is trained using historic data, its coefficients are constantly updated with each new data point [119, 122]. This model fitting or updating consumes time, and requires the most computation power.

The drawback of existing online models is that the method is not scalable with data. Fitting on a huge network, using high frequency data requires a huge amount of computation power and time. Fitting models may not be possible, in such a scenario, over the whole network before the next data point is available. Therefore, the model

may never be fit in time to start making predictions. This is arguably the most probable cause why even recent literature — published in times when big data and huge computational power are readily available — does not have a combination of large network, high frequency data and short-term predictions [122].

By decoupling the fitting from the prediction steps, models can be made such that forecasting is fast and easy, while fitting occurs at a time convenient so that it does not impact predictions, and can be done on moderately powerful computers with sufficient time to spare. This is the fundamental idea behind the synthetic time series framework proposed in this chapter. The repeatability of traffic data is used to effect the synthetic time series, which is used for fitting the models and then making predictions. Figure 9.1 shows vehicle speeds observed on a segment of road from three consecutive Wednesdays in September overlayed on top of each other over a 24-hour period. The repeatability of traffic patterns is clearly apparent in this figure.

Although forecasting traffic patterns is a very complex problem due to the countless factors that influence the motion of vehicles on road, some of which cannot be included in a model such as driver temperament and driving styles [139]. Even the factors that can be included as auxiliary variables, like upstream and downstream conditions, weather and incident information, etc. need reliable and interrelated databases. However, when data from repeatable periods is overlayed, a mean trend becomes apparent [see 61, and also figure 9.1], which can be modeled, while relegating the variations from each period to the error component. Autoregressive Integrated Moving Average (ARIMA) models are inherently designed to model just such a time series, where the mean trend is captured by the coefficient, while the error captures the noise in the time series data. The use of historic data sits squarely within the use case of the synthetic time series framework.

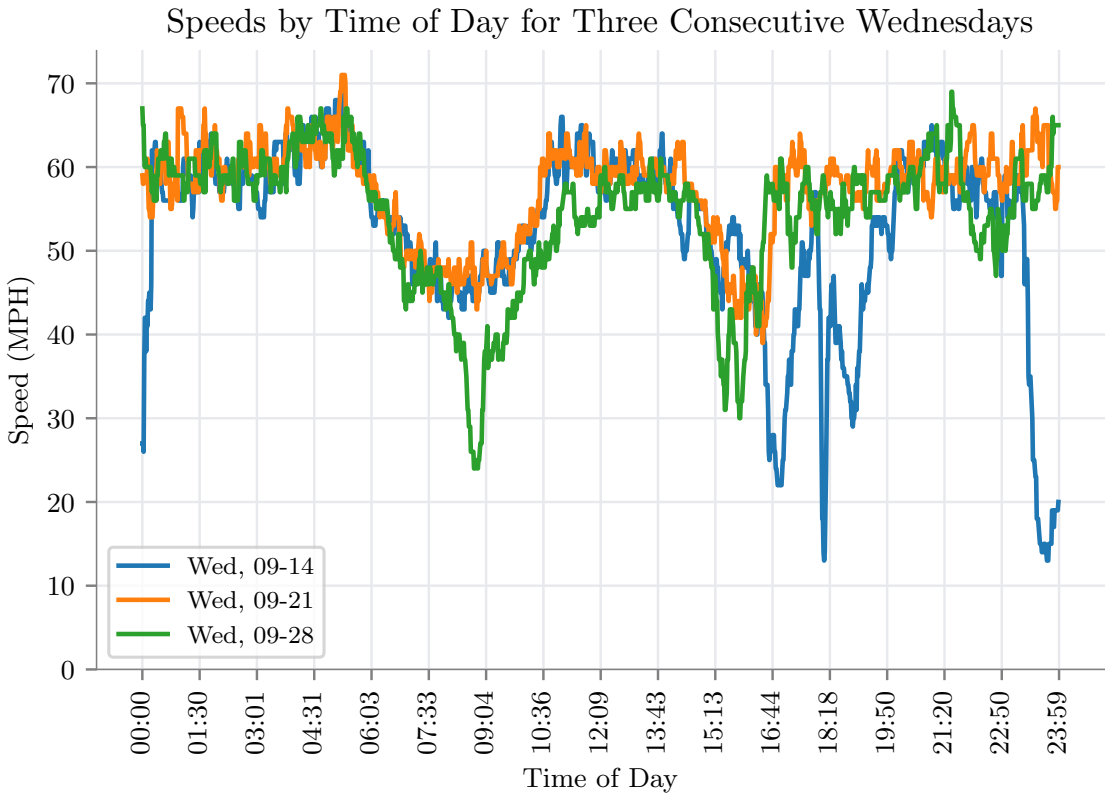


Figure 9.1: Overlayed Traffic Speed Observations from a TMC Segment

9.1.2 Synthetic Time Series

Suppose that the data for a day were completely available beforehand. Models fit to these data would be able to accurately forecast traffic from any point in time for that day for any given horizon. However, since observations for a day are only available as the day progresses, data for the day could be considered incomplete. The synthetic method ameliorates this problem by using data from another time period, where the complete set is available, to inform estimates for the current period. To elaborate, suppose short-term traffic speed predictions on a roadway segment for September 21 from figure 9.1 are required. It can be seen that data profile from September 14 is very similar to data profile from 21^{st}. Therefore, models can be fit to data from 14^{th}, the most reasonable one selected to be used for predicting traffic speeds for 21^{st}.

It can be seen that the synthetic method used as outlined above accords some favorable advantages to the models. Firstly, the models are fit on data from the whole day, they are trained using the entirety of expected traffic patterns. Hence, they can reasonably predict the future short-term trend of traffic patterns based on a few real-time observations. Literature contains many examples where models fit on partial data consistently either over or under estimate during peak and off-peak periods [52, 130, 137, etc.]. This limitation is overcome in a natural way by the synthetic method.

Secondly, in the example above, the models were fit to data from a week before the day of interest. This gives ample time to fit multiple models and select the most reasonable one for a segment of the roadway. The available time also allows for models to be fit for each individual segment in the network. Traffic characteristics can vary widely not only between different roadway classes, but also between different sections of the same roadway. Therefore, a single large and complex model for a network may be outperformed by small and simple models for individual segments that form the network. In addition to simplifying the models used, this also allows easy introduction

of auxiliary variables to provide supplementary and complementary information like incidents, weather, upstream and downstream conditions, etc. to the simple models.

Finally, the synthetic method is extremely flexible. It is compatible with almost all models presented in the literature so far, the only exceptions being some data-driven models, like Neural Networks, that use bulk data to train the models. The only drawback of the synthetic framework is that it requires panel data across time, and that data should be similar across panels. However, this hurdle can be overcome to varying degrees by careful modeling, and transfer of coefficients across time and data sets — limited by the linkages that can be established between datasets.

When compared to data driven approaches, the synthetic method has the additional advantage that a smaller dataset can produce sufficiently reliable estimates. Researchers have postulated that about a year's data is required for training reliable data-driven models [115]. However, on the network studied (figure 11.2) data for a whole year at one minute frequency for each segment measures in gigabytes. Such big data, although more readily available today, requires expert database administrators and programmers to handle and utilize in data-driven models. However, as demonstrated here, data from a single day is sufficient to predict traffic for the same day in the next week, vastly reducing the data requirement. Of course, given the flexibility of synthetic method, one may train a data-driven model within the framework and apply it instead.

9.2 Requirement of Performance Measurement

The National Performance Monitoring Research Data Set (NPMRDS) was procured by the Federal Highway Administration Office of Operations in 2013 initially serving as a research data set for sponsored programs, but with the anticipated emphasis placed on performance measures in the MAP-21 legislation, rights to use the data set to compute performance measures was secured for State Department of Transportation (DOT)

and Metropolitan Planning Organizations (MPO). NPMRDS is a form of commercial Global Positioning System (GPS) probe data, meaning the traffic conditions reported are derived from vehicles self-reporting speed, position, and heading based on GPS electronics on a periodic basis. It differs from commercially available data feeds in that FHWA specified that no smoothing, outlier detection, or imputation of traffic data be performed. As a result, NPMRDS contains unique characteristics with respect to statistical distribution of reported travel times. These unique characteristics render traditional processing techniques ineffective to obtain accurate performance measures.

A method to handle the challenges posed by the nature of NPMRDS, and compute meaningful performance measures from it is proposed. First, the challenges in processing NPMRDS data are exposed and a method to overcome them is designed. Then, the results from the proposed method are compared with traffic data from commercial probe data sources as well as a reference re-identification data sourced at two case study locations. The case studies indicate that the method shows the ability to more accurately capture performance measures from NPMRDS using techniques originally developed to accurately reflect travel time and travel time reliability on interrupted flow facilities.

Quantifying congestion is the largest form of highway performance measurement currently in practice, as evidenced by the majority of documents released by various departments of transportation [31, 47, 106, 116]. The ability to compute congestion measures for the whole network is also being stressed by initiatives like the MAP-21. Traditionally, the data used for computing these measures involved floating car trips during peak and off-peak periods. The inability to scale these data collection techniques meant that congestion measures were spotty in time, and restricted to only the most important of road networks.

With the wave of data revolution that is spreading through transportation, computing congestion measures has become more wide-spread, often encompassing large

networks, and various road classes. However, not all data are created equal. Where floating car runs could capture travel times in any road setting, the new GPS based probe data suffers from limited fidelity on roads with friction to traffic flow like signalized intersections, driveways and parking. Loss of accuracy is caused by the filtering that is applied to probe data because the source of the data cannot be confirmed. Pedestrians and cyclists with GPS devices like phones can unwittingly contribute their movements to probe data, and these need to be filtered out. Similarly, vehicles experiencing trouble, or stopped at rest stops are excluded from observations because they are no longer part of the traffic flow. However, trouble arises because the algorithm cannot distinguish between vehicles stopped at signals and vehicles stopped due to mechanical failure, driver decision, incidents or weather related dangers.

The stress laid on performance management is even more acute due to the budget constraints facing almost all departments of transportation. MAP-21 essentially conditions the distribution of federal funds on setting and meeting performance targets. In order to facilitate computations of these performance measures, and to ensure all departments are producing results using the same method, the FHWA procured and made available a national travel time dataset, called the NPMRDS [28]. Additionally, through the Notice of Proposed Rule Making (NPRM) process, the FHWA has provided the algorithm and formulas to be used to compute performance measures from the NPMRDS [26]. These rules, based on the work documented in this dissertation, went into effect on May 20, 2017.

The NPMRDS is available to all state DOT and MPO free of cost. This dataset provides travel times on road segments spanning the National Highway System (NHS) and the Strategic Highway Network (STRAHNET) at 5-minute intervals, for every day of the year. The NPMRDS reports different travel times for automobiles and trucks, to enable measurement of freight performance separately. NPMRDS is made available monthly, as an archive of travel times for segments in the NHS state by state. With

the availability of the NPMRDS, small state DOT and MPO were suddenly faced with the challenges of processing big data. In fact, the first three quarterly webinars on NPMRDS focused on the issues of data size, and software that could process the data [28, 29, 30]. As the NPMRDS slowly brought about a change in the experience of personnel hired by DOT and MPO, a reliable and straightforward, yet unbiased, technique for computing performance measures was required.

NPMRDS is specified to be a raw dataset, with minimal filtering applied to the observations. Also, it does not impute missing data, leaving large holes in the dataset, especially during night hours. Both of these attributes make it a difficult dataset to work with. An algorithm capable of handling these problems was developed to provide a simple method to compute performance measures from NPMRDS. The method adapts the FHWA guidelines to compute the Planning Time Index (PTI), Travel Time Index (TTI) and the Buffer Time Index (BTI) for use with percentiles obtained from the NPMRDS data over the required period [31].

The developed method is easy to encode in a computer program and have small computational cost. Moreover, by comparing to Bluetooth re-identification data collected from the side of the road, it is shown that the method is not biased, and yields the correct results, even for interrupted flow facilities. Additionally, the fidelity of the NPMRDS data on different classes of roadway was also assessed, and found that NPMRDS without filtration and other cleanup at source performs better than cleaned data, especially on roadways with high variation in travel times, like arterials.

9.3 Conflating Two Geospatial Datasets

The advent of mobile devices with embedded global positioning systems has allowed commercial providers of real-time traffic data to develop highly accurate estimates of network-level vehicle speeds. Additionally, methods have been developed to monitor the network performance in real-time, and even make short-term predictions. Traffic

speed data have far outpaced the availability and accuracy of real-time traffic volume information. Limited to a relatively small number of permanent and temporary traffic counters in any city, traffic volumes typically only cover a handful of roadways, with inconsistent temporal resolution. A conflation methodology was developed to address this data gap by coupling a commercial dataset of typical traffic speeds (by roadway and time of week) from TomTom to the U.S. FHWA's Highway Performance Monitoring System (HPMS) database of Average Annual Daily Traffic (AADT) counts by roadway. The novelty of the developed method is the ability to conflate independent road geometries spanning the whole country using big data techniques. The resulting product is a national dataset providing traffic speed and volume estimates under typical traffic conditions for all U.S. roadways with AADT values.

The ubiquitous availability of speed data is revolutionizing the operation of transportation networks, especially roadways. It is now possible to get a bird's-eye view of current conditions on a whole network, spanning a large area [61]. Rerouting vehicles based on current roadway conditions to avoid congestion and other delays is almost a regular occurrence, especially with advances in route guidance systems built either into vehicles or using smartphones and other mobile devices. Dynamic message signs are increasingly being used to alert travelers of current road conditions downstream so that delays are minimized proactively. These Advanced Traveler Information Systems (ATIS) are being informed by ubiquitously available speed data [137].

Similarly, for planning, a panel of speed data collected over a period helps identify problem areas, bottlenecks, and other issues that may each require different solutions. For example, areas of recurring congestion on certain sections of an urban interstate may prompt a detailed study into the intersection design, the number of lanes, lane widths, demand, and other factors to find workable solutions. Most of these studies, however, require more than just speed data. Such studies rely, perhaps more heavily, on data such as volume of vehicles during peak hours, mixture of vehicle types, geometric

design of the highways, and the rate, type, and severity of incidents.

Researchers and planners are increasingly finding that these studies are made difficult not by the availability of datasets, as most of them are easily acquired, but by the inconsistencies in how different datasets are packaged and reported. No unified underlying geospatial framework exists to which all sets conform, despite global standards and organizations, such as the Open Geospatial Consortium and Technical Committee 211 of ISO. Private companies in the transportation domain continue to push their own independent geospatial data reporting frameworks or continue adhering to older standards [61]. Public agencies, on the other hand, suffer from a shortage of skilled people capable of digitizing existing maps and keeping the current digital maps updated [61]. Two approaches to affect spatial conflation were developed and are documented in this dissertation.

The issue tackled here is the spatial and temporal conflation of the HPMS dataset with Multinet, a commercial speed dataset purchased from TomTom Inc., a multinational vendor of maps and other cartographic services. The spatial conflation is complicated by the lack of a unifying framework used by both datasets, which makes join conditions difficult to resolve automatically. The HPMS dataset is represented by single line segments that follow either one of the carriageway in roadways. On the other hand, Multinet is based on a proprietary segmentation scheme that follows the centerlines of both carriageways in roadways with dual carriageways. There is no parity in the segment lengths either, and consequently segments from the datasets do not coincide anywhere in the entire United States.

Specific algorithmic tools developed to address these challenges are presented. The conflation was not perfect, but the average error in total conflated segment lengths was just under 7 %, which, for the whole United States is commendable. The developed algorithm can best be described as a scoring algorithm, where points are rewarded in direct relation to the strength of the match between segment metadata of the

two datasets. Most of the metadata were generated using the geometry of the line segments themselves, such as the parallel and perpendicular distances between specific points along the line and the angle of separation between them. Some criteria were borrowed from the few fields that were roughly common between the datasets, such as roadway name and shield numbers.

After joining the datasets using the spatial join, the AADT values are transferred from the HPMS to the TomTom Multinet dataset. However, the AADT values are a single number per segment, representing the average annual daily traffic expected on that segment. Using a technique developed by Texas Transportation Institute (T.T.I.) the AADT values were disaggregated to 15-minute intervals for each typical day of the week. The splitting algorithm due to T.T.I. requires typical speeds by day of week, and direction of travel, which would not be possible without the join connecting the AADT and the typical speeds from TomTom Multinet.

9.4 Merging Congestion and Incident Data

Congestion and incidents have an intrinsic relation with each other. Congestion is shown to depend on the demand to capacity ratio of any roadway at a given time. Incidents can take away some of the capacity of the roadway, thus altering the ratio and feeding congestion levels. Conversely, congestion increases the risk of incidents as drivers unaware of the downstream conditions might collide with slow moving vehicles in congestion. The efforts to find the impact of various incidents on congestion was developed to answer such questions. The focus is to correlate the incidents on the freeway system of Maryland with congestion as defined in the Maryland State Highway Mobility Report. The data used in this study was collected in 2011, and spans the whole year. It is found that incidents have a varying effect on congestion. Although the probability of incidents increases with increasing number of vehicles, incidents are more location specific rather than volume dependent. Locations prone to incidents

show larger impact of incidents on congestion as compared to locations with fewer incidents.

Effective improvement of roadway performance does not solely involve adding new capacity to resolve the existing problems of congestion. Often, such action causes evanescent improvements, and gains are quickly overcome by the latent demand. Therefore, it is crucial to understand the underlying cause of traffic jams, and differentiate between those caused by capacity constraints from those caused by bad geometry or design constraints. Here, we divide congestion into two broad categories: recurring, or caused by more demand than the capacity, or non-recurring, caused by incidents.

The distinction between the two causes of congestion is important to tailor the correct solution to the problem. The idea of Traffic Incident Management (TIM) is based solely on the ability to make this distinction. TIM refers to handling of incidents rapidly, so that the disruption caused by it is minimized. Providing TIM patrolling in areas experiencing recurring congestion might not be as effective as providing it in areas where motorists face a higher chance of experiencing distress. Various departments of transportation in the United States understand and prioritize providing aid to motorists so that issues can be quickly resolved with minimal disruption to the commute of others.

TIM forms a considerable portion of the annual Maryland State Highway Administration Mobility Report [76]. Similarly, the mobility reports published by other departments of transportation, such as the Federal Highway Administration (FHWA), the Texas Transport Institute (T.T.I.), Washington State, and others, lay considerable emphasis on executing TIM [31, 47, 106, 116]. Further, given the budget constraints, effectively providing targeted incident management services, as compared to a blanket application, is more important. Consequently, departments of transportation are interested in knowing the best areas and routes for directing highway patrol so that motorists having difficulties can be identified, reached and assisted with minimal loss

of time.

The presented study spatially and temporally joins the congestion and incident databases to find the areas where incidents play a larger role in causing and prolonging congestion. The data used was compiled throughout 2011. Congestion, in this study, is defined where traffic speeds drop below 60 % of the freeflow speeds for an extended period of time. The congestion is given a score based on the duration, length of road congested and the frequency of occurrence. Incidents are recorded by the Maryland Coordinated Highway Action Response Team (MD-CHART), and include a number of categories, some of which are alerts. The location of the head of the bottleneck, and the location of the incident are spatially joined to infer the relationship between them. Temporal analysis then follows, to determine how many occurrences of the bottleneck was caused or prolonged by incidents.

Note that the conflated datasets can be used with the synthetic time series framework for forecasting traffic. Use of volume and incident information in the framework would further enhance the reliability of the predictions. However, such a task is relegated to a study beyond the scope of this dissertation. In this second half of the dissertation, a literature review of the existing methods is presented in chapter 10, followed by a description of the datasets used in chapter 11. The developed methods are documented in chapter 12, while the results are presented in chapter 13. Finally, chapter 14, provides a summary and concludes the second half of the dissertation.

Chapter 10: Review of Performance Measurement

The field of performance measurement is quite vast, and each technique developed and documented in this dissertation has volumes worth of literature. Important and seminal works published in the literature associated with each method are described in this chapter. A note, however, that the literature in this chapter is from the time when the methods were originally developed. Given steady improvements in data collection, handling and cleaning, and also improvements in the available tools and algorithms to perform the analysis, current state-of-art might be better than the developed methods.

10.1 Literature on Traffic Speed Prediction

The vast trove of knowledge available on the subject of traffic data forecasting possibly warrants a library dedicated to archiving the published literature. Over time, numerous attempts have been made to gather and summarize the literature in review papers [7, 50, 65, 83, 110, 118, 119, 121, 122, 137]. In fact, there have been special issues published dedicated to just the topic of short-term forecasting [140]. From these review papers, literature of historic significance to the proposed model is briefly outlined. Readers wishing to gain further knowledge in short-term traffic forecasting are referred to read the cited review papers, and special edition journals.

The main trend that can be observed from the review papers is the evolution of prediction models. Initially, statistical models were favored, mainly because of the cost of computation [2, 20, 46, 72, 130]. However, since the late 1980s, more models

have been steadily replaced with machine learning procedures [1, 7]. At first, statistics based machine learning techniques were employed such as Kalman filters, and Bayesian Inference [85, 118, 135]. However, those were replaced with purely machine learning based models such as k-Nearest-Neighbors, Neural Networks, etc. [50, 65, 110].

Parallel to the models, the data available and accessible evolved and grew. Initially, forecasting was based on very local collected data either automatically using loops [130], or manually. This limited the uses and range of the methods to be focused on just a part of a road. Such methods are useful only for academic importance, and cannot be used for delivering any ITS products. However, with the advent of data, mainly the GPS based probe data, the situation changed [61]. In fact, it is arguably this abundance of data that caused the boom in adoption of data-driven approaches [32, 83, 115]. More importantly, this motivated the application of predictions to more than a section of road, but to entire roadways, and networks [15, 58, 79, 85, 126, 128, 133].

Recent literature discusses very little about parametric statistical models such as ARIMA, and is focused more on data-driven approaches [111]. However, most of the recent studies deal with the whole network, instead of just a section or a road [32, 52, 73, 117]. Unfortunately, the same cannot be said of studies on parametric methods, specifically ARIMA and derivations [12, 129].

The main drawback in using parametric models for a large network is the heterogeneity of the roadway characteristics. Fitting and calibrating a parametric model to a vast and diverse network requires a huge amount of work, and the results may well be outstripped and outperformed by a neural network approach [32, 111]. In the documented study, the problem is solved by simply treating each segment as an individual entity and fitting a different model for each. That way, the best ARIMA order — based on lowest Bayesian Information Criterion (BIC) — is applied to each segment for each of 10 weekdays (a total of 26 540 models are selected out of 716 580

possible models), guaranteeing that the predictions for that segment for the same weekday in the following week are the best possible. Of course, the approach can be improved by considering the impact of the upstream and downstream segments, incident data, and weather data, which is part of planned future work.

10.2 Performance Measurement Literature

Departments of transportation and research institutions that have some experience with probe data and big data were the first to use NPMRDS. Since the dataset cannot be used for general research, except ones that directly lead to performance measures, there is little literature available, and most of it is presented in the quarterly NPMRDS webinars held by FHWA [28]. Probably the first publicly available document was a presentation by Wisconsin Traffic Operations and Safety Laboratory, presented during the second quarterly NPMRDS webinar in February, 2014 [28, 29, 96, 114]. The authors primarily discuss performance measures from NPMRDS, along with representation of the data on a map.

Another presentation in the same webinar, by University of Maryland highlighted the differences in the TMC segmentation scheme and map realizations used by NPMRDS and the I-95 Corridor Coalition's Vehicle Probe Project (VPP), concluding that direct comparison between different sources need to be done carefully to account for the differences in the segment properties [29]. Another study that compares NPMRDS data with Bluetooth re-identification and VPP probe data was conducted at CATT, and was presented at the 2014 ITS World Congress [62]. University of Minnesota and Minnesota Department of Transportation visualized data from NPMRDS on the map layer provided with the data [13]. No processing of the archived NPMRDS travel time data was carried out for this visualization. The American Transportation Research Center (ATRI) later used NPMRDS to compute the congestion and delay price paid by the trucking industry [91]. Freight truck data from NPMRDS was used in the study,

in addition to data from ATRI's Freight Performance Measures (FPM) database.

The Upper Midwest Reliability Resource Center maintains a Travel Time Reliability Reference Manual online [11]. In the manual, NPMRDS data was compared with probe data from INRIX to find that NPMRDS reports a lower mean for travel time with a higher variation as compared to data from INRIX. In the third quarterly NPMRDS webinar, Iteris shared their work for implementing MAP-21 performance measures [30]. This was the only study where extensive processing of NPMRDS data was done prior to computing performance measures.

10.3 Literature on Geospatial and Temporal Conflation

Conflation, as used here, means a spatial join rather than data fusion, as is commonly used in literature. Moreover, the developed conflation methods can be viewed as a join in a relational database system, where joins are used to transfer data from one table or database to another. The spatial join is used to transfer the AADT values from the HPMS dataset to the Multinet dataset. Additionally, another type of join is used to transfer incident information to congestion data. Literature that deals exclusively with conflation of road geometries, either as lines or points, covers a wide range of methods, with no one-size-fits-all method. This is perhaps not surprising because each conflation effort has unique challenges and requires a custom-tailored solution.

Road geometry conflation is mainly born out of the need to integrate open access datasets with existing datasets used by private or public agencies. This linkup is increasingly seen as a cost-effective way to keep datasets up to date with information generated by drivers so that they can be used to provide the most applicable real-time solution to issues faced by drivers [14, 24, 132]. The major source of open geospatial data is Open Street Maps (OSM), which provides rich details about the real-time urban landscape. One recent paper published on road geometry conflation uses urban blocks as polygons surrounded by road segments. The goal of the conflation is to connect

OSM with government data-sets. The polygons from the two sources are matched together in a two-step approach before matching the surrounding road segments. The authors claim that this process has a very low computation cost because the number of urban blocks is fewer than the number of road segments. However, this method is limited to urban areas only [24].

Yang et al. proposed a probabilistic relaxation method to again match the OSM data to government datasets. The relaxation method performs a heuristic search of the neighboring space from an initial set created using the dissimilarities between the geometries. The probability matrix is iteratively updated based on selected criteria, until a match is found [132]. Chen et al. developed an algorithm to attribute floating car data to map segments. The authors argue that existing algorithms found within in-vehicle navigation systems cannot accurately associate low-frequency data to map segments, and propose a dynamic program supported on multiple criteria as a solution [14].

Other examples of road geometry conflation in the literature deal with comparing and benchmarking the accuracies of datasets rather than transferring or updating data. Chiefly, these studies also relate to OSM, as it is crucial to understand the caveats in this modern, free, and crowdsourced dataset, given the rising importance placed on it. Important studies in this domain are works by Girres and Touya [42], Haklay [45], and Neis, Zielstra, and Zipf [80].

The history of automated road geometry conflation can be traced as far back as 1988 to Saalfeld, who used successive approximations to overlay two map geometries [102]. Noteworthy papers have since included a statistical approach [125] and an approach matching semantics prior to overlay to ensure the overlay was as accurate as possible [127]. All these papers used map features, like lines, points, or nodes, for the conflation. A rule-based method, much like the one proposed here, was first proposed by Cobb et al. [19]. This method also used the metadata in addition to geometric

attributes like shapes and locations for the conflation. The decade between 1998 and 2010 is filled with notable contributions by authors using a combination of different data attributes, including geometric and semantic, to achieve a conflation that suits the needs of the study. Altogether, over a hundred relevant papers were produced during this decade [101, 136]. This literature also provided ideas for the spatial and temporal conflation of incidents and bottleneck data.

Chapter 11: Description of Datasets used for Performance Measurement

A lot of datasets have been used for performance measurement, and were processed differently for each method before applying the developed algorithm. In this chapter, an overview of the datasets used is provided first, followed by the description of preprocessing for each method, as appropriate.

11.1 GPS Probe Based Datasets

GPS probe based data are now at the core of and an important part of the available transportation data. There are multiple datasets that are directly measured probe data, and even more which distill some initial results for consumption in other models. The following four datasets are based on GPS probe data, and have been used in the different methods documented in this dissertation.

11.1.1 Vehicle Probe Project (VPP)

The data used in this project is obtained from the Vehicle Probe Project (VPP), which started as a result of contracts awarded by the I-95 Corridor Coalition since 2008. Contractually, the data is received in real-time at one minute intervals. The data comes from probe data vending companies, of which INRIX, HERE and TomTom are the three major players in North America. The data is based on vehicle speeds collected by the probe data vendors from fleet operators, individual users and other

companies which deal with data collection and reporting [61].

Vehicle speeds reported in VPP rely on GPS and telemetry technologies. The GPS devices in the vehicles compute their location, compass bearing, and speed. This information is then periodically transmitted such that the vehicle leaves a bread-crumb trail of locations. By attributing these speeds to directional roadway segments based on reported location and direction, then statistically aggregating these observations, a product like the VPP is produced. Usually these data are collected as part of a tacit agreement where the user gets real-time traffic, routing or other information, while the company gets the GPS data. Probe data vendors often use proprietary algorithms to clean, filter and impute missing data before reporting segment speeds at desired frequencies — one minute for the VPP [61].

Traffic Message Channel (TMC) segments are used as geographic boundaries to aggregate the data. The TMC is one of the oldest roadway segmentation schemes, where the roadway is divided at intersections to form the segments. In the case of the VPP all valid GPS observations attributed to a TMC segment in one minute are aggregated together and reported for that segment at that minute. There may be some instances of time when GPS observations are not available, like in the midnight hours, when traffic volume is lowest. Also, if an observation cannot be clearly identified as an outlier, confidence in that observation is low. In such instances VPP data are bolstered using historic observations in the same time period or by the free flow speed of the roadway. The data has fields to identify these modified records and indicate the level to which they have been altered. However, original (raw) readings from vehicles are not available.

11.1.2 National Performance Management Research Data Set (NPMRDS)

The NPMRDS is a national dataset made available by the FHWA to all state DOT and MPO [28]. The University of Maryland (UMD) has access to the dataset through an agreement with the Maryland State Highway Administration (SHA). The data is archived alongside other probe data received via the VPP. The NPMRDS covers all of the National Highway System (NHS) and Strategic Highway Network (STRAHNET). The NPMRDS is split into segments from intersection to intersection, as per the TMC specification. Speeds for trucks, automobiles and combined traffic are reported for each segment at five-minute intervals for every day of the year.

The NPMRDS is inherently compiled from GPS probe data. GPS devices equipped with telemetry in vehicles are usually programmed to periodically transmit the location, speed and heading computed by the GPS hardware to a central processing station, usually the offices of the GPS device vendor, or, for fleet operations, the vehicle coordinator. These data are then attributed to road segments given the location and heading of the vehicles. Usually, like for the VPP, the observed speeds are filtered to exclude obvious outliers, like pedestrians and cyclists with GPS enabled phones, stopped vehicles, and so on. The cleaned speeds are then aggregated over the reporting window, and reported. In the absence of real observations from vehicles, historic observations or educated guesses of the real conditions might be reported.

The NPMRDS, however, contractually prohibits filtering of speeds, other than wrong way travelers, pedestrians and cyclists. Additionally, imputing historic or estimated speeds are also prohibited during periods of no observations. These two characteristics impart difficulties in using the NPMRDS directly for a finished product. Namely, the lack of outlier and noise filtering makes estimates unreliable, as stopped vehicles are often included (e.g. trucks parked at stops with electrical services turned

on). The gaps in the data caused by non-reporting during periods with lean traffic cause problems to models that rely on an uninterrupted time-series for analysis (e.g. scanning for congestion, vehicle trajectory tracing models, and others). Such problems were difficult even for professionals, using state of the art algorithms which relied on a continuous supply of data, to overcome.

11.1.3 Bottleneck Dataset

The bottlenecks are identified and ranked by the Bottleneck Ranking Algorithm developed by the Center for Advanced Transportation Technologies (CATT) at UMD. This algorithm is applied to the probe speed data also archived by the VPP. The function underlying the algorithm is summarized in figure 11.1. A bottleneck is suspected when the speeds observed from a TMC segment drop below 60 % of the freeflow speed of that TMC segment. If the speed remains below the threshold for 5-minutes or more, the bottleneck is confirmed, and added to the log of bottlenecks originating at that TMC. As the conditions improve, and the speeds rise above the 60 % threshold, and remain there for 10-minutes or more, the bottleneck is considered clear, and the entry log is closed. The duration and suspected length (sum of lengths of all upstream segments also similarly congested) of the bottleneck are recorded.

If two bottlenecks merge, or the head of the existing bottleneck moves downstream — that is, downstream segments start getting congested — then the downstream most congested segment is considered as the head of the bottleneck. If the bottleneck propagates into a spur in the road, where two streams of traffic merge, the merging point between two roads act as the head of the bottleneck on the merging road. The bottleneck impact factor is computed by multiplying the average length and duration of the bottleneck with the number of occurrences during the queried time frame,

$$BIF = N \times \bar{T} \times \bar{L}, \tag{11.1}$$

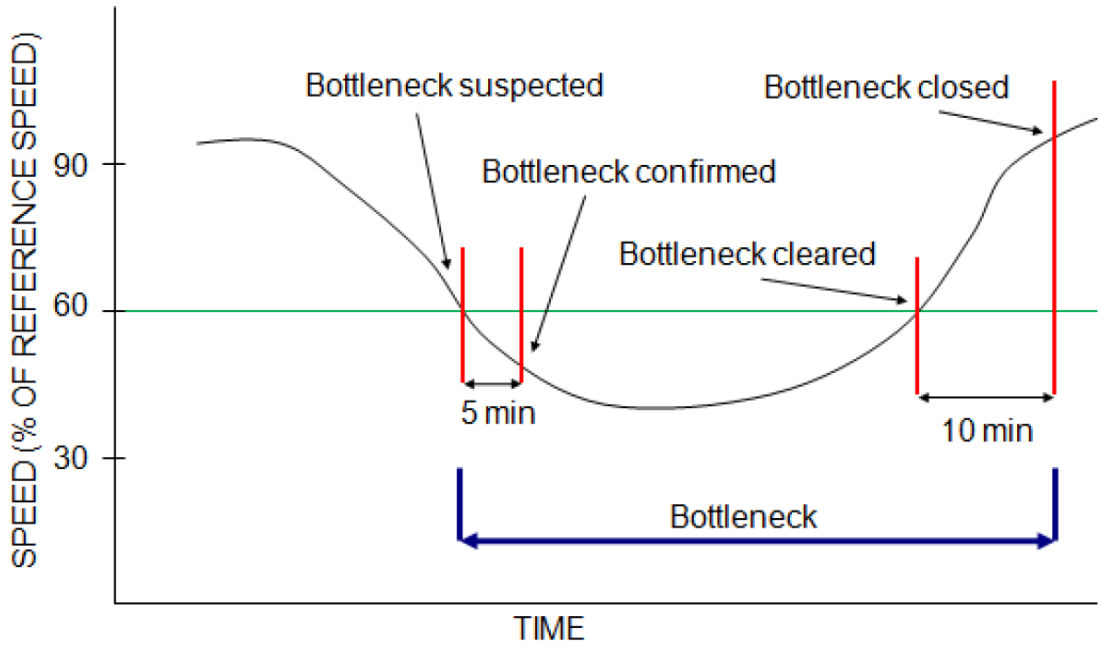


Figure 11.1: Graphical representation of the function to detect bottlenecks

where

BIF is the bottleneck impact factor in mile-hours,

N is the number of bottlenecks recorded in the time frame of interest,

$\bar{T}$ is the average duration of all N bottlenecks in hours, and

$\bar{L}$ is the average of maximum lengths of all N bottlenecks in miles.

Although the location of the head of a bottleneck is known, and the maximum length and duration of each recorded bottleneck with the same head are also available, no details about the upstream segments are provided. That is, the bottleneck head locations are not mapped to TMC segments. Therefore, tracing the length of each occurrence from a given head is not possible. Especially, the evolution of the bottleneck cannot be traced through time, as there is no record of the change of length with time. Therefore, incidents that impact the bottleneck along the length cannot be included in the analysis. However, from theory, if an incident upstream in the bottleneck constricts the flow, it may momentarily cause the bottleneck to show a longer length, but

the reduced flow downstream of the incident might cause the bottleneck to dissolve. Therefore, a major incident will usually cannibalize the parent bottleneck, and spawn another, which may not be a frequent bottleneck given that it is not caused as a result of shortage of capacity [100].

Additionally, an inconsistency sometimes arises with the date-time stamp when the algorithm attempts to merge different TMC segments together. The algorithm may fail to properly aggregate parts of the bottleneck, resulting in the same bottleneck occurrence getting fragmented. As an example, if a bottleneck starts at 4PM and ends at 7PM, there could be a fragment from 6:30PM to 7PM. These multiple records are deleted from the table to avoid duplicate counts. The deleting does not affect the bottleneck count as it just removes subsets of a larger bottleneck.

11.2 Geospatial Datasets

There are two main datasets that have a geospatial component. Additionally the TMC linear referencing system is described in this section.

11.2.1 Highway Performance Monitoring System (HPMS)

The HPMS is a federal dataset that aggregates road geometries submitted by state and regional transportation agencies. The datasets cover all major roads, under federal, state or local jurisdiction. Major decisions about allotment of funds for roadway improvement, maintenance or expansion are based on analysis of HPMS data. The HPMS public release of geospatial data in shapefiles format is an open, free dataset obtained from the FHWA Office of Highway Policy Information. The details of this dataset are available online [82].

The data associated with each link in the HPMS describes the physical properties of the roadway, like functional road class, number of lanes, width of lanes and shoulders,

thickness of pavement, and other such properties. It also contains information about the road number or name, where available, and the Average Annual Daily Traffic (AADT). The roadway segments are usually arbitrarily split, unless when the roadway crosses state or county lines, or when AADT or other physical properties of the roadway changes significantly. Also, a vast number of segments are exactly as long as one U.S. chain length: 160 m (or 0.100 mile). The total length of the HPMS network all over the country is about 2×10^{10} meter (1.200 million miles).

11.2.2 TomTom Multinet

TomTom Multinet has two data subsets: a geospatial map layer with road segment geometries and metadata, and a table that provides typical weekly speed profiles on all segments. This is a commercial dataset and is available only by purchase. The geospatial data exist in shapefiles and contain the free-flow speed along with other information, such as road number and name (where available), Functional Road Class (FRC), length, number of lanes, whether part of a toll road, etc., for each segment. The Multinet segments are divided arbitrarily using proprietary algorithms such that resolution of the speed data is preserved. The segments are uniquely identified and referenced to the map by the OpenLR open standard proposed by TomTom. The Multinet road network is an order of magnitude longer than the HPMS. It has a total length of 10.600×10^{10} m (6.600 million miles) over all of U.S. Multinet covers every road, including the ones within parking lots, and both directions of roadways in the dataset.

The typical speed profiles for the week are given as a percent of free-flow speeds in the geospatial datasets. These percentages are provided at 5-minute epochs in a week, making 288 such epochs in a day. The typical speed during a 5-minute epoch is obtained by the product of the percentage at that epoch and the free-flow speed, for a given day of the week. There are three different free-flow speeds in the

geospatial dataset for each segment. The first is the speed typical for the whole week, the second is the typical speed during weekdays, and the last is the free-flow speed during weekends. There were minor differences between the three speeds, mainly because these speeds are computed from observed global positioning system probe readings [61]. It was ensured that the typical profile percentages were multiplied by free-flow speeds from the correct day, i.e., weekday or weekend where available, and the combined speed where not available. Alaska did not have a profile table; consequently, only a spatial join was performed for Alaska without the disaggregation.

11.2.3 Traffic Message Channel (TMC) Codes for Linear Referencing (LR)

The roots of the TMC standard emerge from the ALERT-C for conveying (encoding and broadcasting) information on radio sidebands. Currently, TMC codes are governed by Traveler Information Services Association (TISA). As a regulatory authority, TISA is responsible for maintaining the standards book for TMC codes, and other segmentation schemes such as ALERT-C and TPEG. Further, TISA keeps a look out for segment encoding standards created by the industry, and standardizes them if necessary.

The standards maintained by TISA only serve as the guideline to create a TMC table. The specifications are laid out in International Standards Organization (ISO) standard documents: ISO 14819-1, ISO 14819-2, ISO 14819-3, and ISO 14819-6. These standards merely supply the procedure to segment the roadway, and the method to assign codes to them. The segments run from one break in access, ingress or egress, to another. Therefore, there are two types of TMC segments, internal to and spanning the length of an intersection (between on egress break and the next ingress break), and external spanning between two intersections (ingress break of the previous intersection to egress break of the next intersection).

The North American TMC Code Alliance (NATMCCA) or the North American Location Code Alliance (NALCA) is responsible for maintaining a TMC table for U.S. and Canada. The table contains a list of TMC codes, along with fields which bind codes to the actual location on a map. The binding fields contain the start and end coordinate points and the road geometry of the TMC code, and other pertinent information which can uniquely identify the location of the segment. These tables are created in accordance with the TISA standards mentioned above.

Since the standards are at best a guideline, TMC segments from different data vendors differ in their encoding. Although the segment might have the same code, the length of road network represented might be different. For example, unlike the TMC representation used by TomTom, the TMC segmentation used by HERE Inc. does not have the internal segments, and they are merged into the downstream external segment. That is, the TMC segments reported by HERE Inc. run continuously from one egress along a roadway to the next. These differences cause compatibility issues when comparing data reported to TMC segments.

11.3 Other Datasets

There are two more datasets that have been used for the performance measurement methods documented in this dissertation: the Bluetooth re-identification data, and the incidents dataset obtained from the Maryland Coordinated Highway Action Response Team (CHART).

11.3.1 Bluetooth Traffic Monitoring (BTM)

The Bluetooth re-identification Data (Bluetooth data) were collected using Bluetooth sensors placed at the sides of the roads [107]. The sensors are powered by a rechargeable battery and last up to two weeks on a single charge. They are deliberately placed at

the ends of TMC segments so that the Bluetooth and GPS probe data can be directly compared. The Bluetooth sensors read and store the MAC address of passing visible Bluetooth devices. The time of detection is also stored with the MAC address. A Bluetooth sensor makes a sweep every few seconds, therefore, a single device can be recorded multiple times as long as the device is within Bluetooth range.

Re-identification involves the process of identifying the same MAC address in both upstream and downstream sensors. Once the same devices are detected, then a difference in recording time gives the travel time between the upstream and downstream sensors. Since a sensor can record a device multiple times, either the first or the last observation is selected to determine the travel time. The data used in this study was compiled by computing the difference between the last observations of the same address at upstream and downstream sensors. This data, is also plagued with noise, outliers, and non-targets, like pedestrians and bicyclists. A filter based on the Inter Quartile Range (IQR) is applied to the raw re-identification data [6].

The Bluetooth data so collected are representative of the actual traffic conditions, and is equivalent to data collected by the floating car, license place matching, and similar methods. Therefore, the Bluetooth data is used as benchmark, and the VPP and NPMRDS datasets are compared against it to gauge accuracy and fidelity. Comparing with the Bluetooth reveals that the proposed method of computing performance measures from NPMRDS is not biased.

11.3.2 Incident Dataset

Incidents are recorded by the Maryland Coordinated Highways Action Response Team (MD-CHART). CHART officials are usually first responders at the scene of an incident due to the large network of patrolling. Also, local fire, police and other first responders also directly communicate with CHART to coordinate responses. The main objective of CHART is to improve operations on the freeways of Maryland. CHART also

administers the 511 driver information system in Maryland which provides traveler information via phone, radio and Internet. CHART is one of the first organizations to be notified of any incident or event on the highway system.

The accuracy of incident locations is fundamentally limited due to the prevailing practice of recording incidents in the CHART system. Operators enter the details of the incident location by noting the nearest intersection in term of crossroads. The software then automatically assigns the coordinates of the center of the intersection to the incident. Note that the intersection may not physically connect the roads, especially at interstate overpasses or underpasses. As such the geodesic coordinates stored with the incident record only provides location accuracy to the nearest crossroad. This lack of precision, however, does not impact the analysis, as all incident data are attributed to the nearest TMC. Hence, as long as the incident is geocoded sufficiently to identify the nearest TMC on which it occurred, its geographic placement within the TMC is inconsequential.

Some incidents have incorrect geographic coordinates, and incidents usually appear outside the borders of the state of Maryland when mapped. Other incidents have no reported location data as they are typically warnings related to weather, system alerts, etc. processed through CHART. Such incidents are excluded from the analysis. Another inconsistency is with the start and end date-time stamps which are sometimes incorrectly recorded. An erroneous date-time stamp does contribute to some errors in the final analysis results. Obvious erroneous incidents, like accidents lasting just minutes, have been removed from the database.

11.4 Preprocessing Data

The different methodologies required different preprocessing steps before the actual analysis documented in chapter 12. Study-wise preprocessing performed on the datasets are documented in this section.

11.4.1 Preprocessing VPP Data for Forecasting

Only VPP data are used in the forecasting study. Data from weekdays in three weeks in September 2016 are examined in this study. The assumption underlying the synthetic method is that the model coefficients do not change in one week, models fitted to data from the week of September 12 are used to predict speeds for the week of September 19, and models fitted to the week of September 19 are used to predict speeds for the week of September 26. A total of 1 440 observations (rows) comprise the data for one day from one TMC segment. The total dataset contains data from 5 weekdays for 3 weeks for 2 654 segments. Therefore, the total number of records used are more than 57×10^7 (21 600 records per segment, or over 19×10^7 observations per week).

Predictions for every minute up to 30-minutes in the future are produced for each segment. The predictions dataset thus has 30 times the data used for fitting. The authors are unaware of an online algorithm or model that would efficiently scale with this big data. All models proposed and studied in existing literature would require a huge amount of processing power to be fit and then make forecasts on such a large dataset. The synthetic method, however, takes under 3 hours on an 8 core computer to fit 27 $ARIMA(p, d, q)$, where $p, d, q \in 0, 1, 2$, models to all these data, and make predictions.

The network studied is shown in figure 11.2. It comprises about 2 000 lane-miles, ranging across almost all classes of roadway (from collectors to freeways/interstates), in Maryland. A total of 2 654 individual, directional segments completely describe this network. The most reasonable model is selected for each segment individually from the 27 possible orders for the Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA model does not have any auxiliary covariates, and the time spacing of the data is 1-minute. Once a model is selected in the synthetic framework for a segment,

predictions are made using the incoming stream of data for that segment. Predictions up to 30-minutes in the future are made using the selected model for each reported data point, received once per minute. This procedure is repeated for each day in the three weeks studied here.

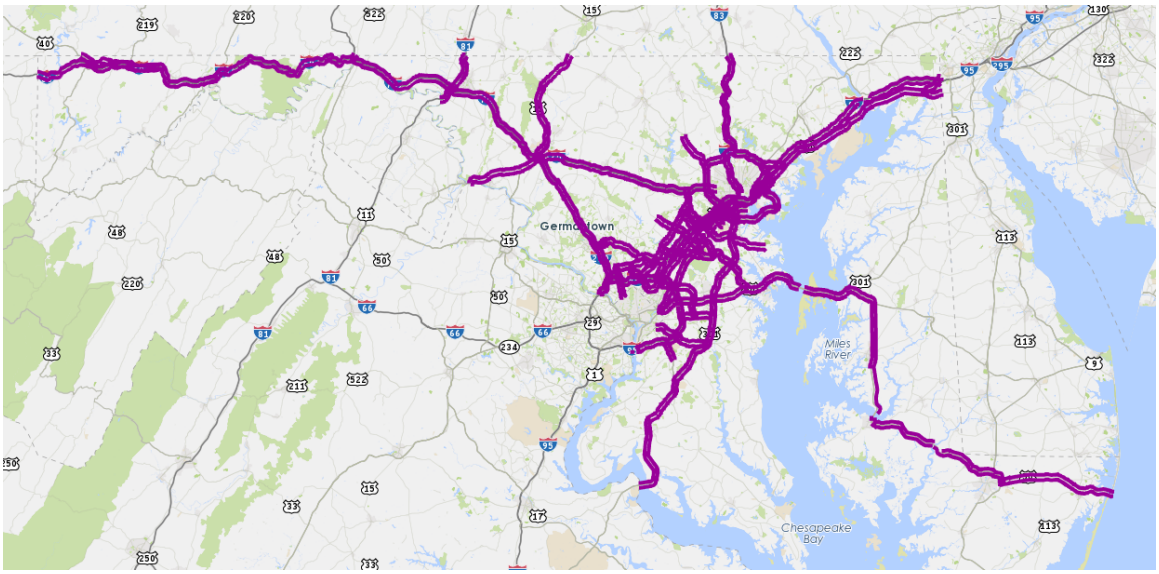


Figure 11.2: Maryland Important Roads Network

Owing to the advantages of the synthetic time series framework, the method is also very scalable. 27 ARIMA models are fitted to the 2 654 roadway segments comprising the network studied. The total number of models thus fit are 71 658 per day. Since data are reported at 1-minute intervals, amounting to 1 440 records per segment per day (over 3.820×10^7 data points per day for all segments) this is truly an exercise in big data. The two weeks in September to which ARIMA models are fitted totals to over 16×10^7 records. Despite the size of the data, the synthetic method allows ample time to fit the 27 models to each segment, models are fit on data from each day for all segments, and select the most reasonable one.

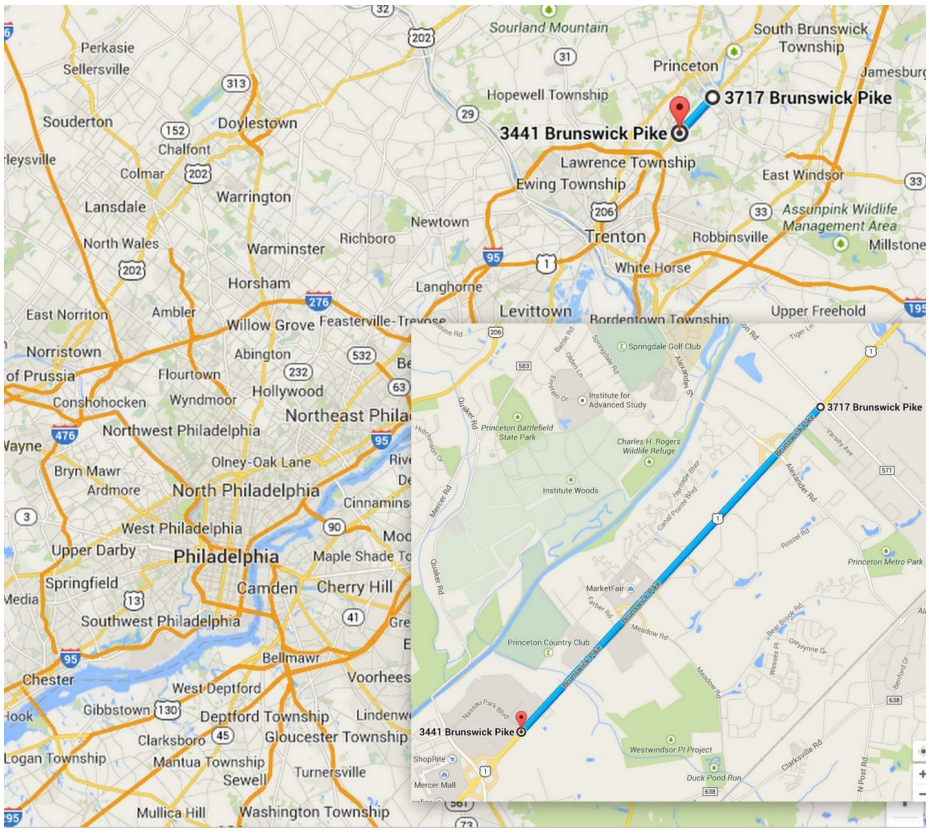
The novel contribution of the documented method is also the size of network handled. This is an unachievable task for models fit in real-time, as data points are received every minute. On a network as large as the one studied, it would be

unthinkable for models to finish computing before the next set of data arrive, especially on conventional hardware. Consequently, predictions on networks this large are neither recorded in literature [7, 50, 65, 83, 110, 118, 119, 121, 122, 137], nor are conducted by giant tech companies with very huge compute hardware infrastructures. Additionally, the synthetic method produces forecasts with reasonable prediction errors, which is primarily attributable to the fact that the model is trained using the complete set of data, containing the expected traffic patterns.

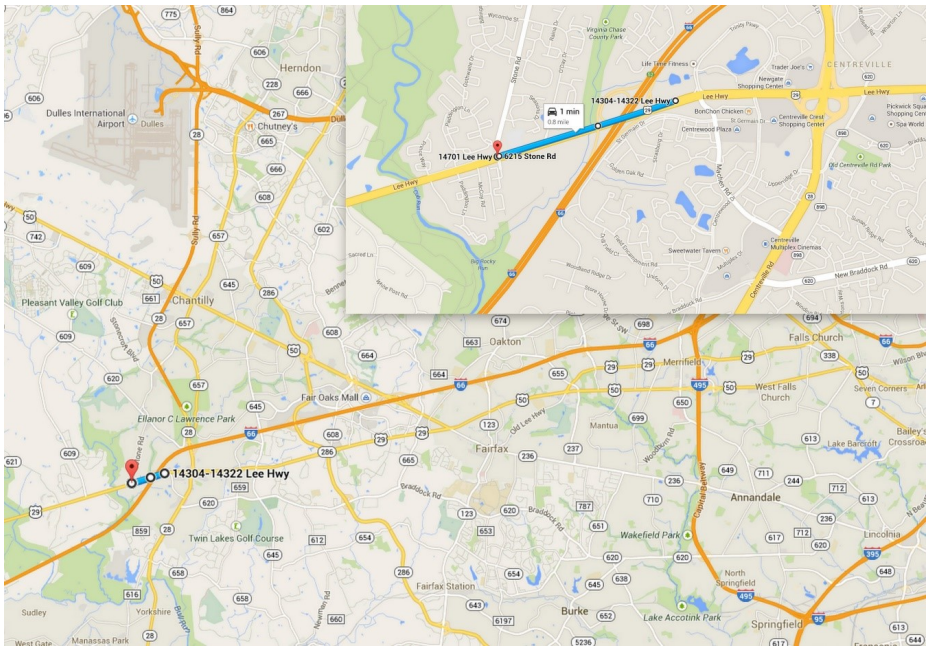
Due to communication artifacts, and other delays, VPP data records are not timestamped at the exact minute, but can be received at any second in a minute. Sometimes, no record is received for a minute, but the preceding or succeeding minute might have two observations. Even more rarely, no data is received for an extended period of time. For the purposes of this study, it was deemed to have the data timestamped perfectly at the minute, which is also a requirement of the ARIMA models. To achieve this, speeds were linearly interpolated at the exact minute using adjacent observations.

11.4.2 Preprocessing Datasets Used for Performance Measurement

The NPMRDS, VPP, and Bluetooth data were collected for two weeks (10 weekdays) for two case study segments. The data for each case study were collected at different two week periods. The case study segments are from two arterials, with different characteristics. One arterial, NJ11-0009, has fewer frictions to traffic, less signals and intersections and higher speed limits as compared to the other segment, VA08-0012. This causes traffic to behave on NJ11-0009 as if it were a freeway, and on VA08-0012 as if it were a congested downtown arterial. Table 11.1 presents a summary of the characteristics of both segments, while figure 11.3 shows their map locations.



(a) Geographic location of segment NJ11-0009



(b) Geographic location of segment VA08-0012

Figure 11.3: Geographic locations of performance measurement case study segments

Table 11.1: Summary of the Segments Chosen for Study

Path ID	Dataset	NJ11-0009	VA08-0012
Road Name	All	U.S.-1	U.S.-29
Direction	All	SB	WB
Number of Lanes	All	4	4
Number of Signals	All	2	4
Segment Length (miles)	Bluetooth	3.010	0.780
	VPP	3.010	0.780
	NPMRDS	3.440	1.280
Number of TMC Segments	VPP	1	3
	NPMRDS	1	2

11.4.3 Preprocessing Geospatial Datasets before Conflating

All roadways with measured non-zero AADT values from 50 states and Hawaii in the U.S. were conflated in this study. Figures 11.4a and 11.4b show the network from both HPMS and TomTom Multinet datasets before the conflation process. Figure 11.4a is a birds-eye view of Denver, Colorado, while figure 11.4b is a zoomed-in section around an interchange. In figure 11.4b, individual segments are demarked by black dots. The intersection a little southwest from the centers of the maps in figure 11.4b is I-70 and Denver West Colorado Mills Parkway. I-70 is clearly seen represented using bidirectional segments in the HPMS map, while Multinet shows each direction of travel on I-70 using separate one-way segments. Southwest of the intersection is a huge shopping complex; the dense network of Multinet segments show the rows of parking lots. The figures place a perspective on the details captured in the Multinet dataset.

Figure 11.4b also explains the major difficulty in the conflation process. North of the intersection is Denver West Parkway, a road parallel to the interstate. The conflation should ensure that the Denver West Parkway segments are joined to the corresponding segments from HPMS and not to the Interstate segments, and vice-versa. The developed algorithm of computing parallel and perpendicular distances

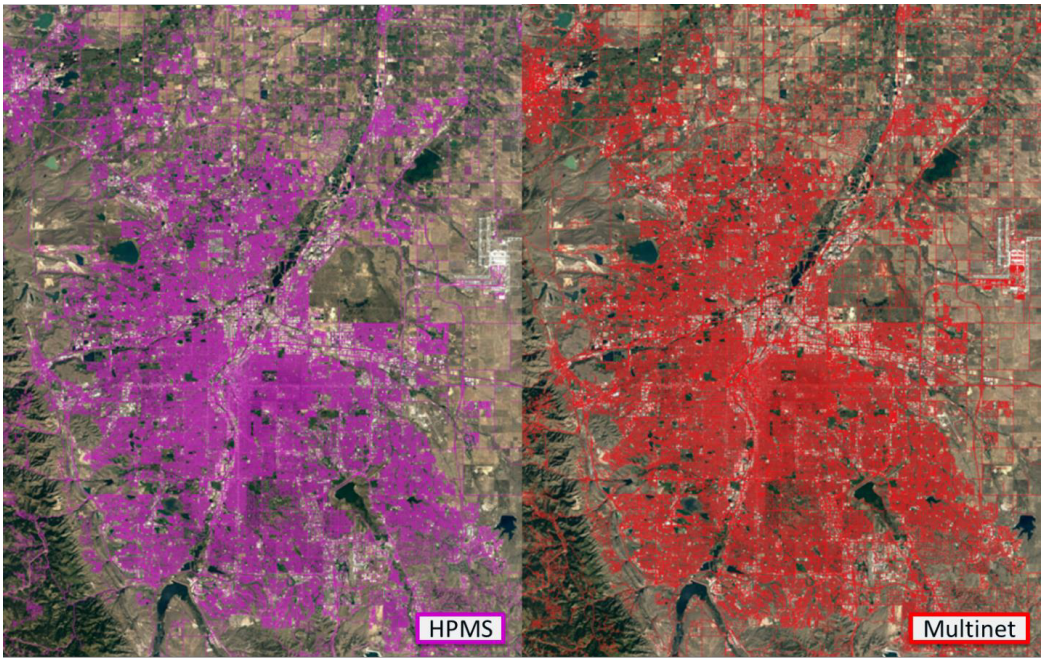
between the segments and scoring them based on closeness overcomes this difficulty to an extent. However, some pre-processing was required before applying the algorithm, as explained below.

Common Fields

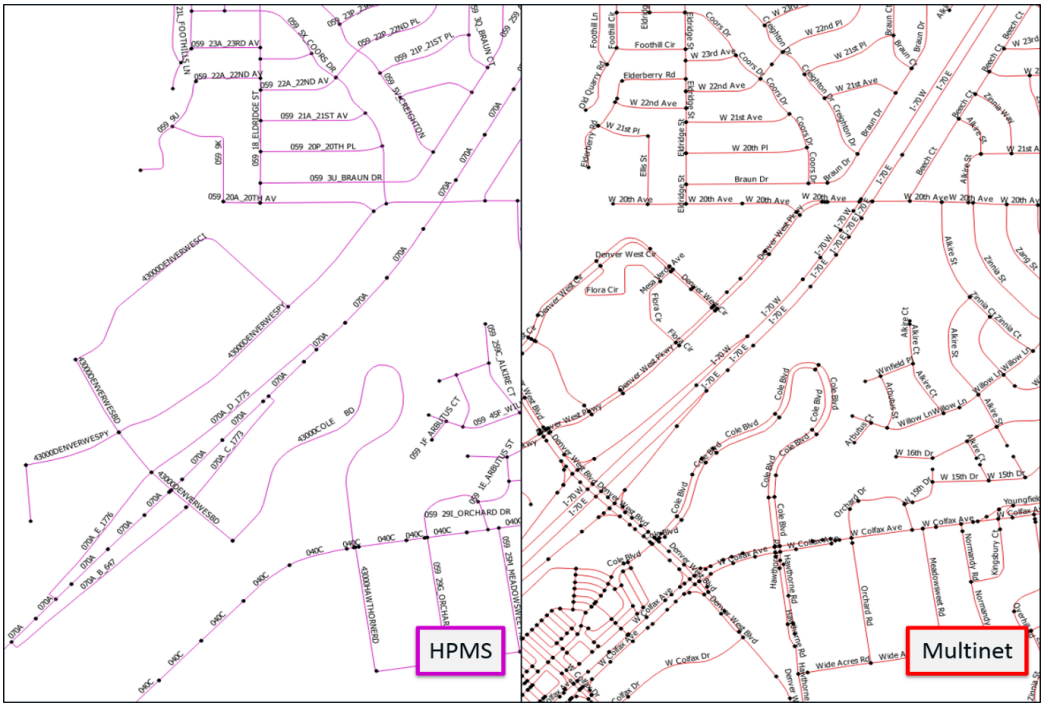
Some fields are common in the two datasets, like road numbers and names. However, the information and its representation format varied widely. For example, the Multinet data might include all shield numbers from a segment on a road with multiple numbers, whereas the HPMS might just include one, or even none, partly because the segments in the latter are often determined by a change in the AADT values, and roads with other shield numbers may merge or diverge within a segment. The shield numbers may themselves have different notations, like the inclusion of hyphens, spaces, back or front slash, or other separators to separate the class identifier and the actual road number (e.g. I-95, I95, I 95, I/95, etc.). The most difficult problem, however, was the lack of road names and shield numbers altogether for many of the segments.

The datasets also encoded whether a segment was part of a toll road. This field was useful for an accurate join, however, the road miles that are tolled are insignificantly small as compared to the total length of roads. Another field recorded the number of lanes on the highway, but the count for HPMS included both directions of travel while Multinet counted the lanes on one carriageway only. On roads with an equal number of lanes in both directions, this is potentially not a problem; however, on one-way streets or roads with uneven number of lanes, there is no clean way to match the lane counts from both datasets.

The last common field is the FRC. However, both datasets use completely independent classification systems. The HPMS uses FRC values from 1 to 7 while Multinet represents FRC from 0 to 11. Often, even the same road may have different FRC values. For example, all grade-separated highways are FRC 1 in the HPMS whereas



(a) Birds-eye View of Denver, Colorado, as rendered by geometries in the HPMS and Multinet datasets.



(b) Birds-eye view of a small area in Denver, Colorado, zoomed in to show geometry details.

Figure 11.4: Comparison of Networks before Conflation

in Multinet, they may take values from 0 to 2, depending on whether the highway is an interstate, a national U.S. road, or a state highway, and the length for which it is grade separated. Further see table 12.1 for a summary of matching fields.

Data Preprocessing

Pre-processing applied to the HPMS splits long segments into shorter components to harmonize the geometric distance computations. On the other hand, FRC classes 7 and above (lower in significance than minor collectors) were removed from the Multinet dataset in addition to filtering out ferries and other segments not part of roads. The following subsections detail the HPMS pre-processing, creation of a geospatial index, and computations of the absolute angles of the segments with respect to the equator. The angles are used to compute the parallel and perpendicular distances between segments from the two datasets.

Splitting HPMS Segments: Not only did the lengths of the HPMS segments vary widely, but the average segment length was also very different from the mean Multinet segment length. Since the conflation algorithm mainly depends on calculating the spatial separation between the segments, having coherent segment lengths is of paramount importance. As the difference in lengths of two adjacent segments increases, the distance between their centroids can take a larger range of values, reducing confidence in eliminating bad join candidates.

For example, consider a hypothetical HPMS segment, which is about 1 km long (0.600 miles). Adjacent to this segment are tiny (by comparison) Multinet segments, each 100 m long (328 ft). The distance between the centroids of the HPMS segment and the Multinet segment can vary between 0 m (0 ft) near the center and 550 m (1 804 ft) near the ends, assuming minimum overlap. This is a large difference in lengths for a typical road network, as the road can curve away, intersect with other roads, etc. within that distance. Therefore, the number of join candidates increases

to the point where selection becomes difficult, even impossible.

Consequently, the best strategy was to split the HPMS segments into smaller pieces. All HPMS segments longer than 200 m (656 ft) were selected and broken down into segments 160 m (528 ft) long, which is the standard U.S. chain length, and forms the basis of the HPMS map geometries. HPMS segment lengths, consequently, tend to cluster at multiples of this length. Some states merged these chain-length segments together, only splitting them when the AADT values changed. Splitting at 160 m also reduced the occurrence of very short segments formed by the left-over bits, which reduces computational burden. The splitting cut-off was derived from a histogram of the Multinet segment lengths, which peaked at 200 m.

Longer line segments are often stored as multilines in map geometries, which are just a collection of shorter, regular line segments. The shorter lines encode a start and an end, between which the segments are usually rendered as straight lines. This allows a line to easily follow a curve in a piecewise approximation. To preserve the geometry, the splitting algorithm first separated all multi-lines into constituent segments. Any constituent segments longer than 200 m were then subdivided into 160 m segments by linearly interpolating between the start and end coordinates. Vincenty's formula for computing distances on geodesics was used to compute all lengths throughout this chapter [120]. Vincenty's method works better when points are very close together.

The last two digits of the HPMS identification code are the state Federal Information Processing Standards (FIPS) code. The rest of the code can repeat in different states. Therefore, to create a unique identification code after splitting, a three-digit number (with zero-padding as required) was inserted before the FIPS code. The three digits were a linear count of the subsections created from a given HPMS segment. Some segments were split into more than 200 sections. Hereafter, HPMS refers to the new database formed by the split HPMS segments.

Creating a Geospatial Index: A geospatial index serves to limit the search space

for matching segments. For example, to join segments on the west side of Kansas, segments on the east side need not be searched. The index was created by applying a buffer to the HPMS segments and storing the buffer extremities in an R-tree [44]. R-trees work by bounding smaller boxes within larger boxes, similar to real-world addresses (see figure 11.5). During a lookup, the R-tree is traversed until all the buffers that enclose a provided point are identified. The buffer radius around HPMS segments was fixed at 500 m (1 640 ft), which in hindsight is a very large distance. A buffer of 100 m (328 ft) would have been sufficient and would have significantly reduced the runtime of the complete algorithm.

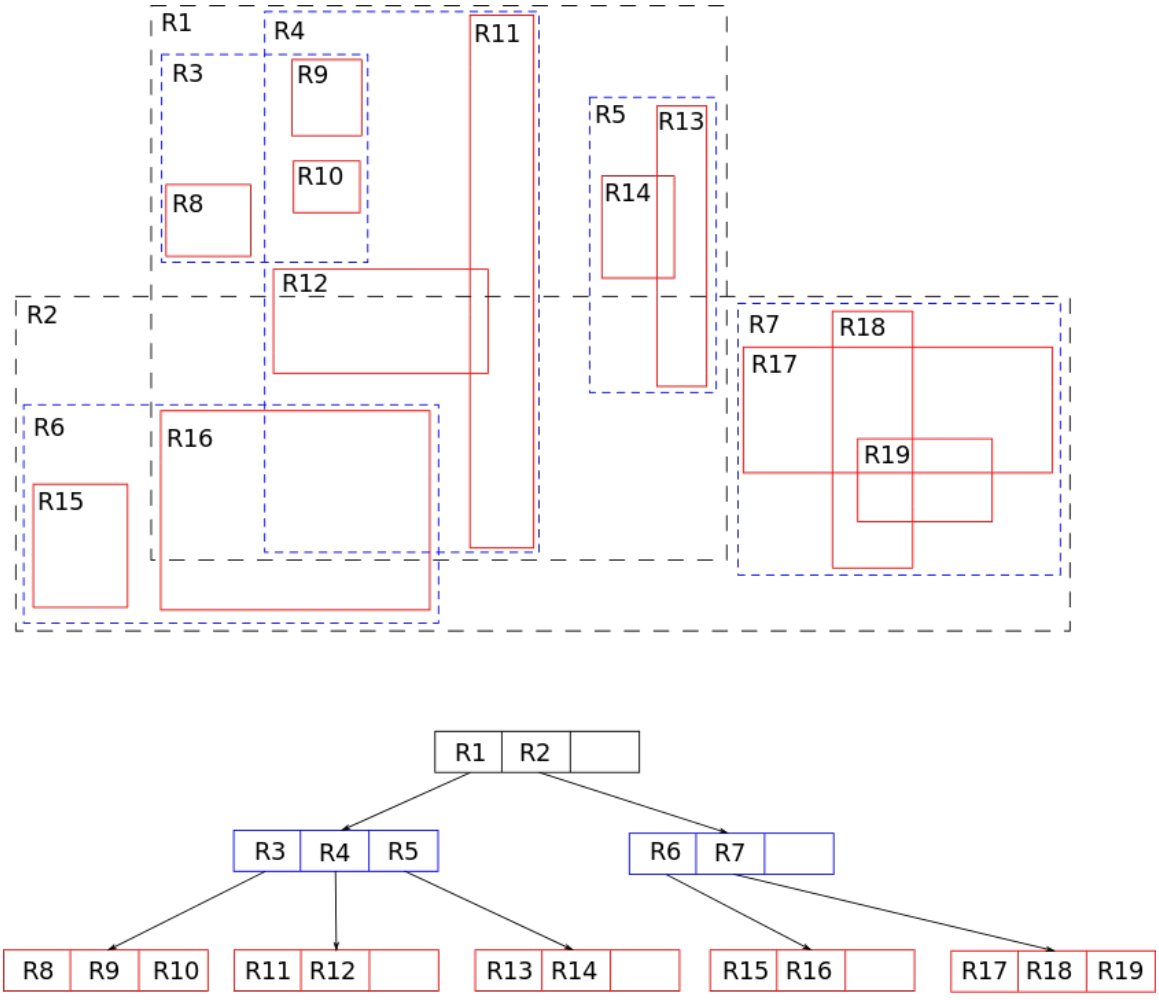


Figure 11.5: Schematic Representation of R-tree Indexing [108]

The goal of the conflation was to create a one-to-one join between each Multinet

segment to the best candidate HPMS segment. Therefore, the algorithm exhaustively combs each valid Multinet segment to find the ideal join partner. However, the join was not forced; without an ideal HPMS segment, the Multinet segment was dropped from the join. Since the lookup was from Multinet to HPMS, the geospatial index was created only for the HPMS dataset. Candidate HPMS segments were selected by querying the index, and join strength was determined only for selected segments.

Computing Multinet Typical Speeds: The profile speeds were used for the temporal disaggregation of the AADT values transferred through the spatial join. Speeds were attributed to each Multinet segment through a simple relational database join between the geospatial dataset and the typical week profile tables. The latter tables repeat each Multinet segment ID seven times, once for each day of the week. Therefore, the join was pivoted so the appropriate free-flow speed column was joined to the correct day: weekday free-flow speeds to weekday profile rows, and for weekends, given a Multinet segment. If the speeds were unavailable for a weekday or weekend day, the generic week free-flow speed was used.

11.4.4 Merging Incidents and Congestion

Both incidents and congestion data were subjected to filtering by time and space, and descriptive analysis as documented. Only data from 2011, collected on the most important roads in Maryland are used.

Preprocessing Bottleneck Data

Data collected during 2011 was used in the study conflating congestion and incidents. The congestion is measured by (11.1). Table 11.2 provides a summary of all bottlenecks recorded in Maryland during 2011, while figure 11.6 shows the locations of the heads of all bottlenecks over the same space and time. As evident in the figure, the majority of bottlenecks appear to happen along the interstate and important routes spanning

the length of Maryland. This is reasonable because the interstates carry most of the traffic, and are bound to be first affected by increasing demand.

Table 11.2: Summary of Bottlenecks

Quarter	Unique Occurrences	Total Occurrences
1	729	26 728
2	752	46 094
3	754	49 515
4	740	48 921
Total	879	171 258

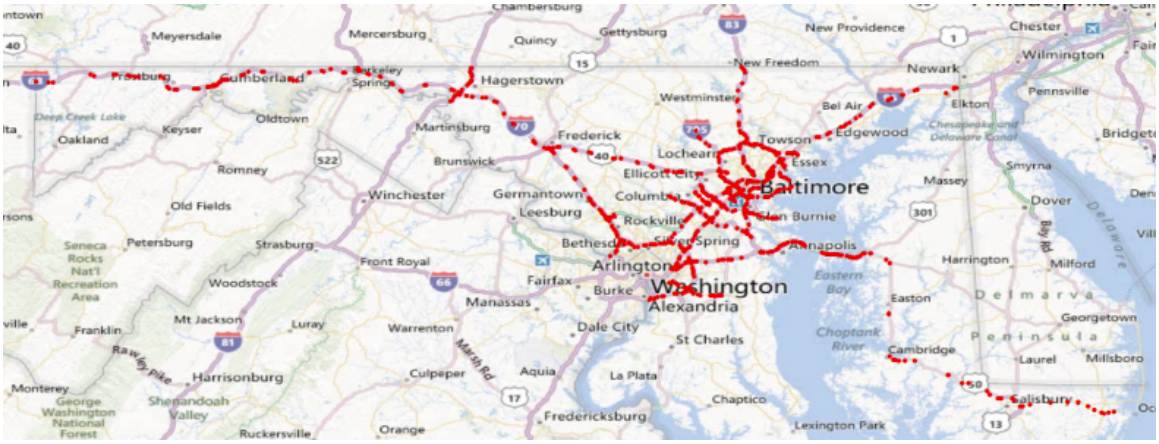


Figure 11.6: All bottlenecks in Maryland recorded in 2011

Preprocessing Incident Data

Table 11.3 provides a summary of the incidents recorded by CHART in Maryland during 2011. Figure 11.7 plots all the incidents recorded by CHART during 2011 on a map.

CHART groups incidents into 17 classes, some with up to 4 sub-classes. These form a complete picture of the incidents and alerts that affect the roadways of Maryland. Table 11.4 provides a full count of all incidents, by type, that affected the road network of Maryland. The counts are divided by quarter, and are further separated to show

Table 11.3: Summary of Incidents

Quarter	Number of Incidents
1	20 085
2	23 356
3	27 510
4	24 073
Total	95 024

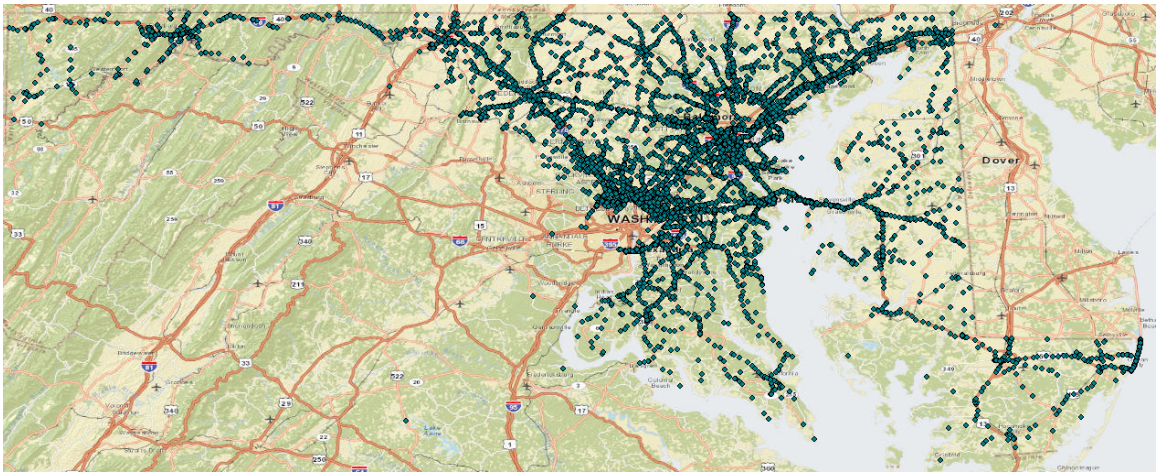


Figure 11.7: Map of all Incidents Recorded by CHART in Maryland during 2011

the count of impacting and total incidents. Impacting incidents are those that lie within one mile of the head of any bottleneck in Maryland.

Table 11.4: Highlighted Incidents Summary

Standardized Type	Quarter 1		Quarter 2		Quarter 3		Quarter 4	
	Impacting	Total	Impacting	Total	Impacting	Total	Impacting	Total
Alert	924	2 715	507	1 475	506	1 717	457	1 373
Animal Struck	49	138	156	373	87	235	240	750
Collision	1 856	2 445	1 961	2 436	2 000	2 507	2 267	2 892
Congestion	141	259	371	415	393	504	271	317
Disabled Vehicle	4 449	5 478	10 346	11 769	10 303	12 788	9 179	10 620
Disturbances	9	28	8	34	99	374	12	55
Emergency Roadwork	63	78	89	127	101	136	95	124
Flood	1	1	5	29	1	1	0	0
Incidents	4 148	6 000	384	576	366	997	405	738
Obstructions	693	1 062	1 280	1 721	1 215	1 968	993	1 305
Road Construction	42	61	0	0	0	0	0	0
Road Maintenance Operations	671	843	2 093	2 574	2 356	3 893	2 604	4 593
Special Event	15	24	99	173	44	134	48	75
Traffic Signal Not Working	269	803	441	1 518	667	2 118	361	1 097
Vehicle on Fire	76	98	107	127	100	129	84	102
Weather Condition	15	52	5	8	3	6	11	30
Work on Underground Services	0	0	0	1	0	3	1	2
Total	13 421	20 085	17 852	23 356	18 241	27 510	17 028	24 073

Chapter 12: Methodology and Algorithms for Performance Measurement

The methodology used for each study in performance measurement is quite different from the others. Each section in this chapter is dedicated to explaining the method used to affect the goal of the study.

12.1 Synthetic Time Series Method

To explain the new synthetic modeling approach for real-time prediction, consider a given road segment and a given day of a week (e.g., Wednesday). The predictive models are likely to differ from one road segment to another or from a given day of a week to another, but the general concept and methodology will remain the same. The basic idea of the proposed synthetic method is to use historical data for both complex model building and model fitting for each road segment separately and simply use the fitted model for forecasting speed in real-time. Thus, the forecasting will be instantaneous.

The underlying assumption made by using this framework is that traffic patterns do not vary for the same day on different consecutive weeks, although, the second half of this assumption may be altered as per convenience. Patterns may be assumed to repeat monthly, or fortnightly, or even annually. Further, the underlying model may be constructed to use auxiliary information such as weather forecasts, history of traffic incidents, or average traffic speeds from some other time periods. The unwavering

part of the assumption is that conditions do not change significantly from a historic observation set to the current observation set. In this study, the repetition interval is taken as one week, and data from the whole day (1 440 records) forms the observation set.

The entire day can be viewed as a collection of T time points. For example, as consistent with the data used here, $T = 1\,440$ if the speed data come in one minute interval. Let $y_{t,w}$ denote the speed at time t in week w. Our goal is to predict speed at a future time point $y_{t+h,w}$ based on the observed data in the previous week $w-1$ and the observed data at or before time point t in the current week w. Consider the following general time series model:

$$\psi_w(B)(1-B)^d y_{t,w} = x_{t,w}^T \gamma_w + \eta_w(B) z_{t,w}, \tag{12.1}$$

where

$\psi_w(B) = 1 - \psi_{1w}B - \cdots - \psi_{pw}B^p$,

$\eta_w(B) = 1 - \eta_{1w}B - \cdots - \eta_{qw}B^q$,

B is a back shift operator: $B^d y_{t,w} = y_{t-d,w}$,

$z_{t,w}$ are white noises that follow normal distributions with zero means and constant variance σ^2,

$x_{t,w}$ is a $s \times 1$ vector of known auxiliary variables,

γ_w is a $s \times 1$ vector of unknown fixed coefficients,

$\psi_{1w}, \cdots \psi_{pw}$ and $\eta_{1w}, \cdots \eta_{qw}$ are unknown model parameters.

Note that above model is the standard Autoregressive Integrated Moving Average (ARIMA) model, denoted by $ARIMA(p, d, q; \phi)$, where $\phi = (\psi, \eta, \sigma^2)$, with the regression term $x_{t,w}^T \gamma_w$ added to borrow strength from external set of relevant auxiliary variables such as weather or accident data available at the current time or historical traffic data from previous weeks. The above class of distributions of speed is flexible

enough to model the distribution of speed for a variety of road segments.

In a direct forecasting approach, to predict $y_{t+h,w}$ time series data up to and including time point t in the current week w only is used. That is, for each forecasting at time t, first a specific model is selected from the above class of models, that is, orders p, q and difference d are selected, and then different parameters (ϕ) of the selected model are estimated. This is a cumbersome and time consuming process and is clearly not feasible when the same is needed to be done for thousands of road segments in real-time and the procedure repeated as one moves along in the time scale.

A radically different idea of completing all the complex model building and estimation well ahead of time, say, a week in advance and then applying the chosen model with estimated parameters for the forecasting is now proposed. The following modeling synthetic assumptions is made:

$$\begin{aligned} \psi_w(B) &= \psi_{w-1}(B) \\ \eta_w(B) &= \eta_{w-1}(B) \\ \gamma_w &= \gamma_{w-1} \end{aligned} \tag{12.2}$$

That is, it is assumed that the model and the model parameters do not change in a week. Any model may be used here, parametric or non-parametric. The base assumption is that the model specification does not change from one week to another, or from one season to another where trends are expected to repeat. This assumption will allow us to select and fit models on data from the previous week, while the fitted model is then be used to forecast speed for the current week. The emphasis is that the real-time forecasting of speed for thousands of road segments will not pose any problem as model selection and fitting for all the road segments will be done one week ahead of time. Of course, one may question the validity of the synthetic modeling assumption, and therefore, the synthetic assumption is tested using real data.

Herein, the best order from the ARIMA family is selected. The ARIMA family is denoted by $ARIMA(p, d, q)$, where, p and q are the orders of the model while d is the differencing number. In our analysis, allowance for p, d and q to take any value from $\{0, 1, 2\}$ is made, a total of 27 combinations. The best ARIMA model is then specified by the combination of p, d and q that results in the lowest Bayesian Information Criterion (BIC).

All 27 models are fit to each of the 2 654 segments separately for each of the day examined. The framework is demonstrated with weekday data from two weeks in September 2016. Therefore, the models are fit on data from 10 different days. All in all, 716 580 models are fit to the data, out of which the best model for each segment for each day are selected, leaving us with 26 540 models. Predictions are made for the next week for the same day as the selected model. Therefore, a total of 10 days, spanning two weeks are predicted, amounting to over 38×10^7 predicted records.

12.2 Method to Compute Performance Measures from NPMRDS

The existing technique to measure performance on roadway segments was developed by Texas Transportation Institute (T.T.I.). The T.T.I. method has proved to be quite successful when working with cleaned and filtered data like that from the VPP, and forms the basis of the results presented in the Maryland Mobility Report [76]. However, the method computes averages of speeds or travel times before computing the performance measures. Since the NPMRDS has considerable outliers and noise, averages produce biased and incorrect results. The proposed method successfully overcomes the biases due to outliers and noise by overlaying the data from multiple similar periods to increase the density of data and dilute the outliers and noise. The densification caused by overlaying increases certainty in central tendencies of the data. While averages are still susceptible to outliers, rank-based methods like percentiles are immune to the effects of outliers within the Inter Quartile Range (IQR).

Travel Time Index (TTI) and Planning Time Index (PTI) are two of the most common performance measures. The equations for computing TTI and PTI by the T.T.I. method are provided in (12.3) and (12.4), respectively. These equations require averages of travel times to be calculated. Note that the reciprocal of equations (12.3) and (12.4) should be used when working with speeds, instead of travel times, as they are inversely proportional to each other. Since, the NPMRDS reports travel times, the equations pertinent to speeds are not presented.

$$TTI = \frac{\bar{t}}{t_{ff}}, \tag{12.3}$$

$$PTI = \frac{t_w}{t_{ff}}, \tag{12.4}$$

where

$\bar{t}$ is the average travel time,

t_{ff} is the freeflow travel time, and

t_w is the worst travel time.

Note that these metrics are quite difficult to obtain, and required many assumptions and untested roadside surveys of travelers. The proposed methods not only reliably compute the TTI and PTI indexes, but also provide a robust estimate for the freeflow, average and worst travel times.

12.2.1 Percentile Method for Performance Measurement

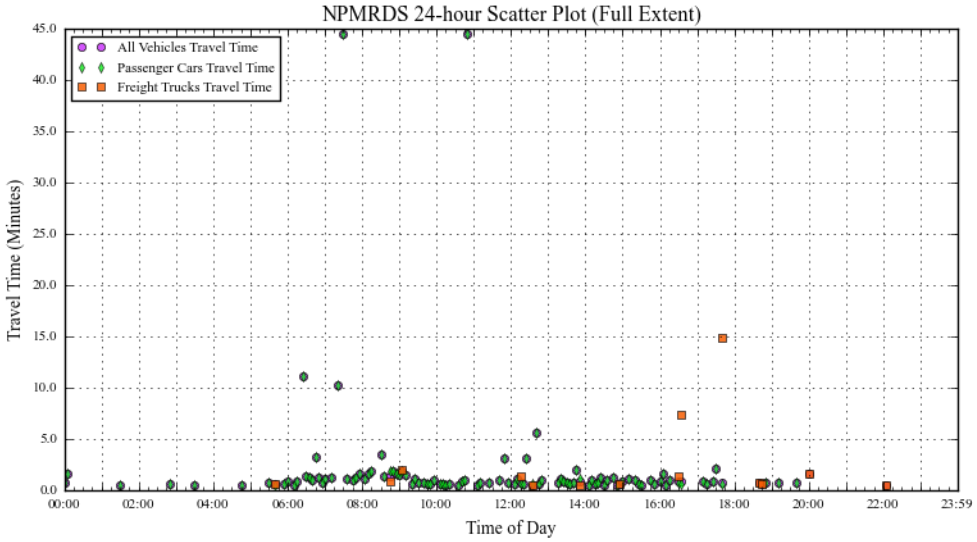
The innovative method proposed involves overlaying data from multiple similar observation windows. The lack of imputation during periods of no observations (during lean traffic) causes considerable gaps in the data. These gaps break continuity algorithms, for example, an algorithm that assesses whether the next observation is an outlier,

based on the trend of current observations will be unable to judge after a long gap. Similarly, algorithms that trace vehicle trajectory will also run into unbridgeable gaps in the data. The solution requires increasing data density, which would serve both functions of helping outlier identification, and providing intermediate data points.

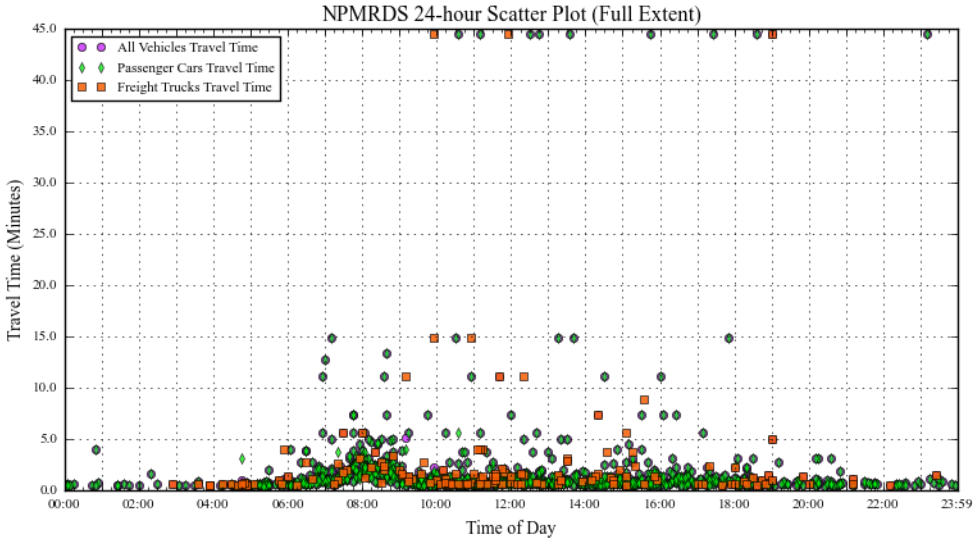
The simplest, direct, and most data-driven approach to increasing density is to simply overlay similar periods of data. This is illustrated using figure 12.1. Figure 12.1a shows NPMRDS data from an interstate segment for just one weekday, while figure 12.1b shows data from 10 weekdays. Immediately, trends invisible in figure 12.1a are clear in figure 12.1b. The morning peak increase in travel times causes a noticeable bump in the data, the outliers are visible at the top of the plot, while the noise prevalent in the afternoon and evening hours are dotted above the actual trend of the data.

To compute performance measures from these data, a simple technique involving percentiles is used. Traffic data has a singular advantage that outliers are easy to filter using percentiles, by discarding data at the extreme ranges. Percentiles of traffic data are also a good estimate of the trend of traffic, and can clearly identify the mean and extreme flows. Especially after overlaying, the trends become more pronounced as compared to the noise, and easier to identify. Quantile plots, also called cumulative frequency distribution (cfd) plots are shown in figure 12.2, where each line represents data from one hour. Every 5^{th} quantile is computed starting at 5 % and going up to 95 %. Even granular quantiles are possible, but it would defeat the goal of noise smoothing.

Using the cfd plot, hourly or daily performance measures can be easily computed. When comparing with the traditional way, the cfd offer some advantages, and one drawback. The advantages are that freeflow travel times, and congested travel times can be easily computed, and are the 5^{th} and the 95^{th} quantile respectively. The drawback is that the median of the quantiles differs slightly from the average of the



(a) Scatter plot of travel times from NPMRDS for single day



(b) Scatter plot of travel times from NPMRDS for 10 days overlayed

Figure 12.1: Scatter plot of travel times from NPMRDS illustrating overlaying method

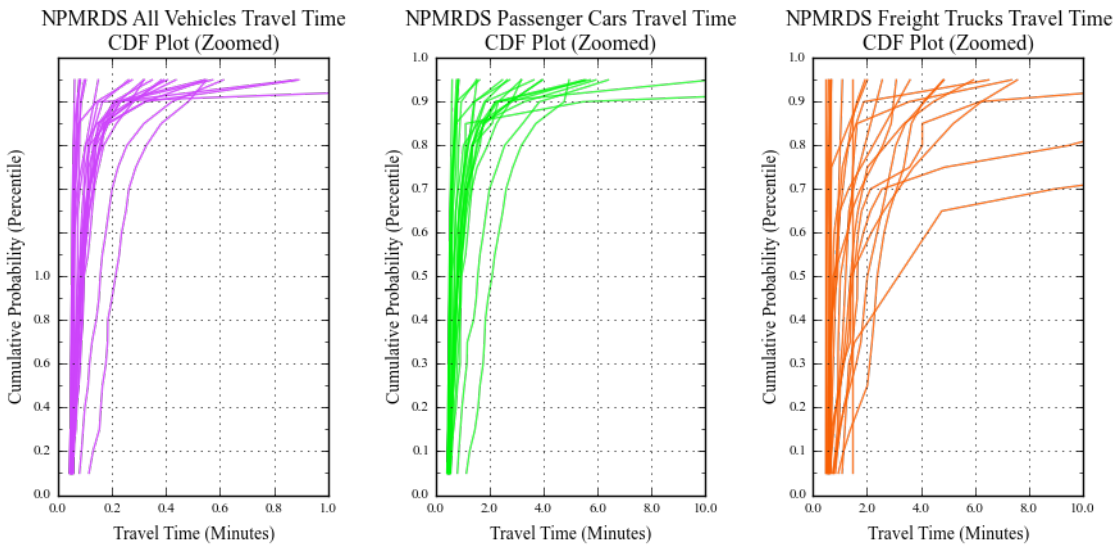


Figure 12.2: Cumulative frequency distribution plots produced from overlayed data in 12.1

data. However, since the NPMRDS is not filtered, average would lead to a more erroneous solution than the median. Equations (12.5) and (12.6) then become:

$$TTI_q = \frac{t_{50}}{t_5}, \tag{12.5}$$

$$PTI_q = \frac{t_{95}}{t_5}, \tag{12.6}$$

where

t_{50} is the 50^{th} percentile, or median, travel time,

t_5 is the 5^{th} percentile, or freeflow, travel time,

t_{95} is the 95^{th} percentile, or congested, travel time, and

the subscript q denotes the measures are computed from percentiles, rather than the T.T.I. method.

12.3 Methodology to Effect Spatial Conflation and Temporal Disaggregation

A scoring algorithm was used to select the best match for the spatial conflation. Since no common fields exist between the datasets, it was not possible to achieve quality conflation just by relying on the metadata for each segment. Further, it was not possible to use the geometries directly due to differences in the way the road segments were encoded. Consequently, a rule based algorithm was developed for the conflation which involved computing geometric measures of similarity among the line segments, in addition to whatever metadata matched.

Appendix A of the Mobility Scorecard [105] published by T.T.I. was referred to for disaggregating the AADT by time of day. The steps in the published appendix were followed exactly, including adjusting the AADT for the day of the week and computing the peak directions and the severity of congestion. The graphs shown in Exhibits A-2 through A-6 in the appendix were used to extract the percent daily volume for a given hour of the day for scenarios applicable to the respective exhibits. These percent points were linearly interpolated to obtain the percent daily volume for 15-minute periods in a day, which were then multiplied by the daily volumes obtained by adjusting the directional AADT for the day of the week. Note that this disaggregation would not have been possible without the join also supplying the typical speeds by day of week for each AADT value transferred to Multinet.

12.3.1 Computing Geometric Information

The Multinet segments were devolved into five points: 0 % (start), 25 %, 50 % (mid), 75 % and 100 % (end) of their lengths. Since Multinet segments are unidirectional and have well-defined starts and ends, the intermediate points fall in the correct order. On the other hand, since HPMS segments are bidirectional, the starts, ends and all

other intermediate points can be equivalently interchanged. The only stationary point along HPMS segments are the centroids. Therefore, the HPMS segments were only devolved into their centroids because using other points would have necessitated the computations be repeated assuming HPMS segments in reverse order, which does not apply universally to all segments.

The five Multinet points were used to build a set of candidate HPMS segments. The lengths and angles of the imaginary lines connecting the five points to the centroids of the candidate HPMS segments were computed. The lengths of the imaginary lines are then resolved into components para-llel and perpendicular to the HPMS segment using the angles. The resolved lengths from all five points are averaged together for each candidate HPMS segment. Additionally, the angle of the Multinet segment with respect to each candidate HPMS segment is computed. The average distances, along with the angles between the segments are used in attributing a score to each candidate HPMS segment.

Computing Angles

Angles between points are easily computed using the inverse tangent (arctan) formula. The inverse tangent function is provided by the Numpy package, a part of the Scientific Python Stack. The two-argument function (atan2) was chosen over the single argument function (arctan) to obtain the signed angle and avoid numeric infinities and division-by-zero errors when the segments are pointing true north, south, east, or west, i.e., when the difference between the start and end latitudes or longitudes is zero [87]. All angles were computed with reference to the candidate HPMS segments. Using HPMS segments was motivated by the bidirectional symmetry of the segments; therefore, the absolute angle can be used. Absolute angles are always between 0 rad to rad, which ensures that all distances are positive when resolved into parallel and perpendicular components.

Scoring Rules

Scores were given based on the computed geometric values and the similarity between the metadata recorded in other fields for each candidate HPMS segment, given a Multinet segment. Table 12.1 presents the criteria and the corresponding points. The threshold for selection was 320 points, which means at least two criteria must be met at full capacity.

Table 12.1: Scoring algorithm tests and points

Test Metric	Test Criteria	Points Awarded
Computed Geometric Data Matching Criteria		
Perpendicular distance between HPMS and Multinet segments	Distance ≤5 m (16 ft) Distance ≤10 m (33 ft) Distance ≤15 m (49 ft)	250 100 10
Parallel distance between HPMS and Multinet segments	Distance ≤20 % of HPMS and Multinet segment lengths Distance ≤30 % of HPMS and Multinet segment lengths Distance ≤50 % of HPMS and Multinet segment lengths	100 50 10
Angle between HPMS and Multinet segments	Angle ≤0.010 rad (0.600°) Angle ≤0.100 rad (5.700°) Angle ≤1.000 rad (57°)	100 50 10
Field Metadata Matching Criteria		
Road shield number of HPMS and Multinet Segments	Are the same	1 000
Difference in FRC between HPMS and Multinet segments	Difference =0 classes Difference ≤1 class Difference ≤2 classes	100 50 10
Number of lanes in HPMS and Multinet segments	Multinet lanes = half of HPMS lanes	300
Toll road indicator in HPMS and Multinet segments	Flagged as toll roads	300
Threshold to select candidate HPMS segment	**Total points should be ≥**	**320**

Note that the points and threshold were fine-tuned by looking at sample joins from a wide range of geographies, including the plains of Kansas, the mountains of Colorado, the coastal roads of Hawaii and California, the density of roads in the Northeast, and the sparsity of roads in the Midwest. The candidate segments scoring above the threshold were sorted in descending order of scores, and ties were broken by secondary and tertiary ascending sorts on the perpendicular and parallel distances. The highest scoring, closest segment was thus selected as the join partner.

Another noteworthy point from table 12.1 is the relative weights given to each test metric and criterion. Perpendicular closeness is weighted very heavily, as are having the same number of lanes and being part of a toll road. Matching names in the metadata of both segments outright receives a huge score to boost it to the top. The points are cumulative, that is, if two segments have a perpendicular distance less than 5 m, then the total points awarded are 360 $(250 + 100 + 10)$, which automatically crosses the threshold. This and the road shield numbers are the only criteria that allow pairing when met in full.

An optimization method was not used for computing the scores or the threshold because of the size of the datasets involved. There were no good candidate areas on which to train the models, as land topology changed drastically over the country. Moreover, the compute times were already quite large, and an optimization algorithm would not have finished in the duration of the project.

12.3.2 Complete Algorithm

The implementation and execution of the following algorithm was performed solely by the author, without any division of labor.

Preprocessing

1. Split the HPMS segments into shorter segments of 160 m (528 ft).

2. Compute the angles with respect to the equator and the centroids of each HPMS segment and store with the HPMS geometries.

3. Construct bounding buffer boxes around the HPMS segments with a 500 m buffer (reduce to 100 m for optimal performance).

4. Build an R-tree index for HPMS bounding boxes.

5. Filter Multinet segments to remove segments with FRC greater than 6 or where the segment is not a roadway, for example, ferry paths.

6. Compute the typical profile speeds by day of week using the Multinet geospatial dataset and profile speed tables. This will be used in step 9.

Conflation

7. For each Multinet segment from the filtered set:

   (a) Devolve the segment into five points at 0 %, 25 %, 50 %, 75 % and 100 % along the segment length.

   (b) For each devolved point:

      i. Compile a list of candidate HPMS segments by looking up the point in the R-tree index from step 3. Go to step 7(h) if no candidate HPMS segments are found.

      ii. Compute the angle and length of an imaginary line connecting the point and centroid of each candidate HPMS segment identified in step 7(b)i.

      iii. Subtract the angle computed in 7(b)ii from the absolute angle of the HPMS segment computed in step 2.

iv. Resolve the length computed in 7(b)ii into components parallel and perpendicular to the HPMS segment using the relative angle computed in 7(b)iii.

(c) Find the average parallel and perpendicular distances from 7(b)iv for all points from 7(a), given a candidate HPMS segment, to obtain the approximate parallel and perpendicular distances between the Multinet and each of the candidate HPMS segment.

(d) Compute the points scored by each join pair produced by the Multinet segment and candidate HPMS segment using table 12.1.

(e) Filter the join pairs for those that score above the threshold in table 12.1. If no pairs qualify, go to step 7(h).

(f) Sort the filtered pairs by decreasing points, increasing parallel and increasing perpendicular distances.

(g) Designate the HPMS segment in the first pair from step 7(f) as the join key.

(h) If any step fails to produce a candidate HPMS segment, attribute $NULL$ as the join key to the Multinet segment.

8. Drop Multinet segment with a NULL join key. Copy the AADT values from HPMS to the Multinet dataset through the join.

AADT Disaggregation

9. For each Multinet segment with an AADT value obtained through a successful join:

(a) Halve the transferred AADT value, for segments part of dual carriageways by referring to the one-way flag in the Multinet metadata.

(b) Factor the AADT from step 9(a), for day of week using Exhibit A-1 from the Mobility Scorecard [105].

(c) Use the typical speed profile for day of week computed in step 6 for steps 9(d) and 9(e) below.

(d) Determine the peak direction (AM, PM, or equal congestion) for each day of the week as detailed in Step 3 of the Mobility Scorecard [105].

(e) Determine the severity of congestion with respect to the off-peak speed profiles using Equation A-1 in the Mobility Scorecard

(f) Using ordinate values for time of day from Exhibits A-2 through A-6 in the Mobility Scorecard, and the adjusted AADT values as computed in step 9(b), and using the peak direction and severity of congestion as determined by steps 9(d) and 9(e) above, find the disaggregated traffic counts for each hour in the week, as detailed in the Mobility Scorecard [105].

(g) Linearly interpolate the typical volumes obtained for each hour in a day from step 9(f) to find the volumes at 15-minute periods in a day. Divide these volumes by 4 to ensure the sum of the volumes for each day remains the same as computed in step 9(b).

10. End Algorithm.

This algorithm was run state by state, in a program that can simultaneously process up to 14 states in parallel on different processors. Texas, the state with the largest number of segments, took the most time to finish–slightly over 5 days and 13 hours, occupying a whole processor. The run time could have been significantly decreased by reducing the buffer around the HPMS segments. The largest penalties to runtime were incurred in computing the angles and lengths of the segments, which required high-precision floating-point geometry calculations.

12.4 Method to Merge Incidents and Congestion Data

This was one of the first methods developed and consequently, is quite basic. The first step was to select only the incidents that impact a bottleneck spatially. Since, the only available information is about the head of the bottleneck, just incidents around the head are selected. The selection buffer was created with a radius of one mile. The radius is computed using Vincenty's great circle formula for ellipsoids [120]. The spatially selected incidents are further filtered by the time of occurrence. Only incidents that meet any of the following criteria are selected. These subselected impacting incidents have been presented in table 11.4.

1. Start within a 30-minute window before the start of the bottlenecks.
2. Start during the span of the bottlenecks. Incidents may last beyond the life of the bottleneck.
3. End during the span of the bottlenecks.
4. Start before and end after the bottlenecks, that is, range of the span of the bottlenecks.

Chapter 13: Results from Performance Measurement Methods

In this chapter the results from the methods developed to compute performance measures are presented. Since each documented method in chapter 12 is different, with different objectives, each section in this chapter presents and discusses the results from the specific method.

13.1 Results from Forecasting Traffic Speeds

The results presented in this section were originally published in the paper by Cirillo et al. [18].

Since visualizing the results for the whole network as shown in figure 11.2 is not possible, aggregate measures are presented. Predictions up to 30-minutes in the future, from any given minute are made, unless insufficient real observations are available. For example, for an $ARIMA(1, 1, q)$ model, a minimum of two observations are required before predictions can be made. Therefore, only after receiving the data for two minutes past midnight, can the data be forecasted. In other words, for segments modeled using $ARIMA(1, 1, q)$, minute three and onwards can be predicted.

13.1.1 Model Selection

The first step in applying the proposed framework is model selection. As mentioned, (see equation (12.2)) the models are selected using data from the same day a week

before the day to be predicted. $ARIMA(p, d, q)$ models are fitted to the data, where $p, d, q \in \{0, 1, 2\}$, and the best model out of the possible 27 are selected. Best model is determined as the model with the lowest Bayesian Information Criterion (BIC).

Table 13.1 gives an overview of the selected ARIMA orders. A model is fit to data from each day for each segment. Since models are fit to a total of 10 weekdays (two weeks) in this study, total number of orders are 26 540. The three most popular orders are highlighted.

Table 13.1: Order Selection

Order	Count	Order	Count	Order	Count
(0, 0, 0)	NA	(1, 0, 0)	5700	(2, 0, 0)	817
(0, 0, 1)	0	(1, 0, 1)	185	(2, 0, 1)	239
(0, 0, 2)	0	(1, 0, 2)	886	(2, 0, 2)	381
(0, 1, 0)	3077	(1, 1, 0)	589	(2, 1, 0)	154
(0, 1, 1)	405	(1, 1, 1)	10995	(2, 1, 1)	1 254
(0, 1, 2)	196	(1, 1, 2)	567	(2, 1, 2)	1 095
(0, 2, 0)	0	(1, 2, 0)	0	(2, 2, 0)	0
(0, 2, 1)	0	(1, 2, 1)	0	(2, 2, 1)	0
(0, 2, 2)	0	(1, 2, 2)	0	(2, 2, 2)	0

13.1.2 Forecasting

Forecasts are made continuously once the minimum number of observations are available. For example, the first two minutes of real data are used to forecast the speed for the third minute in an $ARIMA(1, 1, q)$ model. The forecast for the fourth minute then depends on the predicted value for the third minute, and the observed value of the second minute. Fifth minute is predicted using the forecasts for the third and fourth minutes, and so on. Forecasts are stopped at 11:59PM of each day.

For every minute of the day, up to 30-minutes of future estimates of speed are produced, or, in other words, for each minute, predictions from data received up to 30-minutes ago are available. To avoid confusion, the latter terminology is used, and

the term lag is defined as minutes prior to current minute that were used to predict the speeds at current minute. In other words, a lag of 5 means speed at the current minute was predicted from data received 5-minutes ago. Using these forecasts, statistical properties like the relative error are computed. Relative prediction errors from 5, 10, 15 and 30 minutes of lag are selected for presentation to avoid cluttering the figures. Where feasible in the plots, results from 20 and 25 minutes of lag are also included.

Since each segment can have a different speed profile, and maximum speed (the speed limit of the roadway the segment is on) a single measure for all segments needs to be relative. Therefore, the Relative Root Mean Squared Prediction Error (RRMSPE) is introduced. The RRMSPE is calculated similar to the Root Mean Squared Error (RMSE), but with the additional step that the squared difference between the true speed and predicted speed is divided by the square of true speed for each observation. Mathematically, RRMSPE is given by

$$RRMSPE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2}, \tag{13.1}$$

where

$\hat{y}_t$ is the predicted speed at time t,

y_t is the real observed speed at time t,

T is the total number of time steps. For a day, $T = 1\,440$, or $T = 60$ for an hour.

Note that a single value of RRMSPE is produced for a given lag, segment and day. Therefore, a total of 26 540 RRMSPE values are produced for each lag. RRMSPE is calculated for each minute the data is predicted, however, since only predictions from 5, 10, 15 and 30 minutes are presented, RRMSPE from only these minutes are shown. In all figures, each lag is kept distinct from others, so that errors can be clearly analyzed.

Figure 13.1 shows a box plot for each lag. Abscissa in figure 13.1 shows the lag value, while the ordinate is the RRMSPE measure. The box encompasses the Inter Quartile Range (IQR) for RRMSPE values from all segments for both weeks of predictions, (that is, IQR of 26 540 points). The orange line within a box represents the median RRMSPE value for the given lag. The whiskers of the boxes show the 5^{th} and 95^{th} quantiles. The decision to move the whiskers out to the extreme quantiles was made to demonstrate the accuracy of ARIMA models in the synthetic framework. If figure 13.1 were instead also split by the day of the week, and the RRMSPE values averaged for each lag (that is 2 654 points averaged together), figure 13.2 is produced.

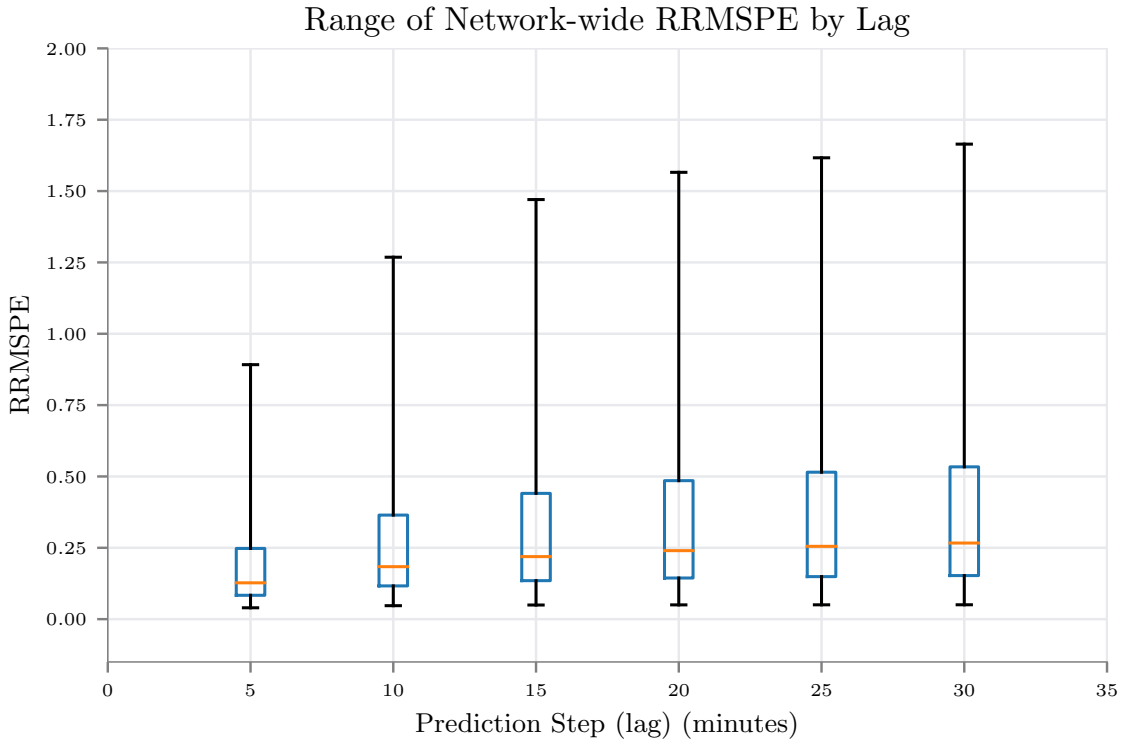


Figure 13.1: Box Plot of RRMSPE Using all Predictions by Lag

Figure 13.2 can be transposed such that the day falls on the abscissa, while a line for each lag is drawn on the plot against RRMSPE value on the ordinate. Figure 13.3 shows this transformation. It is of academic interest that Tuesday and Thursday are two of the most difficult days to predict, although, traffic patterns on those days are

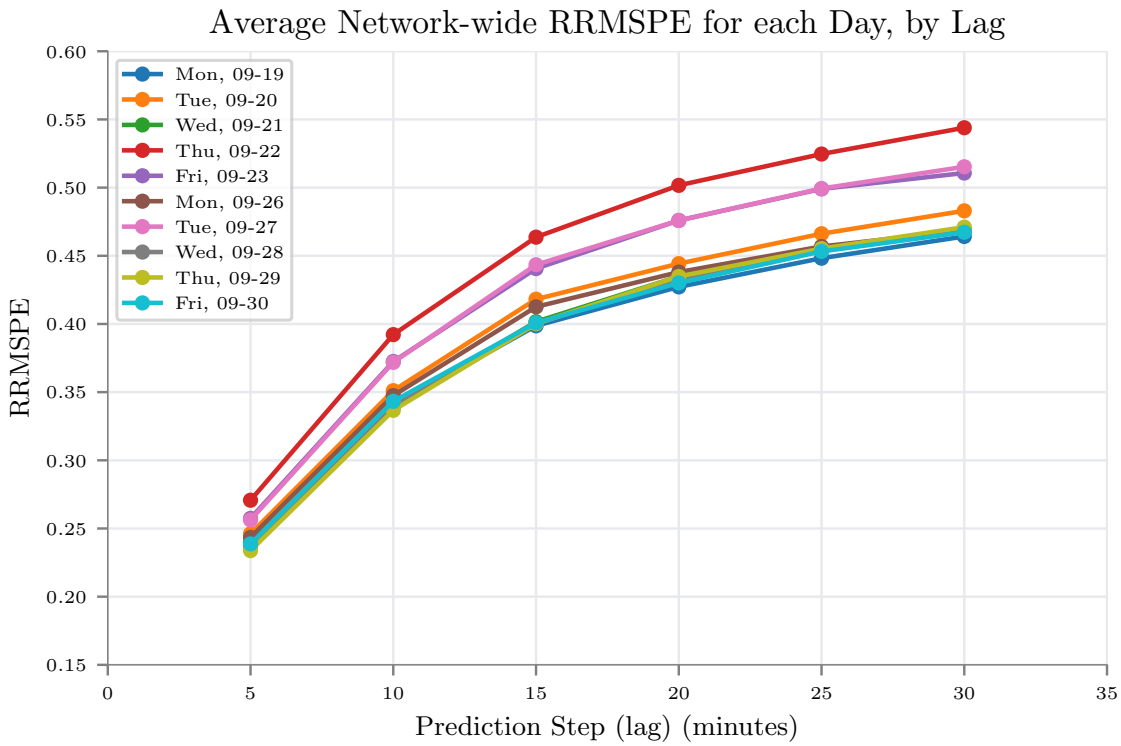


Figure 13.2: RRMSPE for all Segments, each Day by Lag

generally considered very repeatable. In fact, documents like the Maryland Mobility Report compute results using data from Tuesdays, Wednesdays and Thursdays because the traffic patterns are more stable as compared to other days [76]. Conversely, Fridays — with very variable patterns due to weekend demand — seems easier to predict.

In order to understand figure 13.3 better the day is broken down into periods as shown in figure 13.4. Periods are fixed in size, and are defined in 4-hour intervals — therefore 6 periods in a day. The list of periods is: midnight to 4AM; 4AM to 8AM; 8AM to noon (morning peak); noon to 4PM; 4PM to 8PM (evening peak); and 8PM to midnight. Average RRMSPE for a lag is plotted using data from all segments in a given period in a day in figure 13.4. For figures with distinct periods, RRMSPE of a period is computed from 240 data points for each minute in the 4-hour period. Similar to 13.3, figure 13.4 plots the day and period on the abscissa, RRMSPE value on the ordinate, and divides the plot to demarcate each day.

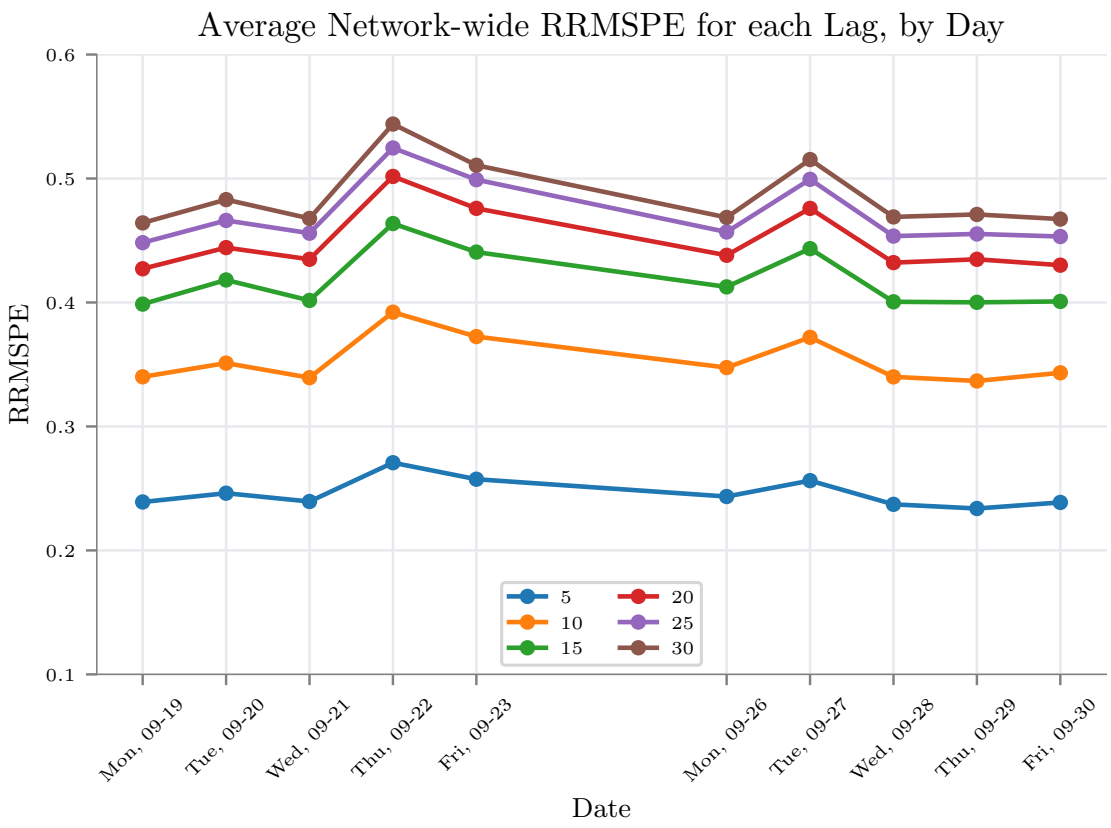


Figure 13.3: RRMSPE for all Segments, for each Lag by Day

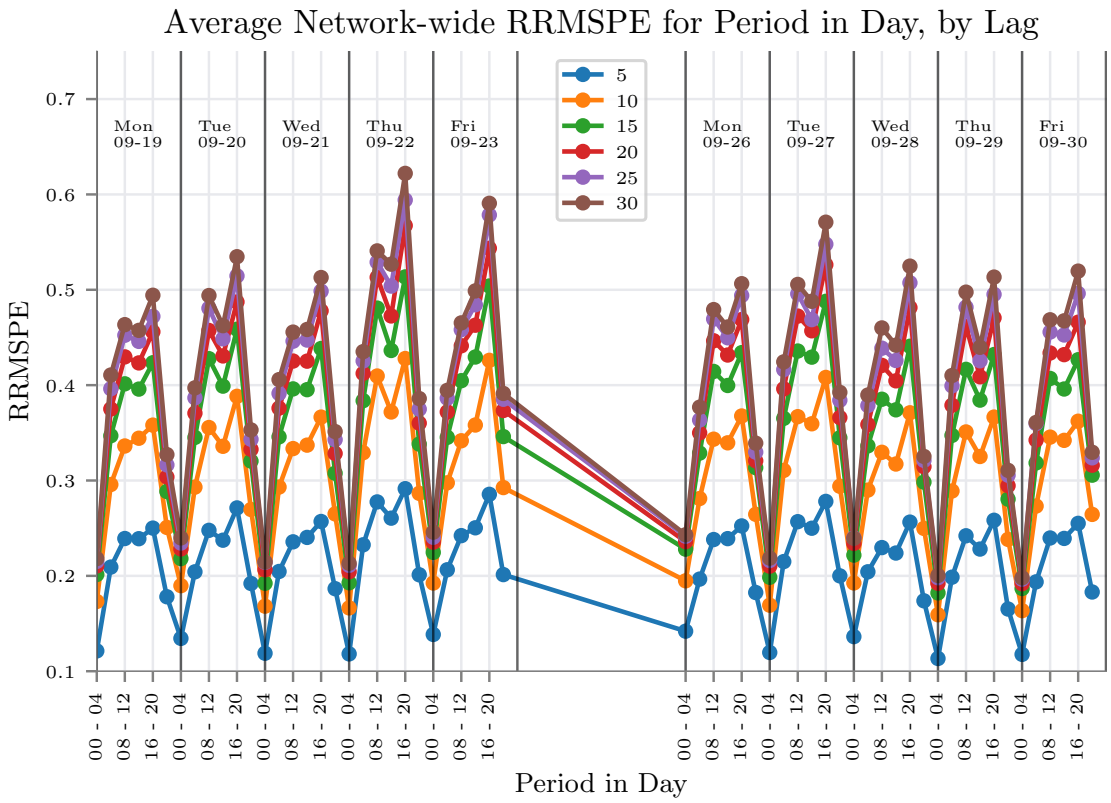


Figure 13.4: RRMSPE for all Segments, for each Lag by Period in Day

Figure 13.4 does not show the spread of the RRMSPE values, just the average value. Figure 13.5 however, expands each point in figure 13.4 into a box, showing the ranges of the data. Note that the whiskers in all box plots (figures 13.1, 13.5 and 13.6) correspond to the 5^{th} and 95^{th} quantiles, while the blue box itself shows the IQR, and the orange line within a box shows the median. Each subplot in 13.5 shows data from one lag (indicated in the subplot title).

Lastly, in order to show the full variation in the data, figure 13.6 is shown. This figure plots a box plot for each minute of the day. Relative residuals for all segments, over all 10 days for the given minute comprise each box (26 540 data points per minute). Relative residuals are given by equation (13.2). Figure 13.6 clearly shows how the predicted values deviated from the real values, for each minute of the day. Similar to all other box plots, the median is shown by the orange line, the blue boxes, barely even at peak periods shows the IQR, while the whiskers shows the extreme 5 % quantile ranges. The time of day is plotted on the abscissa, while the ordinate gives the relative residual value.

$$Rr_t = \frac{\hat{y}_t - y_t}{y_t}, \tag{13.2}$$

where

Rr_t is the relative residuals at time t,

$\hat{y}_t$ is the predicted speed at time t,

y_t is the real observed speed at time t.

The increase in deviation is clearly visible around peak periods in figure 13.6. Further, it is observed, from the larger relative residuals on the positive side, that the models make more optimistic predictions over the whole network. One of the most important observations is that the median line remains consistently near zero, indicating that the majority of predictions are very accurate. Also, note the extent of IQR denoted by the small blue region around the median line, again holding a

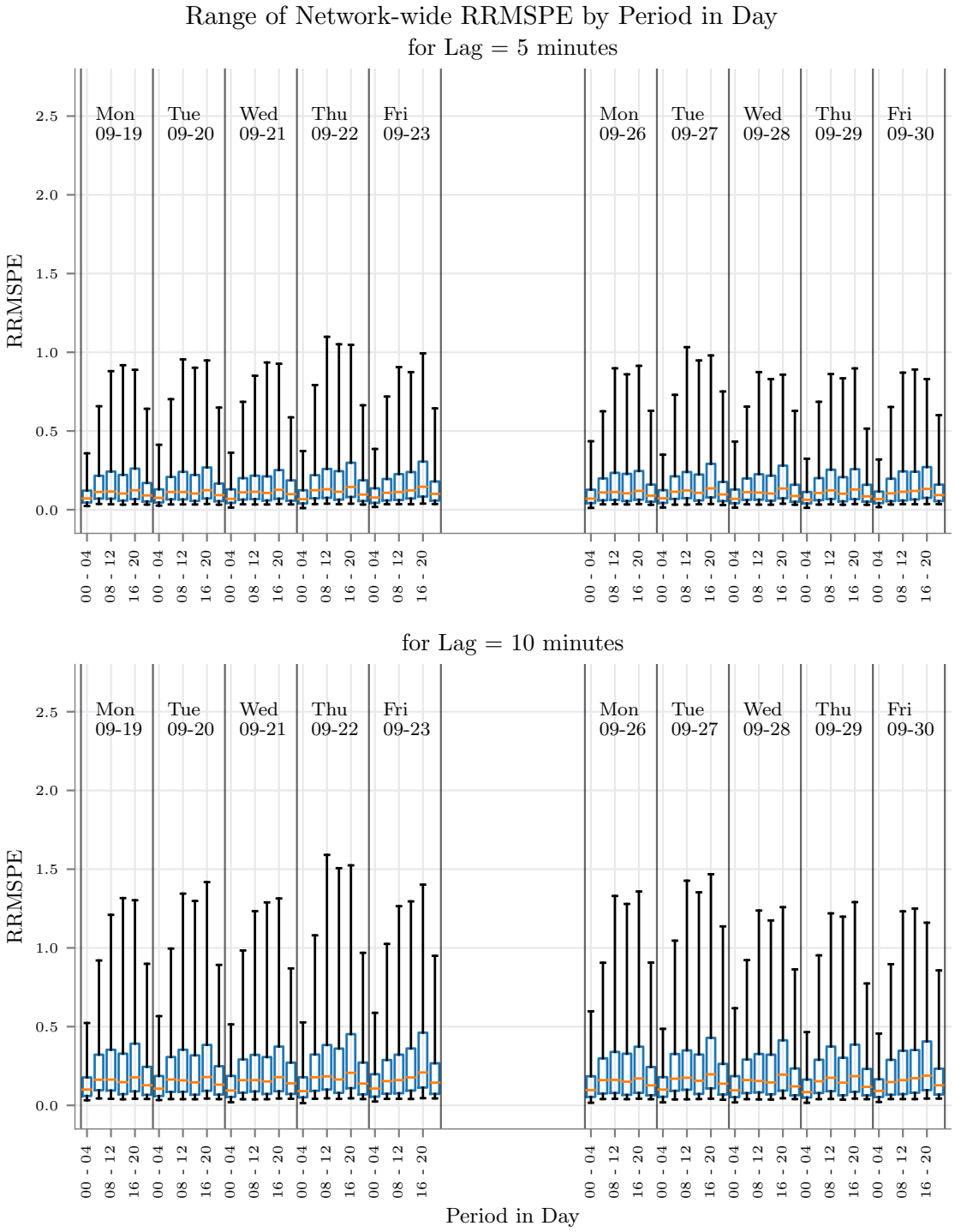


Figure 13.5: Box Plot of RRMSPE for all Segments, for each Lag by Period in Day

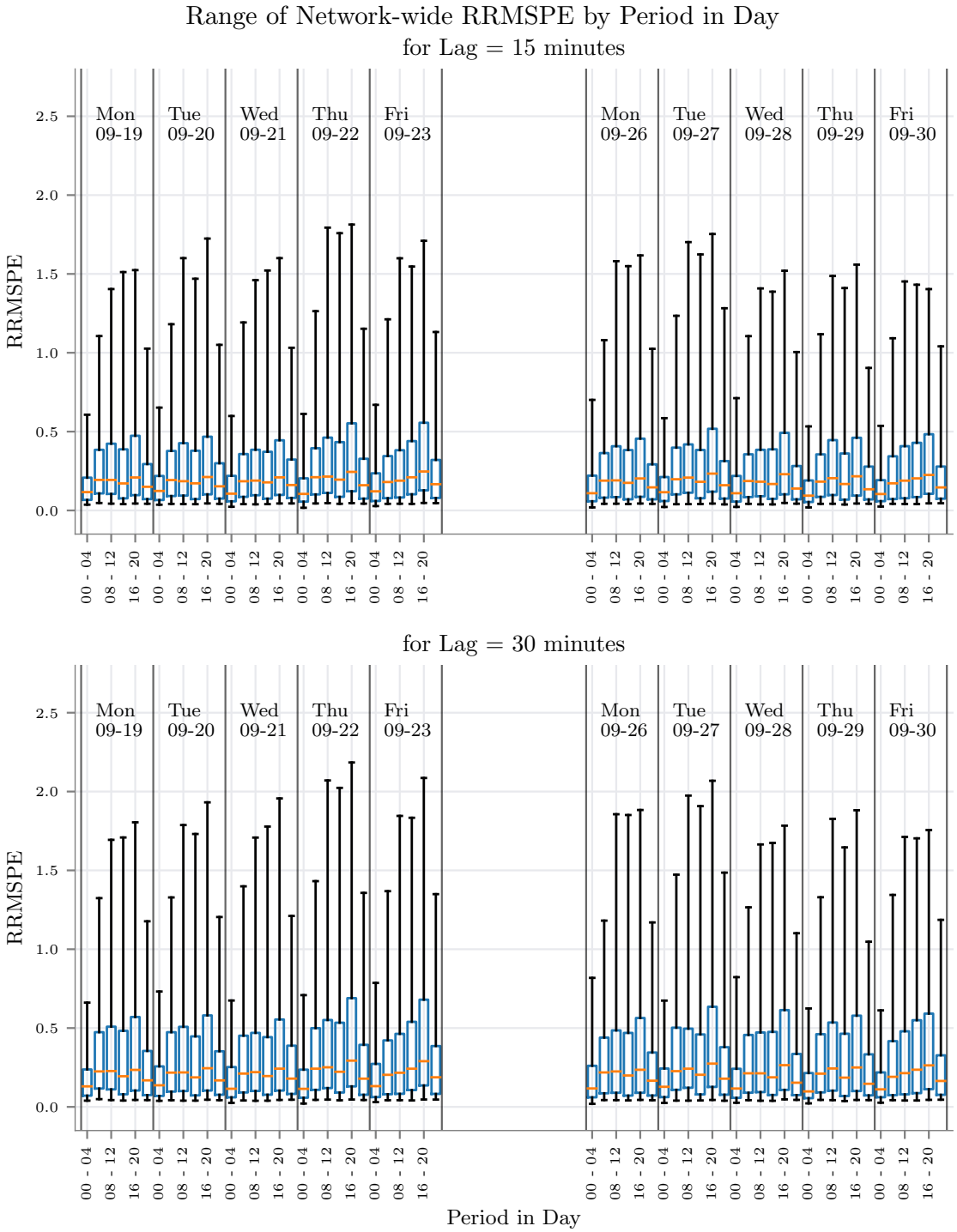


Figure 13.5: Box Plot of RRMSPE for all Segments, for each Lag by Period in Day

testament to the accuracy accorded by the synthetic method to ARIMA models.

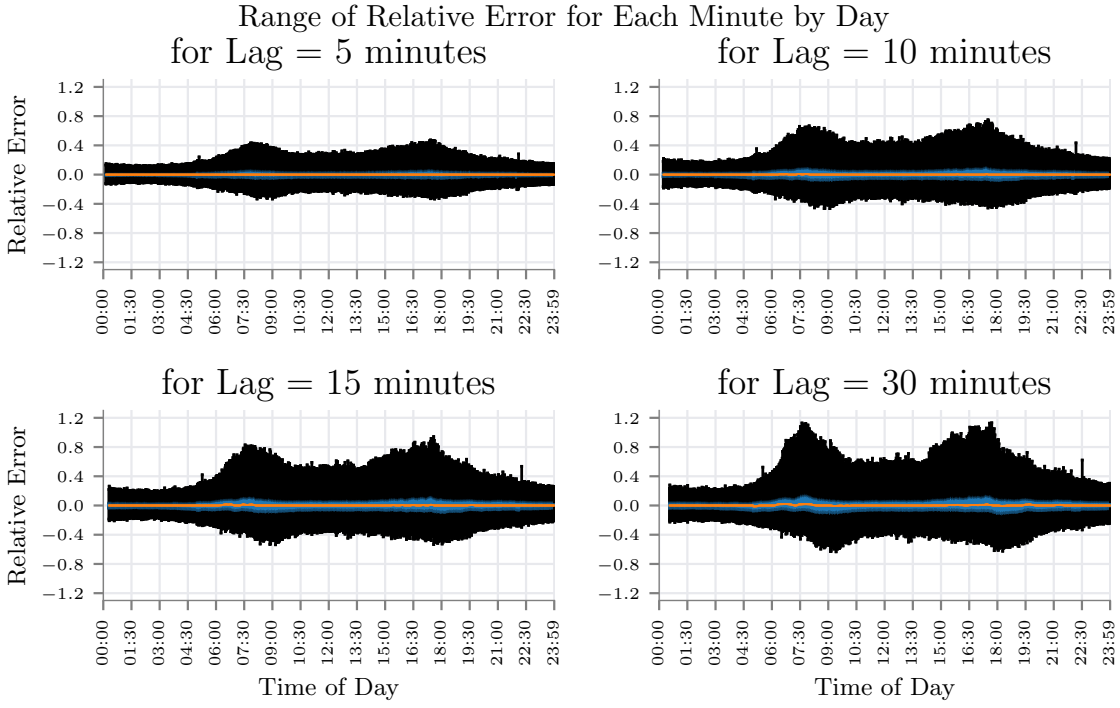


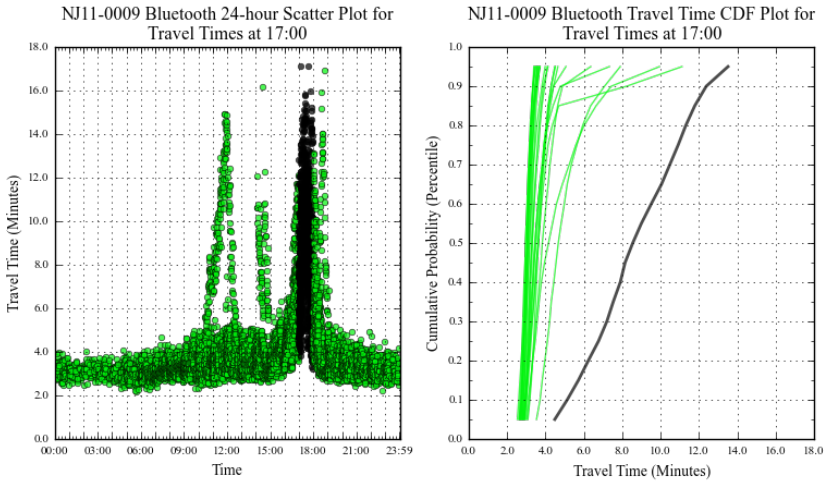
Figure 13.6: Relative Deviation for all Segments, for each Lag by Minute in Day

13.2 Results from Performance Measurement

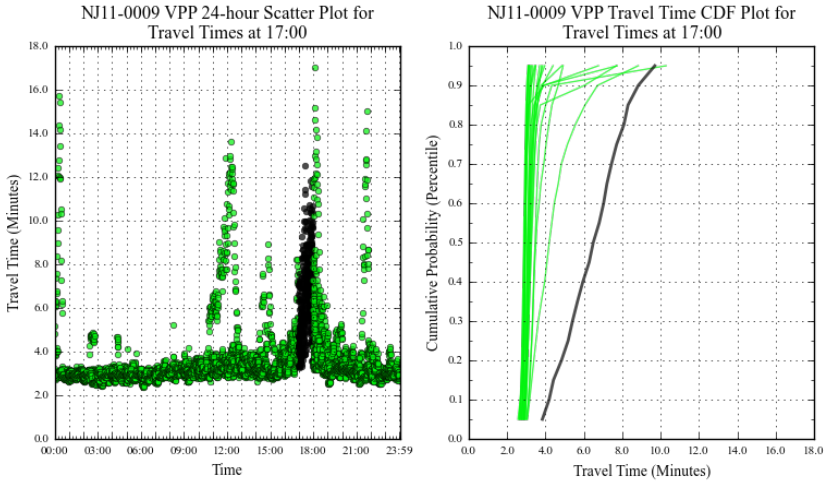
The results presented in this section were originally published in the paper by Kaushik, Sharifi, and Young [61].

The NPMRDS, VPP and Bluetooth data were overlaid for the two case study segments as explained (see 12). The results of overlaying the data, and plotting the cfds are presented in figures 13.7 and 13.8 for segments NJ11-0009 and VA08-0012, respectively. The peak hour is highlighted in all subplots of figures 13.7 and 13.8 to allow direct comparison of the different datasets. Both case study segments have a PM peak hour, with travel times rising to almost four times the freeflow travel times. The peak hour is colored black in the overlay scatter plot and the cfd plot in each subfigure in figures 13.7 and 13.8. In each figure, part a shows the Bluetooth data,

part b shows the VPP data, part c shows NPMRDS all vehicle data, part d shows NPMRDS passenger car data, while part e shows NPMRDS freight truck data.



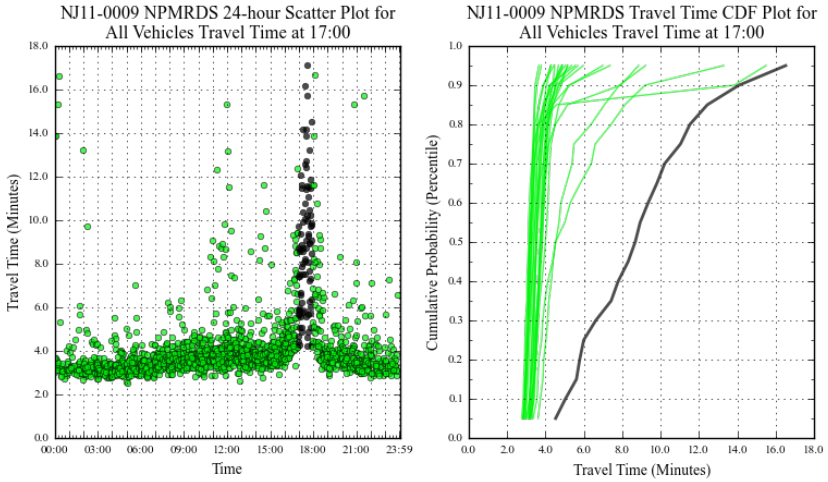
(a) Bluetooth travel times on NJ11-0009



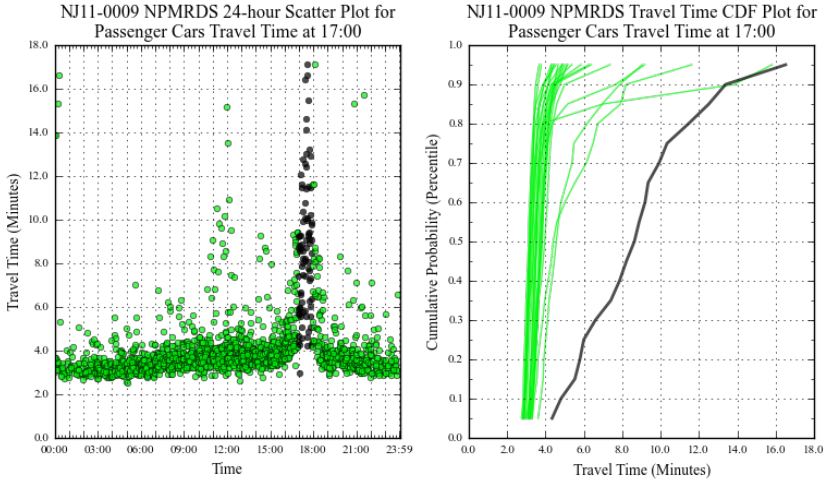
(b) VPP travel times on NJ11-0009

Figure 13.7: Overlaid travel times, and associated cfd plots

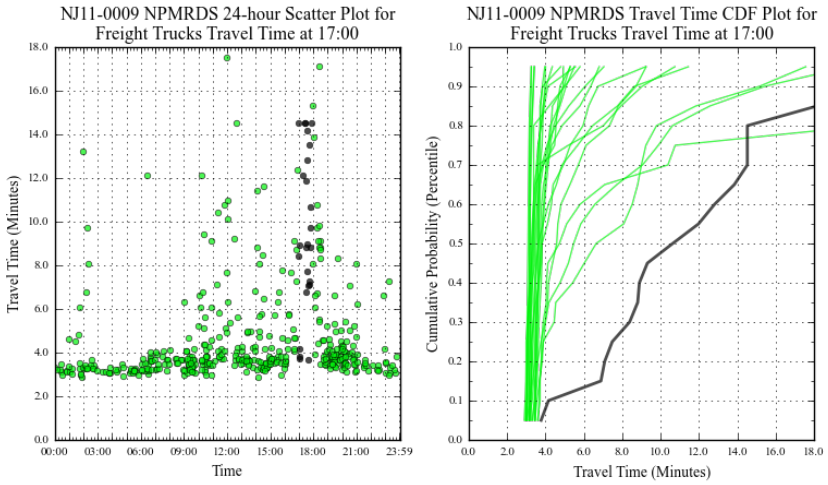
It is clearly seen from figure 13.7 that the three datasets produce almost similar results for the freeway-like NJ11-0009. This is expected in the case of freeways, where filtered probe data works very well. However, as shown by figure 13.8, in the case of the arterial segment, VA08-0012, VPP data, which is filtered and smoothed at source, performs poorly as compared to the NPMRDS or the Bluetooth data. These plots visually help compare the performance of the three datasets. The actual numeric



(c) NPMRDS all vehicle travel times on NJ11-0009 for NJ11-0009

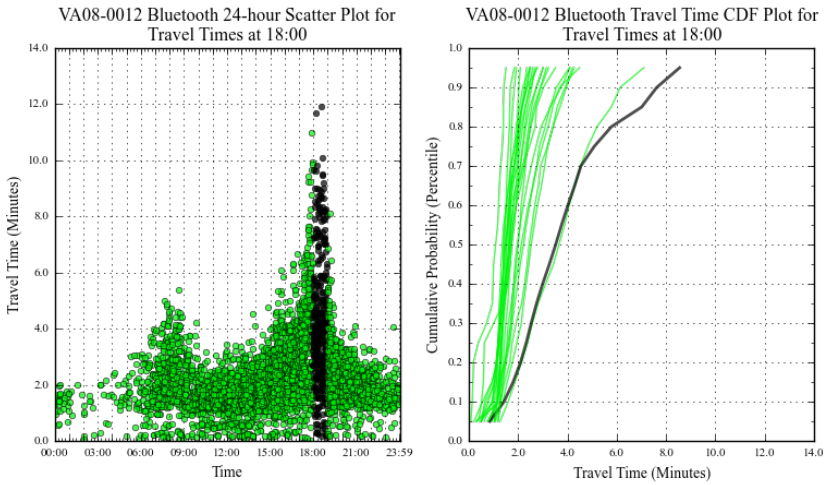


(d) NPMRDS passenger car travel times on NJ11-0009

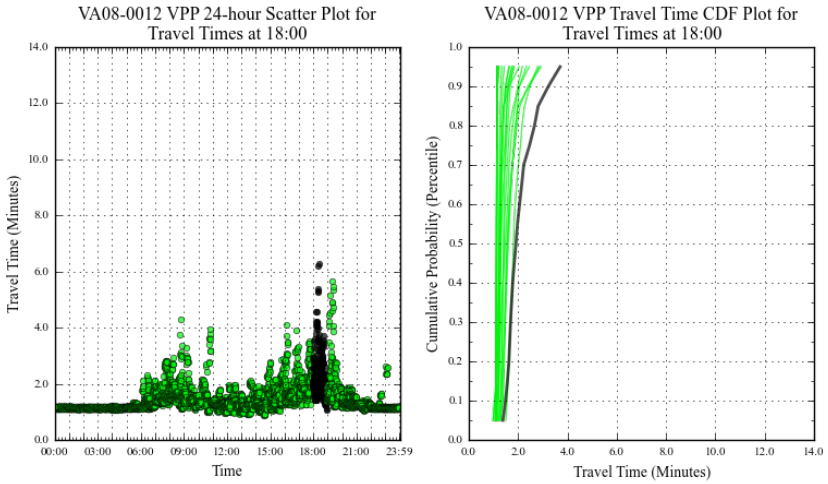


(e) NPMRDS freight truck travel times on NJ11-0009

Figure 13.7: Overlaid travel times, and associated cfd plots for NJ11-0009

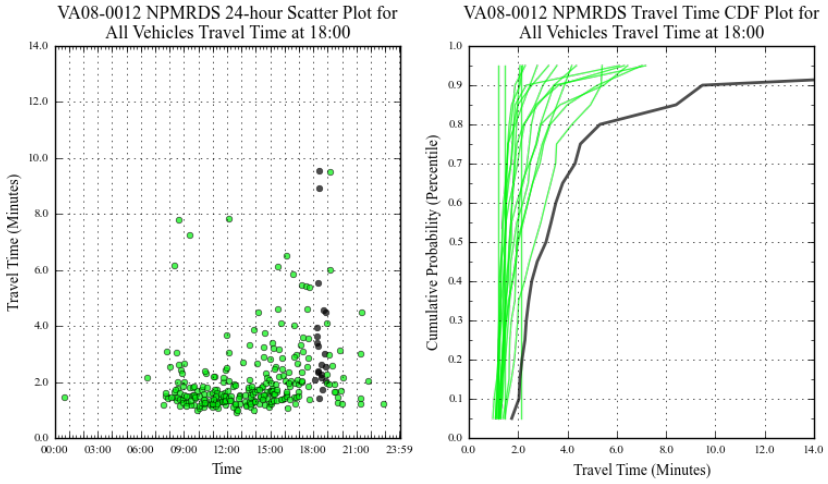


(a) Bluetooth travel times on VA08-0012

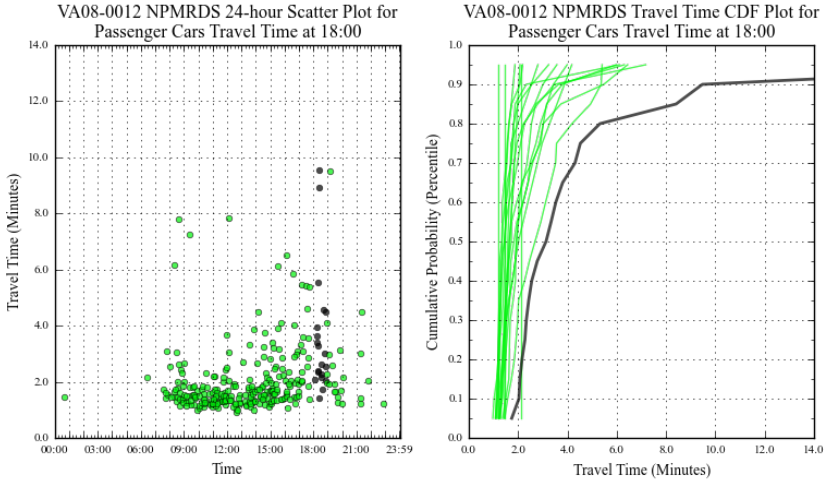


(b) VPP travel times on VA08-0012

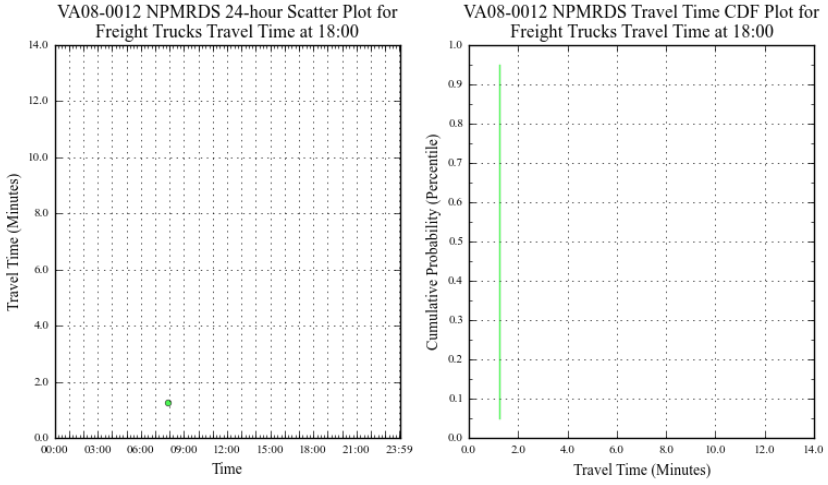
Figure 13.8: Overlaid travel times, and associated cfd plots for VA08-0012



(c) NPMRDS all vehicle travel times on VA08-0012



(d) NPMRDS passenger car travel times on VA08-0012



(e) NPMRDS freight truck travel times on VA08-0012

Figure 13.8: Overlaid travel times, and associated cfd plots for VA08-0012

values can be easily computed using equations (12.5) and (12.6). Due to the simplicity of the calculations, and in the interest of brevity, the numbers are not shown in this chapter.

Additionally, it should be noted that VA08-0012 is composed of two or more TMC segments, depending on the data source. In this case, a shortcut method was used to get the combined travel times on both segments. The method involves finding those times when data was available for all comprising TMC segments simultaneously, and then adding them together to produce the total travel time to traverse the study segment. The correct method to compute this is to trace each observation on the upstream segment, and look for the corresponding travel time at the downstream segment when a hypothetical vehicle is expected to arrive at the downstream segment. However, this method is quite difficult to execute on the NPMRDS data due to the gaps in the data, and therefore was not used. Since travel times from VPP and NPMRDS for all constituent TMC segments must be simultaneously available in the case of VA08, the data density of both NPMRDS and VPP was severely reduced. A way to overcome this is to overlay even more data. Alternatively, the 10-day overlay can be used to construct and impute missing data before tracing a hypothetical vehicle to arrive at the travel time to cross the section of roadway under study.

13.3 Results from Spatial Conflation of HPMS and Multinet Datasets

The results presented in this section were originally published in the paper by Kaushik, Wood, and Gonder [63].

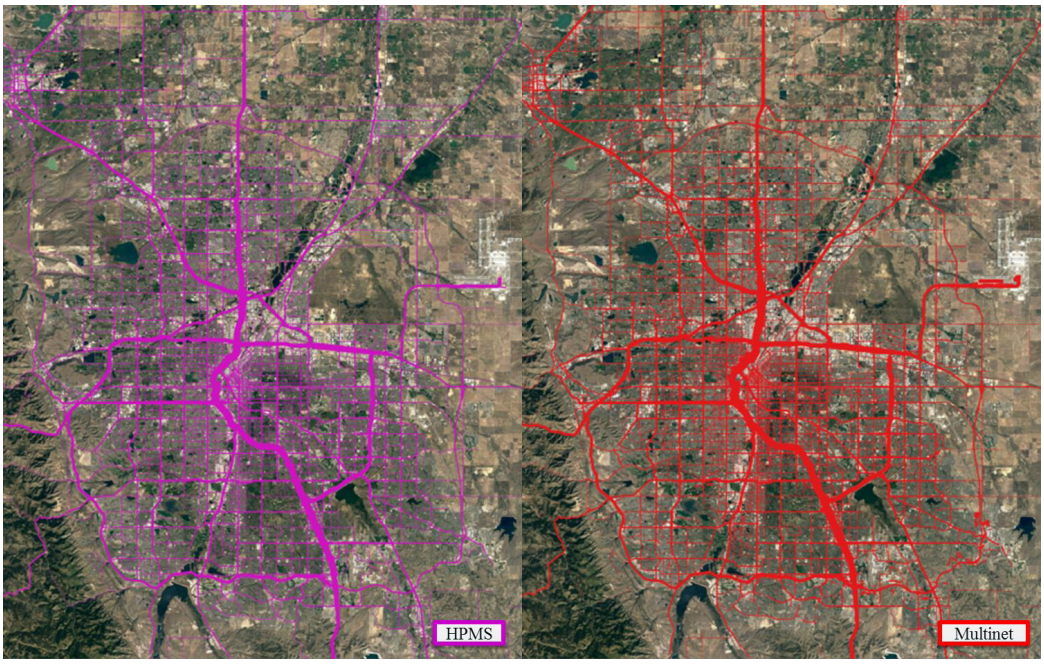
Figures 13.9a and 13.9b show the same geographical areas as figures 11.4a and 11.4b. The line widths have been scaled in proportion to the AADT of each segment. The similarity between the HPMS and Multinet maps scaled by the transferred AADT is clearly visible over a metropolitan region in figure 13.9a, and spotlights the prowess of the presented conflation algorithm. Figure 13.9b, when compared with figure

11.4b, shows that the segments in the parking lot southwest of the intersection have disappeared, along with other segments that did not have a join partner. Further, the line widths of Denver West Parkway, north of the intersection show that it has correctly been conflated with the HPMS segment and was not attributed to I-70. Additionally, the ramps of the Interstate have also been correctly attributed.

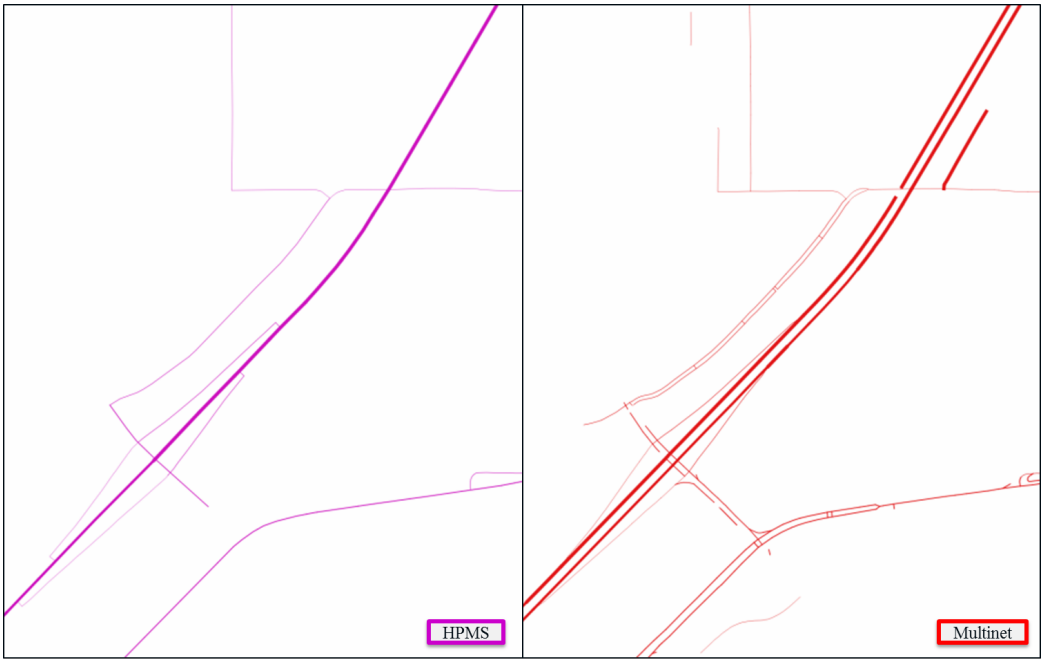
The imperfections become apparent, however, when other segments, parallel but unrelated to the HPMS segments, are thickened by AADT, like the segment in the northeast corner of the map in figure 13.9b, which was attributed the AADT from I-70. Similarly, the segments to the west of the incorrect segment also show "bleeding" of AADT to unrelated segments. These are primarily a product of having a single threshold. The threshold was fixed by considering winding roads, and consequently is far too lenient for straight roads. In addition, some Multinet segments are missing around the intersection, indicating the join failed for these segments.

Table 13.2 presents a global perspective of the errors. The table shows the percent error in the total length of conflated Multinet segments when compared with the total length of HPMS segments for each state. The final value for the whole country shows the average error is just under 7 %, which for 1.200 million miles of complex, real-world roadways is very good by standards in literature.

Figure 13.10 shows the use of the conflated dataset. This figure shows the relation of the speed and volume datasets for I-405N in Los Angeles, California, and I-270N in Maryland. Figures 13.10a and 13.10c are created by plotting the typical week speed profiles by day, and thickening the lines by the traffic volume for each 15-minute period of that day. Figures 13.10b and 13.10d use arrows to present the evolution traffic congestion with respect to time, volume, and speeds. These figures also share a marked resemblance with speed-flow fundamental diagrams.

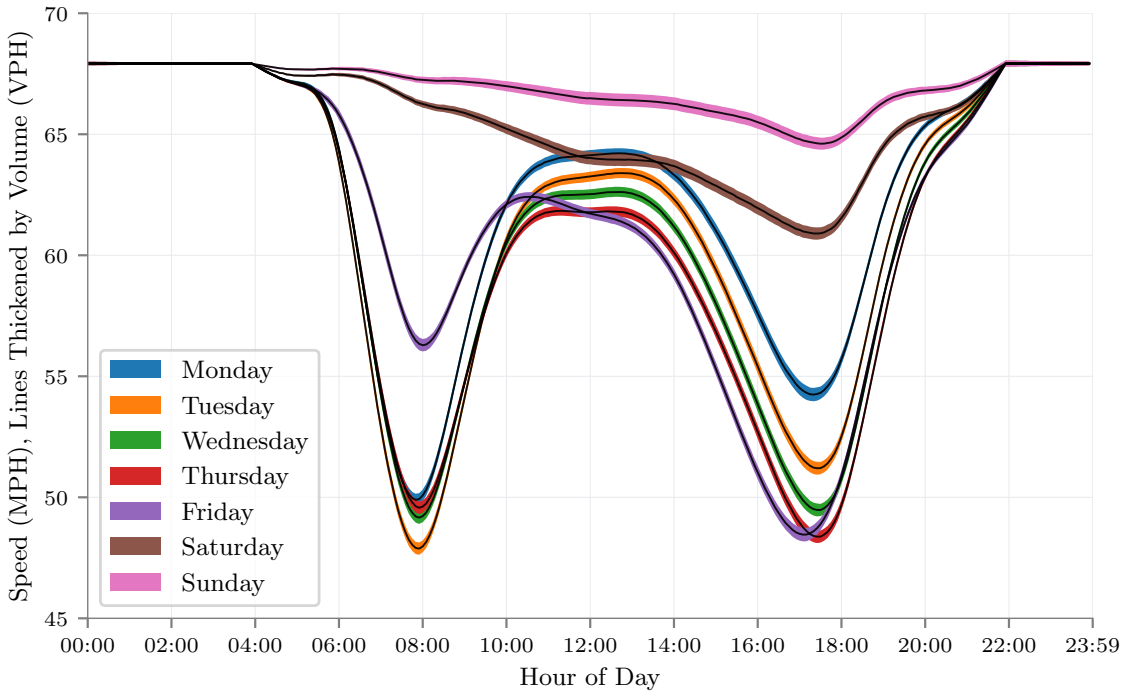


(a) Birds-eye view of Denver, Colorado, post conflation. Lines thickened in proportion to AADT values.

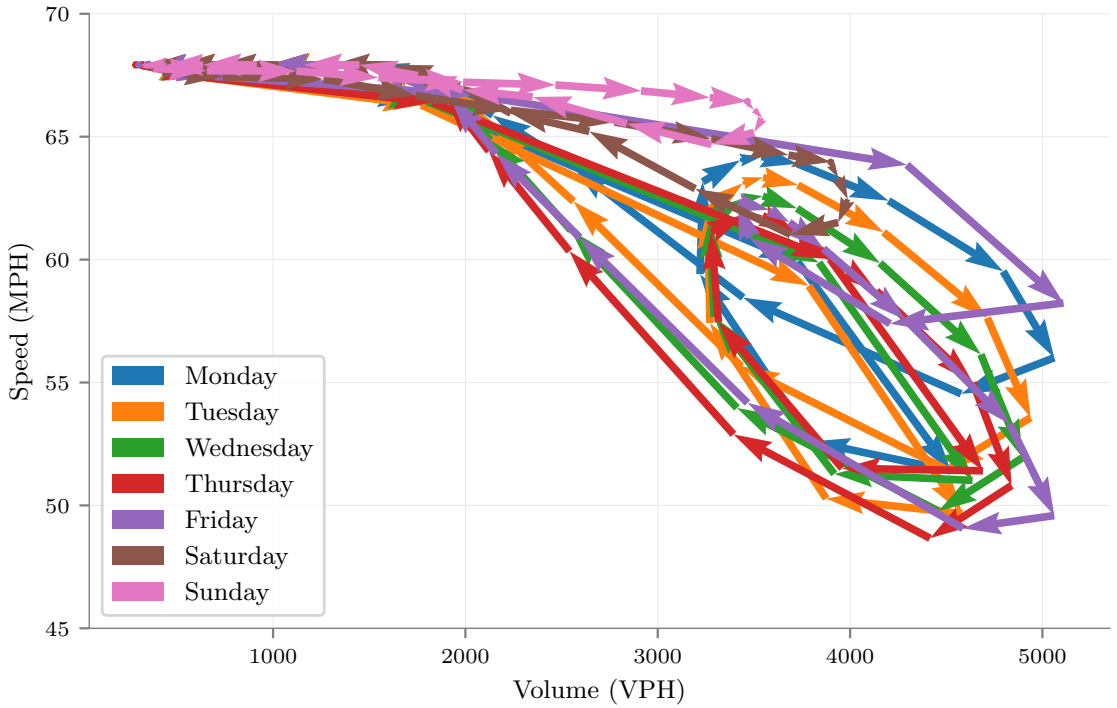


(b) Zoomed-in birds-eye view of an interchange in Denver, Colorado. Lines thickened to show AADT values.

Figure 13.9: Comparison of Network after Conflation

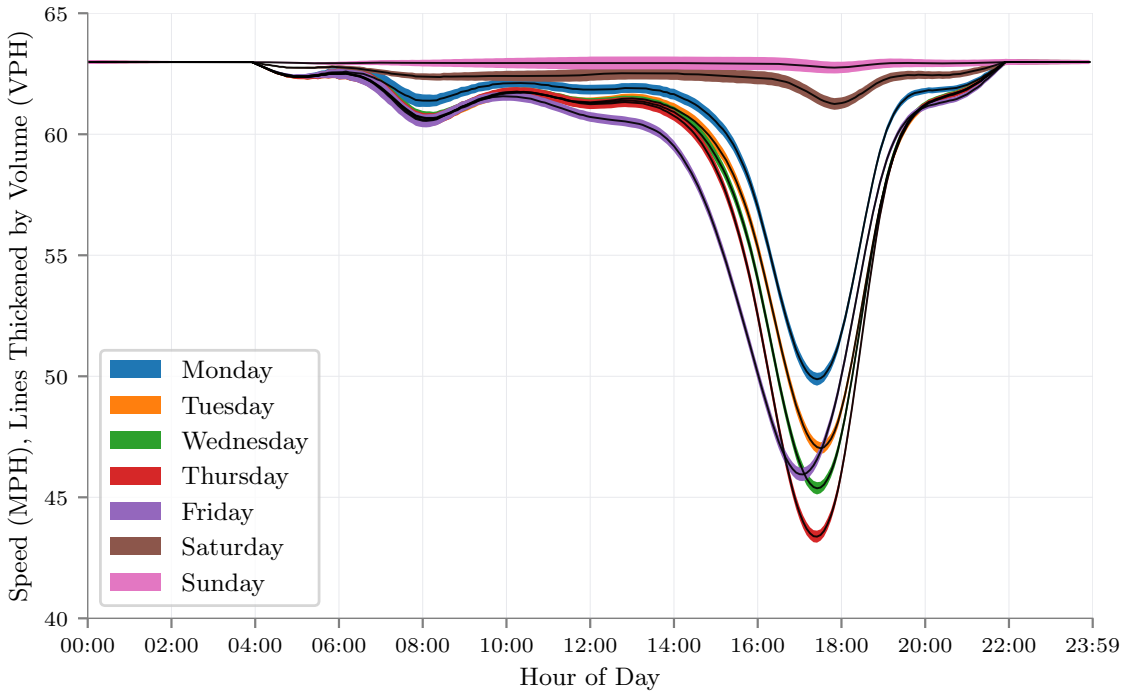


(a) Speed and volume on I-405N in California for a typical week.

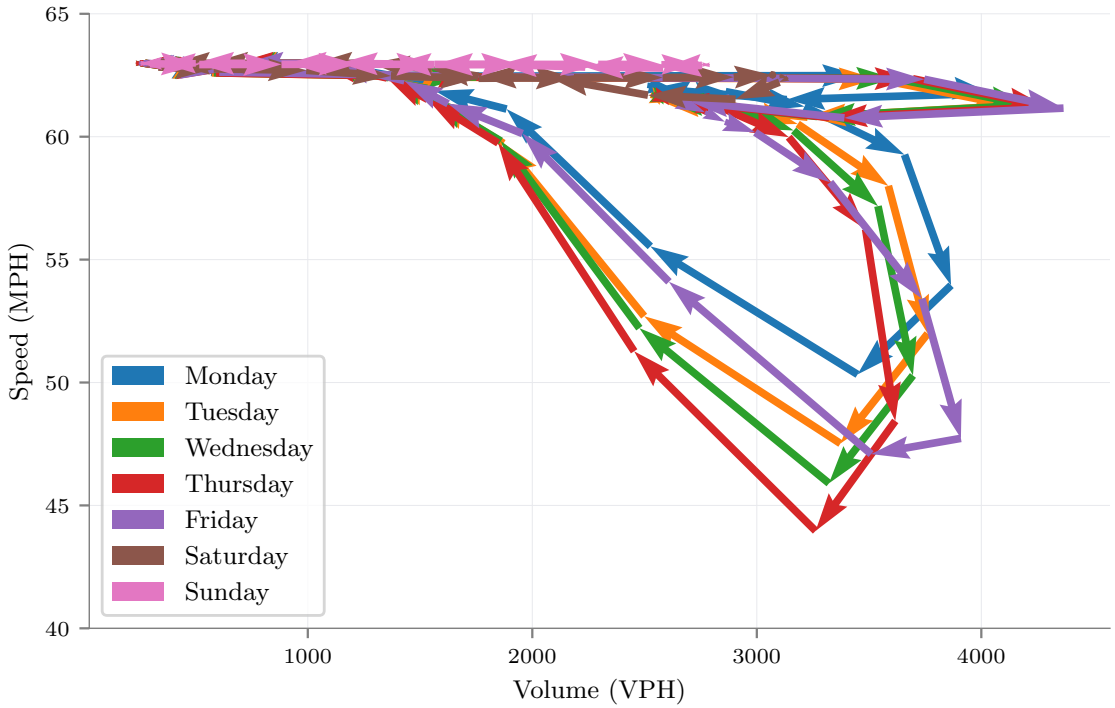


(b) Evolution of congestion during typical week on I-405N in California.

Figure 13.10: Jointly visualizing conflated speed and volume data for two corridors.



(c) Speed and volume on I-270N in Maryland for a typical week.



(d) Evolution of congestion during typical week on I-270N in Maryland.

Figure 13.10: Jointly visualizing conflated speed and volume data for two corridors.

Table 13.2: Total absolute conflation error by state and for the whole country

State FIPS	Total Error	State FIPS	Total Error	State FIPS	Total Error
1	0.044	21	0.080	38	0.033
2	−0.254	22	0.307	39	0.147
4	0.207	23	0.014	40	0.171
5	0.055	24	0.178	41	−0.027
6	0.282	25	0.103	42	0.057
8	0.141	26	0.123	44	0.183
9	0.046	27	0.124	45	−0.310
10	0.004	28	0.075	46	0.083
11	−0.109	29	0.153	47	0.108
12	0.475	30	−0.475	48	0.278
13	0.089	31	0.020	49	−0.068
15	0.156	32	0.051	50	−0.064
16	−0.008	33	0.024	51	−0.291
17	0.230	34	0.347	53	−0.031
18	0.139	35	0.263	54	−0.439
19	0.163	36	0.200	55	0.130
20	0.056	37	0.087	56	0.129
Country Average Absolute Error				0.068	

13.4 Results from Merging Incident and Congestion Data

The results of the incident selection are presented in figure 13.11. In the figure, it appears that every bottleneck has impacting incidents, and the incidents are quite dense. However, since each incident is attributed to a bottleneck, it is a simple matter to count the number of incidents that impact each bottleneck. A chart produced by plotting the count of number of impacting incidents on the ordinate, and the bottleneck impact factor, computed using equation (11.1), on the abscissa for each bottleneck in 2011 is presented in figure 13.12. Different marker types and colors are used to denote the four quarters of the year. The figure clearly shows a decreasing exponential trend: the bottlenecks with a high impact factor are usually impacted by less incidents, while the bottlenecks impacted by a large number of incidents usually have a smaller impact factor. This makes intuitive sense because bottlenecks caused by

incidents are infrequent, and possibly last for shorter durations due to the combined effect of less demand with respect to capacity, and TIM efforts.

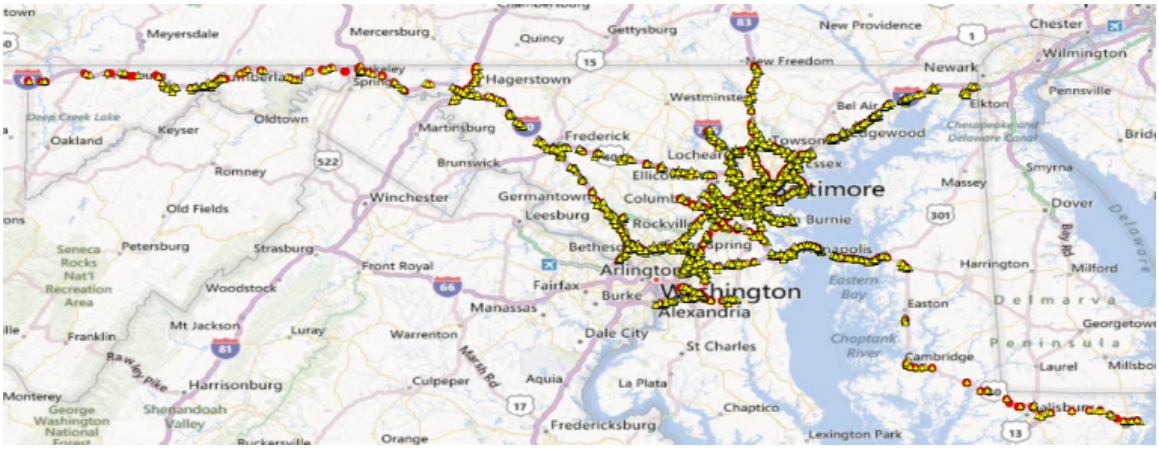


Figure 13.11: Map of all bottlenecks in Maryland from 2011 overlain by incidents also from 2011

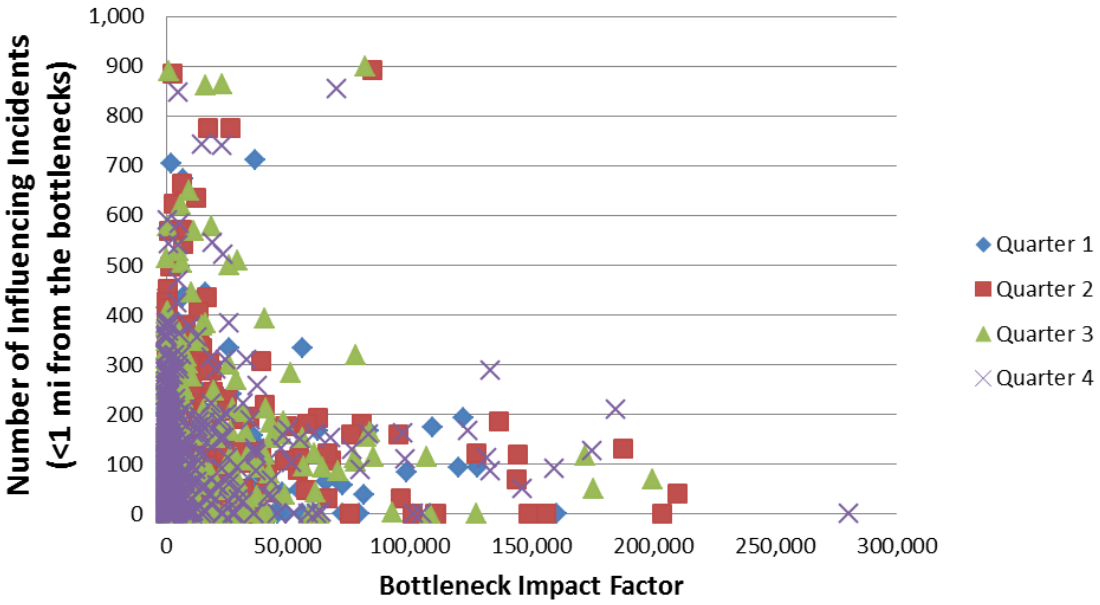


Figure 13.12: Scatter plot of count of incidents and Bottleneck Impact Factor

The interesting results are obtained when figure 13.12 is divided into four quadrants as shown in figure 13.13. The division was guided by Maryland State Highway Administration (SHA), and clearly shows the detailed relation between bottlenecks and incidents. Points in the first quadrant represent bottlenecks with a high impact

factor and incidents, a potential place for heavy motorist misery. The second quadrant comprises of those areas with a large number of incidents, but moderate bottlenecks. This quadrant captures the locations where incidents are mainly responsible for the bottlenecks. Reasonable performance is depicted in the third quadrant, where both bottleneck impact factor and number of incidents are low. Congested conditions with very few incidents are in quadrant four. In order to explore the relation between bottlenecks and incidents, a bottleneck each from quadrants two and four are examined.

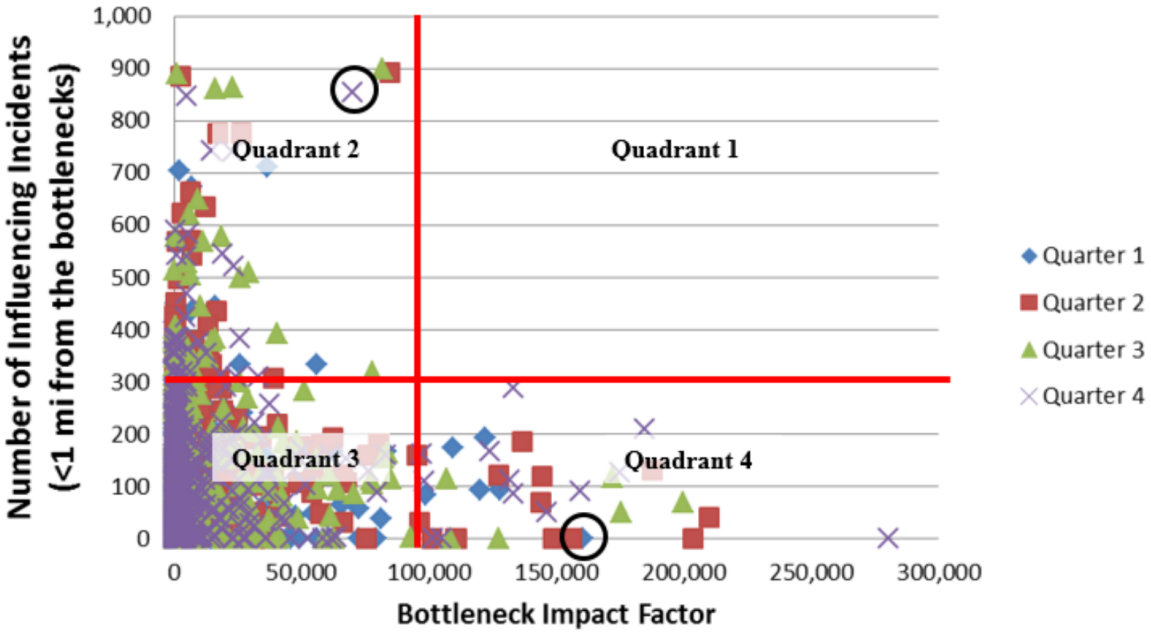


Figure 13.13: Scatter plot of count of incidents and Bottleneck Impact Factor divided into four quadrants

13.4.1 4th Quadrant Bottleneck at I-695 and MD-147

Figure 13.14 shows the location of the bottleneck, with the impacting incidents. The red octagon shows the head of the bottleneck, the purple line shows the direction of the tail, while the yellow triangles represent locations of the incidents. This bottleneck is from the fourth quadrant, and had the worst bottleneck impact factor in the first quarter of 2011. The bottleneck ranking algorithm counted 62 occurrences for this bottleneck in 3 months with an average maximum duration of two and a half hours,

and an average maximum length of 9 miles. The impact factor computed by the bottleneck ranking algorithm is 127 703 mile-hours. However, it was impacted by only 16 incidents during the same time frame.

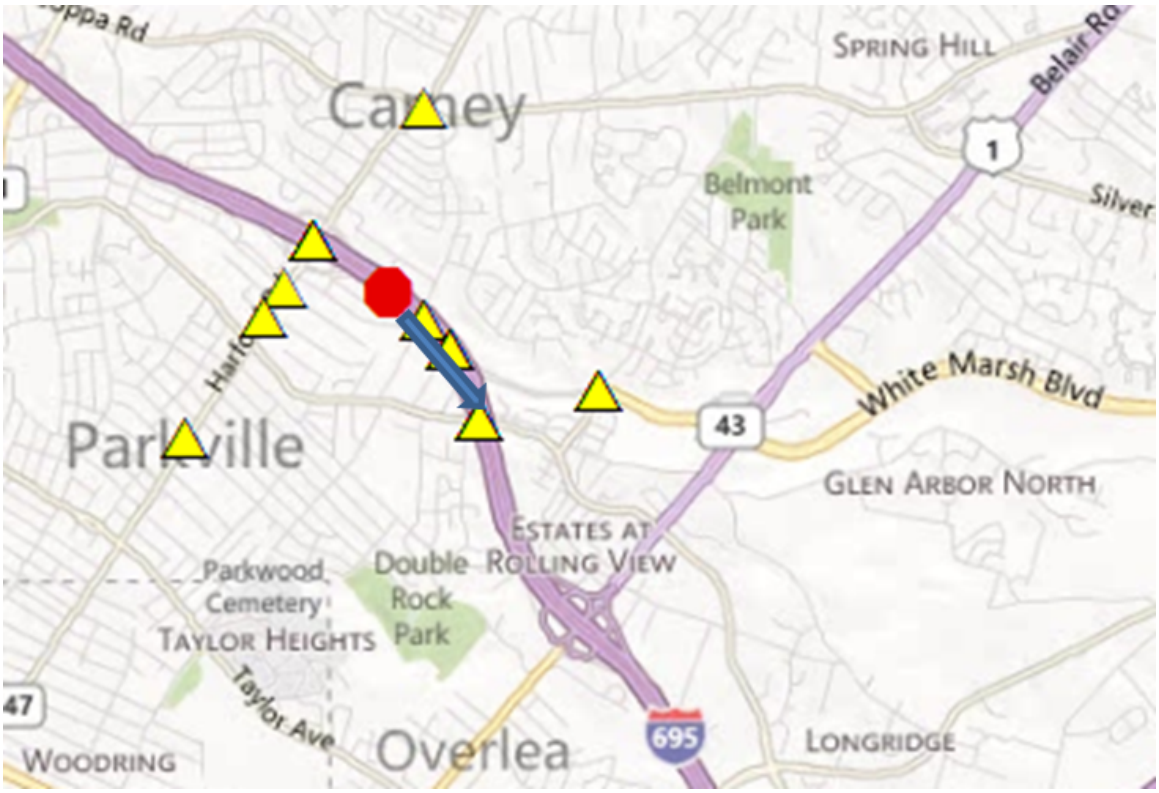


Figure 13.14: Map of the bottleneck at I-695 and MD-147

The detailed records of the bottleneck were cleaned up, and plotted on a time-time graph as shown in figure 13.15. In this figure, the each occurrence of the bottleneck are plotted as days on the abscissa, while the time of day is shown on the ordinate. The purple squares represent the start of the bottleneck, while the tail shows the duration of the bottleneck. Incidents are shown on the graph as red triangles. Figure 13.15 clearly shows that this bottleneck occurs regularly during the evening peak. There are some bottleneck occurrences that clearly fall outside the trend, like on January 19 and 26, February 22, and March 8, which are clearly caused by incidents, due to the presence of the incident triangle at the start of the bottleneck.

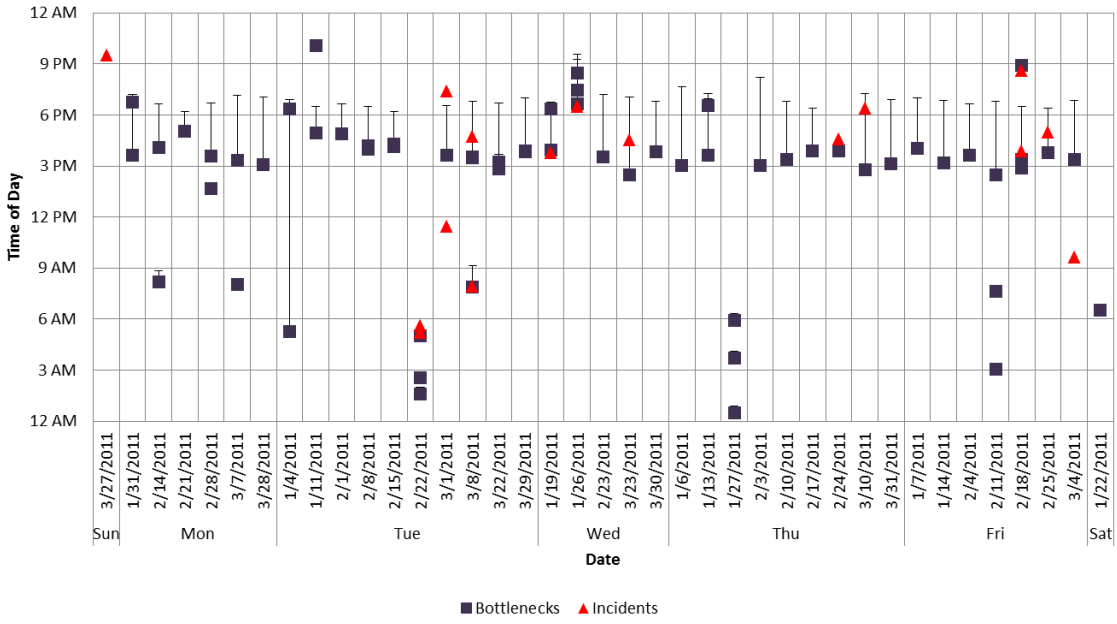


Figure 13.15: Time-time plot showing impact of incidents on bottleneck at I-695 and MD-147

13.4.2 2nd Quadrant Bottleneck at I-95 and MD-43

This bottleneck is from the second quarter of 2011. The bottleneck ranking algorithm counted 47 occurrences for this bottleneck in 3 months with an average maximum duration also of two and a half hours, and an average maximum length of 9.080 miles. Both bottlenecks seems to be quite similar, at the outset, except for the frequency of recurrence. The impact factor computed by the bottleneck ranking algorithm is a moderate 84 869 mile-hours. However, where this bottleneck differs from the one at I-695 and MD-147 is the number of impacting incidents. This bottleneck was impacted by 126 incidents during the same time frame. The location of this bottleneck is shown in figure 13.16. The circled incident shows the location where 96 of the 126 impacting incidents are located.

A similar time-time plot for this bottleneck clearly shows how the incidents impact almost every bottleneck. This plot is shown in figure 13.17. The occurrences of this bottleneck is also during the PM peak period, however, they always appear to be

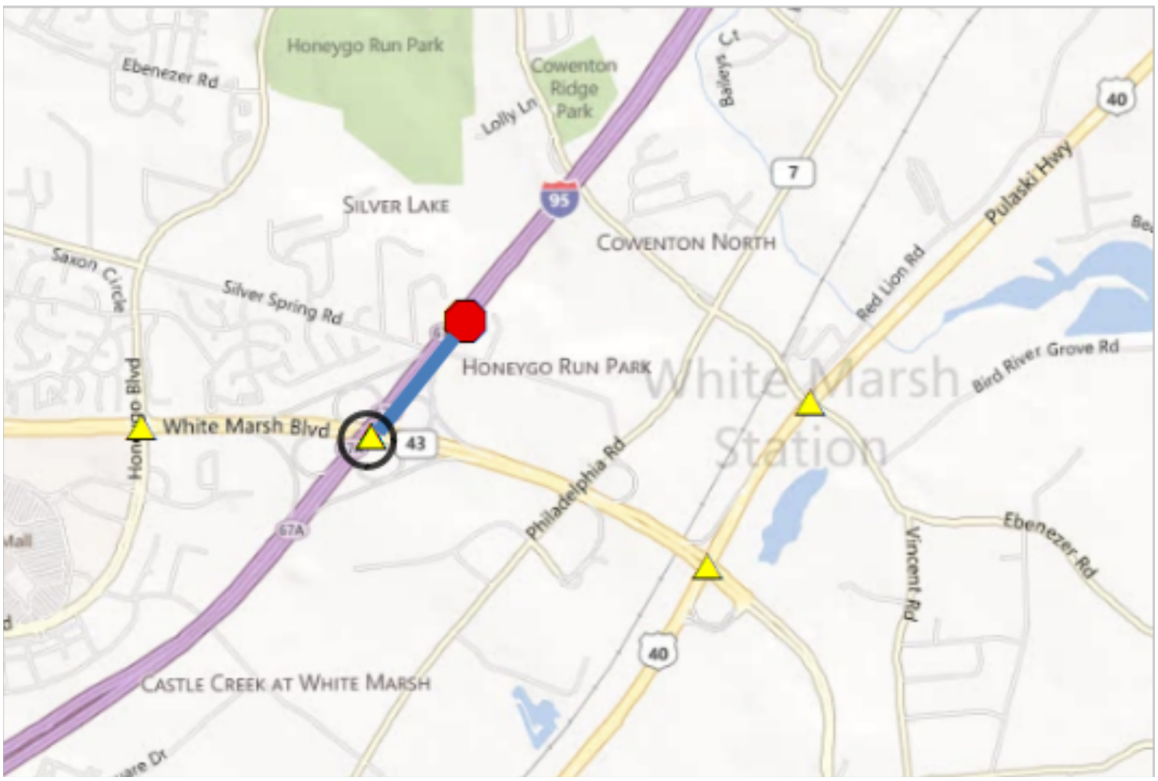


Figure 13.16: Map of the bottleneck at I-95 and MD-43

caused by incidents. The most frequent incident type at this location is Disabled Vehicle, which does not help to understand the reason behind this incident and bottleneck pair. However, taken at face value, a disabled vehicle might cause motorists to slow down, and hence start and cause a bottleneck. In figure 13.17 it can also be seen that some incidents keep a bottleneck active for longer, as they occur before the bottleneck has a chance to disappear.

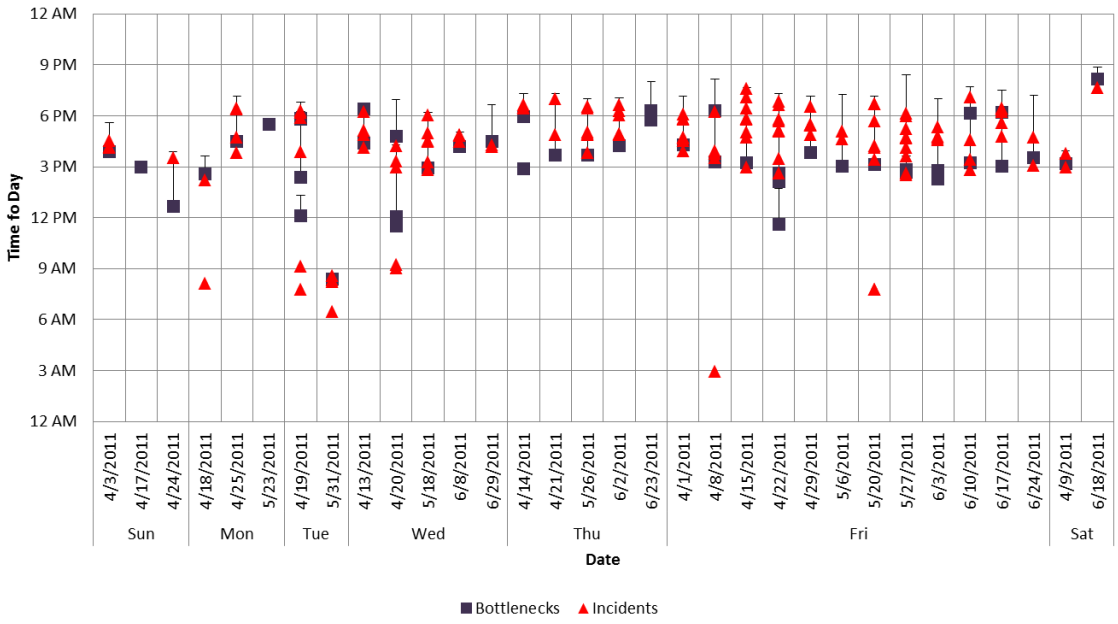


Figure 13.17: Time-time plot showing impact of incidents on bottleneck at I-95 and MD-43

Chapter 14: Conclusions Derived from Performance Measurement

In this chapter, summaries and conclusions are presented from the methods developed for various supply side performance measurement.

14.1 Synthetic Time Series Forecasting Method

The results presented in the previous chapter demonstrate the power of synthetic time series framework to predict traffic data with a good accuracy. Figure 13.4 shows that the maximum average RRMSPE over all periods in the ten days is just about 0.630. This error is acceptable in a naive model, without accounting for the volatility of traffic patterns due to the heavy dependency on environmental and human factors. Also, in terms of deviations from real value, figure 13.6, shows that the IQR is less than 10 %. Therefore, for most of the segments, the predicted value varies less than0.010 of the real value. The authors are unaware of a better result produced using the ARIMA family of models in the literature. In fact, the synthetic method outperforms many other, more advanced and complicated models presented in literature.

The proposed method is a foundation that should be expanded to include a multitude of factors. It is urged that future research in traffic prediction be conducted within the synthetic method, where applicable. The authors have planned certain enhancements to this work. Including various datasets, such as weather and incident data, into the ARIMA models as supplementary variables is part of one such planned

enhancement. Additionally, accounting for network effects, such as upstream and downstream conditions, as auxiliary information in the models is also planned.

Further, the scalability of the synthetic framework will allow for easy expansion of the network to include a larger area. For example, all of the Washington DC, Maryland, and Virginia roadway system can be easily modeled, over a continuous and expanding time window. These results will provide key information to all users of ITS, and support subsystems like AITS, and TMS in the region. Also, the scalability allows for network-wide route guidance systems to be implemented, which would drive better utilization of infrastructure and lower costs to all stakeholders. Lastly, the method can be quickly applied to any region in the world, with data acquisition arrangements, paving way for reliable ITS deployment.

14.2 Working with NPMRDS

The NPMRDS data is characterized and methods to increase its density, overcome the propensity of outliers, and quickly compute performance measures without the need for filtering, smoothing or imputation were demonstrated. The increased data density resulting from 24-hour overlay allows for outlier rejection without complex algorithms. The case studies performed highlight that the proposed method can be used to generate results from NPMRDS that resemble results from Bluetooth data, as evidenced by figures 13.7 and 13.8. Further, since VPP is subject to rigorous filtering and smoothing, it tends to underestimate travel time variation, especially on interrupted flow facilities. The most likely cause is perhaps vehicles stopped in queues (such as at traffic lights) are filtered out as outliers. Investigation of NPMRDS fidelity for use as the basis of performance measurement is recommended using the described methods on a statistically valid sample of segments representing freeways as well as varying functional classes of interrupted-flow arterials.

14.3 Spatial Conflation of Geospatial Datasets

A scoring algorithm for conflation was proposed here and was applied to the entire U.S. network. The total conflation error was found to be less than 7 %, which is commendable when compared with other results in the literature. For example, the probabilistic relaxation method proposed by Yang, Zhang, and Luan [132] has an error of about 5 % for a much smaller urban study area. By constraining the area, the types of geographies are limited, and therefore the scores in table 12.1 can be further tuned for best results. If the geographies are limited, the tuning can be accomplished as an optimization problem. However, the United States has wide ranging and vastly different geographies, which makes using a generic conflation algorithm difficult.

The dataset created by the conflation is being used by the National Renewable Energy Laboratory to answer specific questions about the effects of factors such as congestion, volume of traffic, gradient of roadways, and weather conditions on fuel use and greenhouse gas emissions. Figure 13.10, which present an innovative way to visualize the conflated data, were produced to aid one such study conducted at the National Renewable Energy Laboratory (NREL). A possible avenue for improvement of the conflation algorithm is to make the scores and thresholds more sensitive and directly dependent on the attributes of the segments. Additionally, successful join candidates can also be used to train advanced deep neural networks for better joins and to help resolve the issues with the current conflated dataset.

14.4 Merging Incidents and Congestion Datasets

The correlation between incidents and bottlenecks proved challenging due to the fidelity of the incident database and the variation in the methods used to geolocate incidents and bottlenecks. Even so, the nature and magnitude of the correlation was observed, and is best summarized in figure 13.12 which depicts a relation between the

severity of the bottleneck and recorded incidents.

Further analysis showed a high degree of correlation can be obtained in bottlenecks with a high count of incidents as compared to ones with a low count of incidents. In either case, it was shown that bottlenecks which fall outside the general trend are mostly attributable to incidents. Both of these are evident from figures 13.15 and 13.17 which provide a graphical method to visually assess the correlation between congestion and incidents. Extension and refinement of this graphical method was used to provide SHA with the appropriate patrol locations and paths.

However, the fidelity of the incident database and the unavailability of bottleneck evolution data (how the bottleneck initiated, grew, and merged with other bottlenecks) hinders more concrete and tangible conclusions which may involve a mathematical model to compute the extent of incident impact on bottlenecks. For further extending this branch of study, a rigorous quality control over the incident data is necessary along with more details in the bottleneck data.

Chapter 15: Dissertation Summary and Conclusions

It is necessary to have the domains of supply and demand of transportation work hand in hand to solve a majority of the current problems facing travelers on todays transportation networks. The methods described in this dissertation are geared towards both of these domains, providing separate solutions as required to tackle the most urgent problems and weakest link in existing methodologies. The difference in the problems faced in the supply and demand domains prompt solutions that are distinct. However, data fusion precipitates out of the methods documented in this dissertation, which is not surprising given the vast disparity in datasets available for transportation analysis.

In this dissertation, three-pronged contributions to the analysis efforts in transportation industry, affecting both supply and demand domains, are documented. The first involves a reliable model to generate synthetic population to small areas. Synthetic population is required to estimate the demand for facilities from behavioral models fitted to sample survey data. The second contribution is techniques designed to obtain real-time insights into roadway facilities, involving short-term forecasting to understand the expected trend of performance, and a data-driven model to quickly produce the necessary performance measures. The third contribution is born out of the existence of transportation data in spatially and temporally disparate datasets, and involves merging of road networks to transfer information like volume of vehicles, and incidents from one dataset to another.

The synthesis of population considering various characteristics of people was

demonstrated in the first section of this dissertation. Copulas were used to capture the dependence among various recorded characteristics of people, and then using those to synthesize the population with the same dependence for a different geographic area. The marginals of the synthesized population were then conformed to the totals of the region of interest. Not only does the method tackle the problem of small area estimation, by producing reliable synthetic population to small areas, but it also merges datsets together. In the first section of this dissertation, the American Community Survey (ACS) and the U.S. Decennial Census data were linked together using copulas. Further noteworthy is the handling of the temporal disparity, because the ACS data, were collected in 2016, while the census data were from 2010. The captured dependence represents the more recent data, and so do the number of synthesized individuals in each census tract. Census only supplies the relative frequencies attributes in each characteristic.

The developed method using copulas can be easily extended to merge various other datasets, and simultaneously synthesize population with characteristics sourced from the separate datasets. The National Household Travel Survey (NHTS) collects data from very few respondents in a state, and consequently cannot be reliably used at even the county level. However, using the ACS marginals can be obtained at the Public Use Micro Area (PUMA) level, where NHTS population can be synthesized by fitting copulas to the statewide NHTS data. Examples of other possible datasets that can be included with copula models to synthesize reliable population to the region of interest are administrative data like the Internal Revenue Service (IRS). These data can fill gaps in the data reported by the census, like census does not collect and report income information. Even big data like the GPS probe based datasets discussed in chapter 11 can be used to obtain marginal information like travel times, congestion metrics, etc. in the region of interest.

Performance measurement in transportation is a relatively new terminology, almost

born due to the economic constraints of the current age. The basic idea is to find the utilization of existing infrastructure, and improve upon it by considering the transportation network as a whole. This is radically different from the traditional practice of adding capacity, and roads as per the regional long-term plans. The reason behind the shift to performance measurement and management is to use the dwindling monetary reserves smartly, and make the most out of the existing infrastructure. It also increases sustainability, saves on wasted energy and time, and improves quality of life for the people.

The mandated rules under MAP-21 for continuous reporting of performance measures add to the pressures heaped on transportation agencies trying to get the most bang for their buck. Most mandated measures require fusing data from different spatial and temporal regions. An instance of temporal linkage was demonstrated while predicting short-term traffic, where data from previous periods were linked through a model to the current observed traffic patterns. Similarly, in producing performance measures from the GPS and Bluetooth probe datasets, data from multiple similar periods were combined to increase the density of data, which improves the reliability and accuracy of the produced results. Further, incidents near the head of congested segments were temporally linked with the congestion, and represented graphically to understand the relation between the two.

Although speed data are very useful for immediate insights into some aspects of the traffic, the complete picture is only obtained after incorporating other datasets, like incidents, volume, and weather to name a few. These data help estimate the global and system-wide impacts and externalities due to transportation. Some of these results are mandatory parts of reports generated under MAP-21. Therefore, considerable efforts were invested in producing conflation algorithms to merge the data together. A result documented in the second section of this dissertation from the conflation efforts was the merger of the speed and volume data together to produce

combined estimation of the resources consumed, time and energy spent, greenhouse gases emitted and other externalities of transportation. Similarly, the spatial and temporal conflation of the incident and congestion data provided insights into the causes of congestion, and the impact of further incidents on the congestion. However, the study had a limited scope due to the fidelity of the incident and congestion datasets.

Chapter 16: Proposed Future Extensions to Presented Work

In this dissertation, a two-pronged approach was taken to solve the most pressing needs in the supply and demand domains of transportation. The ultimate goal would be the merger of the two in a large modeling framework such that the impacts estimated in one area immediately predicts the impacts to the other areas. A promising modeling approach in practice with the potential to achieve this goal lies is in the Activity Based Models (ABM). ABM runs many interlinked models mainly aimed at producing the choices made by people as they go about their daily lives. These choices are then used to generate an origin–destination matrix and inform the expected demand on various transportation networks. Models in the ABM are also sensitive to the local and regional congestion and other negative costs due to traveling. Consequently, the framework needs to be run in loops to converge on the final set of expected demand and loads on network links.

Apart from reliable estimates of demand, the output can be compared against the typical trends of facility utilization observed using the real-time data over a sufficiently long period of time to minimize the impacts of current events. This would provide a check and balance to the demand models and assess the reality of their predictions. The merged volume and speed dataset produced by conflation algorithm documented in this dissertation would allow such comparisons to be direct for roadway based travel. The contribution of freight to the total volume and congestion on network links can be estimated by the use of datasets like the NPMRDS which provide separate data on freight vehicles and passenger vehicles.

With real-time and short-term predicted data for links forming transportation networks provided through the GPS big data, and the documented forecasting technique, it can be expected to improve the ABM fitting and performance. The tuned ABM models can then also be subjected to various policy scenarios to predict potential behavior of people and the likelihood of achieving long-term targets like reduction in greenhouse gas emissions, and adoption of more efficient mobility solutions. However, such results are only possible with a comprehensive database of various datasets in a similar temporal and spatial frame that can be leveraged to power the models. The methods developed and documented in this dissertation go a long way towards achieving the universal transportation dataset. The following subsections provide a high-level overview on further tasks required to achieve the goal of the universal dataset.

16.1 Universal Data Synthesis Model

In Small Area Estimation (SAE) literature, a hierarchical models have been developed, which are estimated using hierarchical Bayesian approaches. These models can not only provide the required estimates for variables of interest in small areas, in space, time, or population sub-domains, but also provide the complete posterior distribution function for the variables. These posterior distributions can be obtained as the marginal univariate distribution function of the variable of interest. Combined with the copula approach, the hierarchical Bayesian models allow synthesis of not only the basic socioeconomic and demographic characteristics of individuals but also more advanced variables for which the marginal information may otherwise be unobtainable.

The dataset thus produced would have all the information about people, including the transportation specific characteristics such as vehicles owned, departure time to work, travel time to work, distance traveled for business and leisure activities, income, educational achievements, location of home and work, presence of children, location

of schools for the children, and so on. This comprehensive dataset, along with a dataset that not only provides real-time and short-term predicted speeds, but also volumes, congestion indexes, performance measures and current and predicted weather information would provide holistic data for the most intricate ABM framework, that can predict choices of people, and estimate demand, even in real-time. Moreover, such data and models can be used to predict the likelihood of even the most difficult to forecast events, like incidents. Prediction of incidents would allow proactive positioning and patrolling to ensure not only minimal disruption to other users, but also provide critical and time-sensitive aid to the victims.

16.2 Population from Aggregate Data

Population synthesis through the use of copulas documented in this dissertation only works with sample data, as provided by the ACS. Extending the method to use aggregate data requires the synthesis of the dependence that is lost during aggregation. The uses would be in regions and countries that lack a reliable sample survey framework. Moreover, many surveys are reported in aggregate totals, rather than sample data. Such a method might be produced by the use of Monte Carlo Markov Chains (MCMC) to capture the hidden relationship within aggregated datasets. Conditional drawing through the use of MCMC is required because people's demographic characteristics are closely related to and dependent on each other. Such a study would be innovative, and would unlock the possibilities with data that otherwise find little use outside their specific domains.

16.3 Extending the Forecasting Framework

The prediction framework demonstrated in Chapter 12 has only been demonstrated for short-term forecasting of traffic speed data. Since the framework is flexible and

extensible, it can be modified to support forecasting of a multitude of variables of interest. Also, the method itself can be expanded to improve the prediction accuracy over the short-term. One of the simplest ways to improve the method is to use the conditions of the upstream and downstream segments as auxiliary variables while predicting the conditions on the current segment. Additionally, the data on incidents, volume and weather can be incorporated into the model through spatial conflation methods as documented in this dissertation.

It is also possible to modify the framework to provide predictions over other time horizons. Potential usefulness of having medium-term forecasts apply to the estimation of greenhouse gas emissions, and to the volume data in response to policies and market prices of gasoline. The latter can be supplied as auxiliary variables to the synthetic time-series framework.

Appendix A: Additional Monte Carlo Simulations with Copulas

In order to further reinforce that the copula captures and synthesizes data with the correct dependence structure, we provide results from additional simulations in this appendix. The simulations follow a triple objective:

1. to illustrate that the copula can recapture the parameter from synthesized data conforming to the same marginals as the ACS PUMS;
2. to show that the parameter is still re-captured even with drastic changes in marginals;
3. to empirically examine the sensitivity to the copula parameter value.

The simulations are all carried out using the Gumbel copula. The copula is used to generate two datasets of $n = 118\,583$ observations in $d = 9$ dimensions. The difference between the datasets are the copula parameter, which is set to $\theta = 1.059$ for the first dataset, representing the Gumbel copula parameter estimated from the PUMS data for PUMA 1107 (see table 6.3). Let this dataset be called $\mathcal{X}^{(RP)} = \{\mathbf{x}_i^{(RP)},\ i = 1, \dots, n\}$, where RP stands for "Real Parameter". The second dataset is created with $\theta = 1.500$ to showcase a different and stronger dependence structure, as it is further from one, the value of the parameter corresponding to the independence copula. Let this dataset be $\mathcal{X}^{(SP)} = \{\mathbf{x}_i^{(SP)},\ i = 1, \dots, n\}$, where SP stands for "Simulated Parameter".

The first dataset is inverse transformed using on one hand the marginals representing the PUMS data from PUMA 1107, and on the other with marginals representing the simulated dataset, with 10 different values per marginal, to create two different datasets. Continuing the nomenclature, the first of these datasets is called $\mathcal{X}^{(RPRM)}$, where RM indicates that the marginals of $\mathcal{X}^{(RPRM)}$ are indistinguishable from the marginals of the PUMS data. Similarly, the other dataset is called $\mathcal{X}^{(RPSM)}$, where SM indicates that the marginals are simulated and formed using (5.2). We apply the same two marginal transformations to the dataset $\mathcal{X}^{(SP)}$ to obtain two new datasets $\mathcal{X}^{(SPRM)}$, and $\mathcal{X}^{(SPSM)}$. The proposed copula estimation procedure is applied to all four resulting datasets. Results for $\mathcal{X}^{(SPSM)}$ have already been presented in table 5.5. Results obtained for the other three datasets $\mathcal{X}^{(RPRM)}$, $\mathcal{X}^{(RPSM)}$ and $\mathcal{X}^{(SPRM)}$ are discussed below.

Results obtained by applying the proposed procedure to dataset $\mathcal{X}^{(RPRM)}$ are presented in table A.1. The Gumbel copula parameter is almost correctly recovered. The p-values further reinforce that the results are very similar to those presented in table 6.3. Results obtained by applying the proposed procedure to dataset $\mathcal{X}^{(RPSM)}$ are presented in table A.2. The Gumbel copula parameter is again almost correctly recovered, illustrating that the captured dependence is independent of the influence of the marginal distribution. The results presented in table A.2 clearly show that any copula for which H_0 cannot be rejected is equally adept at reproducing the dependence structure, regardless of the marginals.

Table A.1: Constructed Archimedean Copulas for PUMA 1107

Copula	Parameter	CvM Statistic	p-value
Independent	NA	867.652	0.001
Clayton	0.160	716.146	0.001
Frank	NA	NA	NA
Gumbel	1.113	668.015	0.999
Joe	1.185	679.031	0.999

Table A.2: Constructed Archimedean Copulas for PUMA 1107

Copula	Parameter	CvM Statistic	p-value
Independent	NA	185.441	0.001
Clayton	0.146	145.084	0.001
Frank	0.478	147.829	0.001
Gumbel	1.076	120.433	0.999
Joe	1.134	109.844	0.999

Finally, results obtained by applying the proposed procedure to dataset $\mathcal{X}^{(SPRM)}$ are presented in table A.3. The estimated Gumbel copula parameter is close to the correct one. However, the p-values necessitate rejection of every copula except Joe copula in the results. Also, the Cramér–von Mises statistic values are drastically different from the results presented in tables 6.3, A.1 or A.2, but interestingly, the value is the smallest for the Gumbel copula. This suggests the statistical test proposed by Kojadinovic [66] has still some limitations and deserves more exploration, and that other tests to compare discrete multivariate distributions should be investigated in future research.

Table A.3: Constructed Archimedean Copulas for PUMA 1107

Copula	Parameter	CvM Statistic	p-value
Independent	NA	6 401.343	0.001
Clayton	0.641	3 827.180	0.001
Frank	3.110	2 883.169	0.001
Gumbel	1.484	2 785.526	0.006
Joe	1.726	3 104.049	0.999

Bibliography

[1] Baher Abdulhai, Himanshu Porwal, and Will Recker. "Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks". In: *California Partners for Advanced Transit and Highways (PATH)* (1999).

[2] Mohammed S Ahmed and Allen R Cook. *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*. 722. 1979.

[3] Theo A. Arentze and Harr J.P. Timmermans. "A learning based transportation oriented simulation system". In: *Transportation Research Part B* 38.7 (2004), pp. 613–633. DOI: 10.1016/j.trb.2002.10.001.

[4] Johan Barthelemy and Philippe L. Toint. "Synthetic population generation without a sample". In: *Transportation Science* 47.2 (2013), pp. 266–279. DOI: 10.1287/trsc.1120.0408.

[5] Richard J Beckman, Keith A Baggerly, and Michael D McKay. "Creating synthetic baseline populations". In: *Transportation Research Part A* 30.6 (1996), pp. 415–429. DOI: 10.1016/0965-8564(96)00004-3.

[6] BluSTATs. *BluSTATs version 1.94 Operations Manual*. 2012.

[7] David E Boyce and Huw Williams. "Urban travel forecasting in the U.S. and U.K." In: *Methods and Models in Transport and Telecommunications*. Springer, 2005, pp. 25–44. DOI: 10.1007/3-540-28550-4_3.

[8] Mark Bradley, John L Bowman, and Bruce Griesenbeck. "SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution". In: *Journal of Choice Modelling* 3.1 (2010), pp. 5–31. DOI: 10.1016/s1755-5345(13)70027-7.

[9] United States Census Bureau. *2010 Census*. 2010. URL: https://www.census.gov/programs-surveys/decennial-census/decade.2010.html (visited on 04/2018).

[10] United States Census Bureau. *2016 American Community Survey*. 2016. URL: https://www.census.gov/programs-surveys/acs/ (visited on 10/2017).

[11] Upper Midwest Reliability Resource Center. *Travel Time Reliability Reference Manual*. 2014. URL: http://en.wikibooks.org/wiki/Travel_Time_Reliability_Reference_Manual.

[12] Srinivasa Ravi Chandra and Haitham Al-Deek. "Predictions of freeway traffic speeds and volumes using vector autoregressive models". In: *Journal of Intelligent Transportation Systems* 13.2 (2009), pp. 53–72. DOI: 10.1080/15472450902858368.

[13] Liao Chen-Fu. "Using truck GPS data for freight performance analysis in the Twin Cities metro area". In: (2014).

[14] Bi Yu Chen, Hui Yuan, Qingquan Li, William HK Lam, Shih-Lung Shaw, and Ke Yan. "Map-matching algorithm for large-scale low-frequency floating car data". In: *International Journal of Geographical Information Science* 28.1 (2014), pp. 22–38. DOI: 10.1080/13658816.2013.816427.

[15] Steven I-Jy Chien and Chandra Mouly Kuchipudi. "Dynamic travel time prediction with real-time and historic data". In: *Journal of transportation engineering* 129.6 (2003), pp. 608–616. DOI: 10.1061/(asce)0733-947x(2003)129:6(608).

[16] Tom Choe, Alexander Skabardonis, and Pravin Varaiya. "Freeway performance measurement system: operational analysis tool". In: *Transportation Research Record: Journal of the Transportation Research Board* 1811 (2002), pp. 67–75. DOI: 10.3141/1811-08.

[17] Abdoul-Ahad Choupani and Amir Reza Mamdoohi. "Population synthesis using iterative proportional fitting (IPF): A review and future research". In: *Transportation Research Procedia* 17 (2016), pp. 223–233. DOI: 10.1016/j.trpro.2016.11.078.

[18] Cinzia Cirillo, Ying Han, Kartik Kaushik, and Parthasarathi Lahiri. "Synthetic time series technique for predicting network-wide road traffic". In: *Statistical Journal of the IAOS* 34.3 (2018), pp. 425–437.

[19] Maria A Cobb, Miyi J Chung, Harold Foley III, Frederick E Petry, Kevin B Shaw, and H Vincent Miller. "A rule-based approach for the conflation of attributed vector data". In: *GeoInformatica* 2.1 (1998), pp. 7–35.

[20] Gary A Davis, Nancy L Nihan, Mohammad M Hamed, and Leslie N Jacobson. "Adaptive forecasting of freeway traffic congestion". In: *Transportation Research Record* 1287 (1990).

[21] Paul Deheuvels. "La fonction de dépendance empirique et ses propriétés, un test non paramétrique d'indépendance". In: *Académie Royale de Belgique, Bulletin de la Classe des Sciences, 5e série* 65.6 (1979), pp. 274–292.

[22] W Edwards Deming and Frederick F Stephan. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known". In: *The Annals of Mathematical Statistics* 11.4 (1940), pp. 427–444. DOI: 10.1214/aoms/1177731829.

[23] Paul Embrechts, Filip Lindskog, and Alexander McNeil. *Modelling dependence with copulas*. Tech. rep. Zurich, Switzerland: Département de mathématiques, Institut Fédéral de Technologie de Zurich, 2001.

[24] Hongchao Fan, Bisheng Yang, Alexander Zipf, and Adam Rousell. "A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data". In: *International Journal of Geographical Information Science* 30.4 (2016), pp. 748–764. DOI: 10.1080/13658816.2015.1100732.

[25] Bilal Farooq, Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. "Simulation based population synthesis". In: *Transportation Research Part B* 58 (2013), pp. 243–263. DOI: 10.1016/j.trb.2013.09.012.

[26] Federal Highway Administration and U.S. Department of Transportation. "National Performance Management Measures; Assessing Performance of the National Highway System, Freight Movement on the Interstate System, and Congestion Mitigation and Air Quality Improvement Program". In: 82-FR-5970 (Feb. 2017), pp. 5970–6050. URL: https://www.federalregister.gov/documents/2017/01/18/2017-00681/national-performance-management-measures-assessing-performance-of-the-national-highway-system.

[27] Jean-David Fermanian. "Goodness-of-fit tests for copulas". In: *Journal of multivariate analysis* 95.1 (2005), pp. 119–152. DOI: 10.1016/j.jmva.2004.07.004.

[28] FHWA Office of Operations, Resource Center, HERE, and The Volpe Center. *Introduction to the National Performance Management Research Data Set (NPMRDS)*. 2013. URL: http://connectdot.connectsolutions.com/p27329s6h9l/.

[29] FHWA Office of Operations, Resource Center, HERE, and The Volpe Center. *Second Quarterly NPMRDS Webinar*. 2014. URL: https://connectdot.connectsolutions.com/p36vxid1rr5/.

[30] FHWA Office of Operations, Resource Center, HERE, and The Volpe Center. *Third Quarterly NPMRDS Webinar*. 2014. URL: https://connectdot.connectsolutions.com/p1ubotswuel/.

[31] FHWA Office of Performance Measurement Program. *Urban Congestion Report: Documentation and Definitions*. 2014. URL: http://www.ops.fhwa.dot.gov/perf_measurement/ucr/documentation.htm.

[32] Gaetano Fusco, Chiara Colombaroni, and Natalia Isaenko. "Short-term speed predictions exploiting big data on large urban road networks". In: *Transportation Research Part C: Emerging Technologies* 73 (2016), pp. 183–201. DOI: 10.1016/j.trc.2016.10.019.

[33] Christian Genest and Anne-Catherine Favre. "Everything you always wanted to know about copula modeling but were afraid to ask". In: *Journal of hydrologic engineering* 12.4 (2007), pp. 347–368. DOI: 10.1061/(asce)1084-0699(2007)12:4(347).

[34] Christian Genest, Kilani Ghoudi, and L-P Rivest. "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions". In: *Biometrika* 82.3 (1995), pp. 543–552. DOI: 10.2307/2337532.

[35] Christian Genest, Wanling Huang, and Jean-Marie Dufour. "A regularized goodness-of-fit test for copulas". In: *Journal de la Société Française de Statistique* 154.1 (2013), pp. 64–77.

[36] Christian Genest and Johanna Nešlehová. "A primer on copulas for count data". In: *ASTIN Bulletin: The Journal of the IAA* 37.2 (2007), pp. 475–515. DOI: 10.2143/ast.37.2.2024077.

[37] Christian Genest, Johanna G Nešlehová, Bruno Rémillard, et al. "On the empirical multilinear copula process for count data". In: *Bernoulli* 20.3 (2014), pp. 1344–1371. DOI: 10.3150/13-bej524.

[38] Christian Genest, Johanna Nešlehová, and Martin Ruppert. "Discussion: Statistical models and methods for dependence in insurance data". In: *Journal of the Korean Statistical Society* 40.2 (2011), pp. 141–148. DOI: 10.1016/j.jkss.2011.03.004.

[39] Christian Genest and Bruno Rémillard. "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 44. 6. Institut Henri Poincaré. 2008, pp. 1096–1127. DOI: 10.1214/07-aihp148.

[40] Christian Genest, Bruno Rémillard, and David Beaudoin. "Goodness-of-fit tests for copulas: A review and a power study". In: *Insurance: Mathematics and economics* 44.2 (2009), pp. 199–213. DOI: 10.1016/j.insmatheco.2007.10.005.

[41] Malay Ghosh and Parthasarathi Lahiri. "A hierarchical Bayes approach to small area estimation with auxiliary information". In: *Bayesian Analysis in Statistics and Econometrics*. Springer, 1992, pp. 107–125. DOI: 10.1007/978-1-4612-2944-5_6.

[42] Jean-François Girres and Guillaume Touya. "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4 (2010), pp. 435–459. DOI: 10.1111/j.1467-9671.2010.01203.x.

[43] Jessica Guo and Chandra Bhat. "Population synthesis for microsimulating travel behavior". In: *Transportation Research Record* 2014 (2007), pp. 92–101. DOI: 10.3141/2014-12.

[44] Antonin Guttman. *R-trees: A dynamic index structure for spatial searching*. Vol. 14. 2. ACM, 1984. DOI: 10.1145/602264.602266.

[45] Mordechai Haklay. "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning B: Planning and design* 37.4 (2010), pp. 682–703. DOI: 10.1068/b35097.

[46] Mohammad M Hamed, Hashem R Al-Masaeid, and Zahi M Bani Said. "Short-term prediction of traffic volume in urban arterials". In: *Journal of Transportation Engineering* 121.3 (1995), pp. 249–254. DOI: 10.1061/(asce)0733-947x(1995)121:3(249).

[47] Paula J Hammond. "The 2012 Congestion Report". In: *WSDOT's comprehensive annual analysis of state highway system performance* (2012).

[48] Sam Hawala and Partha Lahiri. "Variance modeling in the U.S. small area income and poverty estimates program for the American community survey". In: *Proceedings of the American Statistical Association, Bayesian Statistical Science Section, Survey Research Methods Section*. 2010, pp. 4655–4663.

[49] Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *Copula: Multivariate dependence with copulas, R package version 0.999-18*. 2018. URL: https://CRAN.R-project.org/package=copula.

[50] Sherif Ishak and Haitham Al-Deek. "Performance evaluation of short-term time-series traffic prediction model". In: *Journal of Transportation Engineering* 128.6 (2002), pp. 490–498. DOI: 10.1061/(asce)0733-947x(2002)128:6(490).

[51] Mahmoud Javanmardi, Joshua Auld, and K Mohammadian. "Integration of TRANSIMS with the ADAPTS Activity-Based Model". In: 4th Conference on Innovations in Travel Modeling. Tampa, Florida, U.S., 2011.

[52] Erik Jenelius and Erik Koutsopoulos. "Urban network travel time prediction based on a probabilistic principal component analysis model of probe data". In: *IEEE transactions on intelligent transportation systems (Print)* (2017). DOI: 10.1109/tits.2017.2703652.

[53] Byungduk Jeong, Wonjoon Lee, Deok-Soo Kim, and Hayong Shin. "Copula-Based Approach to Synthetic Population Generation". In: *PloS one* 11.8 (2016), e0159496. DOI: 10.1371/journal.pone.0159496.

[54] Jiming Jiang and P Lahiri. "Mixed model prediction and small area estimation". In: *Test* 15.1 (2006), pp. 1–96. DOI: 10.1007/bf02595419.

[55] Harry Joe. *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC, 1997. DOI: 10.1201/b13150.

[56] Harry Joe. "Asymptotic efficiency of the two-stage estimation method for copula-based models". In: *Journal of Multivariate Analysis* 94.2 (2005), pp. 401–419. DOI: 10.1016/j.jmva.2004.06.003.

[57] Harry Joe. *Dependence modeling with copulas*. CRC Press, 2014. DOI: 10.1201/b17116.

[58] Yiannis Kamarianakis and Poulicos Prastacos. "Space–time modeling of traffic flow". In: *Computers & Geosciences* 31.2 (2005), pp. 119–133. DOI: 10.1016/j.cageo.2004.05.012.

[59] Shih-Chieh Kao, Hoe Kim, Cheng Liu, Xiaohui Cui, and Budhendra Bhaduri. "Dependence-Preserving Approach to Synthesizing Household Characteristics". In: *Transportation Research Record* 2302 (2012), pp. 192–200. DOI: 10.3141/2302-21.

[60] Kartik Kaushik, Cinzia Cirillo, and Fabian Bastin. "On Modelling Human Population Characteristics with Copulas". In: *Procedia Computer Science* 151 (2019), pp. 210–217.

[61] Kartik Kaushik, Elham Sharifi, and Stanley Ernest Young. "Computing Performance Measures with National Performance Management Research Data Set". In: *Transportation Research Record: Journal of the Transportation Research Board* 2529 (2015), pp. 10–26. DOI: 10.3141/2529-02.

[62] Kartik Kaushik, Elham Sharifi, Stanley Ernest Young, and Babak Baghaei. "Comparison of National Performance Management Research Data Set (NPMRDS) with Bluetooth Traffic Monitoring (BTM) data and I-95 corridor coalition Vehicle Probe Project (VPP) data". In: 21st World Congress on Intelligent Transport Systems. Detroit, Sept. 8, 2014.

[63] Kartik Kaushik, Eric Wood, and Jeffrey Gonder. "Coupled Approximation of US Driving Speed and Volume Statistics using Spatial Conflation and Temporal Disaggregation". In: *Transportation Research Record* 2672.43 (2018), pp. 1–11.

[64] Gunky Kim, Mervyn J Silvapulle, and Paramsothy Silvapulle. "Comparison of semiparametric and parametric methods for estimating copulas". In: *Computational Statistics & Data Analysis* 51.6 (2007), pp. 2836–2850. DOI: 10.1016/j.csda.2006.10.009.

[65] Howard R Kirby, Susan M Watson, and Mark S Dougherty. "Should we use neural networks or statistical models for short-term motorway traffic forecasting?" In: *International Journal of Forecasting* 13.1 (1997), pp. 43–50. DOI: 10.1016/s0169-2070(96)00699-1.

[66] Ivan Kojadinovic. "Some copula inference procedures adapted to the presence of ties". In: *Computational Statistics & Data Analysis* 112 (2017), pp. 24–41. DOI: 10.1016/j.csda.2017.02.006.

[67] Ivan Kojadinovic and Jun Yan. "Comparison of three semiparametric methods for estimating dependence parameters in copula models". In: *Insurance: Mathematics and Economics* 47.1 (2010), pp. 52–63. DOI: 10.1016/j.insmatheco.2010.03.008.

[68] Ivan Kojadinovic and Jun Yan. "Modeling multivariate distributions with continuous margins using the copula R package". In: *Journal of Statistical Software* 34.9 (2010), pp. 1–20. DOI: 10.18637/jss.v034.i09.

[69] Ivan Kojadinovic and Jun Yan. "A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems". In: *Statistics and Computing* 21.1 (2011), pp. 17–30. DOI: 10.1007/s11222-009-9142-y.

[70] Ivan Kojadinovic, Jun Yan, and Mark Holmes. "Fast large-sample goodness-of-fit tests for copulas". In: *Statistica Sinica* 21 (2011), pp. 841–871. DOI: 10.5705/ss.2011.037a.

[71] Siem Jan Koopman, Rutger Lit, and André Lucas. *Intraday stock price dependence using dynamic discrete copula distributions*. Tinbergen Institute Discussion Paper 15-037/III/DSF90. Amsterdam and Rotterdam, The Netherlands: Tinbergen Institute, 2015. DOI: 10.2139/ssrn.2580840.

[72] Moshe Levin and Yen-Der Tsao. "On forecasting freeway occupancies and volumes (abridgment)". In: *Transportation Research Record* 773 (1980).

[73] Li Li, Xiaonan Su, Yanwei Wang, Yuetong Lin, Zhiheng Li, and Yuebiao Li. "Robust causal dependence mining in big data network and its application to traffic flow predictions". In: *Transportation Research Part C: Emerging Technologies* 58 (2015), pp. 292–307. DOI: 10.1016/j.trc.2015.03.003.

[74] Benmei Liu, Partha Lahiri, and Graham Kalton. "Hierarchical Bayes modeling of survey-weighted small area proportions". In: *Proceedings of the American Statistical Association, Survey Research Section*. 2007, pp. 3181–3186.

[75] Lu Ma. "Generating disaggregate population characteristics for input to travel-demand models". PhD thesis. University of Florida, 2011.

[76] Subrat Mahapatra, Matthew Wolniak, and Kaveh F. Sadabadi. *2016 Maryland State Highway Mobility Report*. Maryland State Highway Administration, Dec. 2016.

[77] Albert W Marshall and Ingram Olkin. "Families of multivariate distributions". In: *Journal of the American statistical association* 83.403 (1988), pp. 834–841. DOI: 10.2307/2289314.

[78] Kirill Müller and Kay W Axhausen. "Population synthesis for microsimulation: State of the art". In: *Arbeitsberichte Verkehrs-und Raumplanung* 638 (2010).

[79] Chumchoke Nanthawichit, Takashi Nakatsuji, and Hironori Suzuki. "Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway". In: *Transportation Research Record: Journal of the Transportation Research Board* 1855 (2003), pp. 49–59. DOI: 10.3141/1855-06.

[80] Pascal Neis, Dennis Zielstra, and Alexander Zipf. "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011". In: *Future Internet* 4.1 (2011), pp. 1–21. DOI: 10.3390/fi4010001.

[81] Roger B Nelsen. *An Introduction to Copulas*. Springer, 2006. DOI: 10.1007/978-1-4757-3076-0.

[82] Office of Highway Information Management. "Highway Performance Monitoring System (HPMS)". In: FHWA-PL-98-045 (1998).

[83] Simon Oh, Young-Ji Byon, Kitae Jang, and Hwasoo Yeo. "Short-term travel-time prediction on highway: a review of the data-driven approach". In: *Transport Reviews* 35.1 (2015), pp. 4–32. DOI: 10.1080/01441647.2014.992496.

[84] Ostap Okhrin, Alexander Ristig, and Ya-Fei Xu. "Copulae in high dimensions: an introduction". In: ed. by Wolfgang Karl Härdle, Cathy Yi-Hsuan Chen, and Ludger Overbeck. Third edition. Berlin, Germany: Springer, 2017. Chap. 13, pp. 247–277. DOI: 10.1007/978-3-662-54486-0_13.

[85] Iwao Okutani and Yorgos J Stephanedes. "Dynamic prediction of traffic volume through Kalman filtering theory". In: *Transportation Research Part B: Methodological* 18.1 (1984), pp. 1–11. DOI: 10.1016/0191-2615(84)90002-x.

[86] Stan Openshaw and Liang Rao. "Algorithms for reengineering 1991 Census geography". In: *Environment and planning A* 27.3 (1995), pp. 425–446. DOI: 10.1068/a270425.

[87] Elliott Irving Organick. "A Fortran Iv Primer". In: (1966). DOI: 10.2307/2311791.

[88] Darshan Mukund Pandit, Kartik Kaushik, and Cinzia Cirillo. "Coupling National Performance Management Research Data Set and the Highway Performance Monitoring System Datasets on a Geospatial Level". In: *Transportation Research Record* (2019), p. 0361198119838983.

[89] R Pappadà, F Durante, and G Salvadori. "Quantification of the environmental structural risk with spoiling ties: is randomization worthwhile?" In: *Stochastic Environmental Research and Risk Assessment* 31.10 (2017), pp. 2483–2497. DOI: 10.1007/s00477-016-1357-9.

[90] Fortunato Pesarin. *Multivariate permutation tests: with applications in biostatistics*. Vol. 240. Wiley Chichester, 2001.

[91] Dave Pierce and Dan Murray. "Cost of Congestion to the Trucking Industry". In: (2014).

[92] Abdul Rawoof Pinjari, Naveen Eluru, Rachel B Copperman, Ipek N Sener, Jessica Y Guo, Sivaramakrishnan Srinivasan, and Chandra R Bhat. *Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP Models, Framework, Software Architecture and Application Results*. Tech. rep. 2006.

[93] David R Pritchard and Eric J Miller. "Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously". In: *Transportation* 39.3 (2012), pp. 685–704. DOI: 10.1007/s11116-011-9367-4.

[94] David Robert Pritchard. "Synthesizing agents and relationships for land use/-transportation modelling". Master thesis. University of Toronto, 2008.

[95] Friedrich Pukelsheim and Bruno Simeone. "On the iterative proportional fitting procedure: Structure of accumulation points and L1-error analysis". In: Preprint Nr. 5/2009 (2009).

[96] Peter Rafferty and Chip Hankley. "National Performance Management Research Data Set (NPMRDS)". In: *Wisconsin Traffic Operations and Safety Laboratory* 12 (2014).

[97] J. N. K. Rao and Isabel Molina. *Small-Area Estimation*. Second edition. Wiley, 2015. DOI: 10.1002/9781118735855.

[98] Louis-Paul Rivest, Francois Verret, and Sophie Baillargeon. "Unit level small area estimation with copulas". In: *Canadian Journal of Statistics* 44.4 (2016), pp. 397–415. DOI: 10.1002/cjs.11296.

[99] Louis-Paul Rivest, François Verret, and Sophie Baillargeon. "Estimation of the parameters in copula models for small areas". In: *Statistical Society of Canada, Proceedings of the Survey Methods Section*. 2015.

[100] Roger P Roess, E Prassas, and William R McShane. *Traffic Engineering*. Fourth edition. Pearson, Upper Saddle River, NJ, 2011. ISBN: 9780136135739. DOI: 10.1007/978-1-349-10800-8.

[101] Juan J Ruiz, F Javier Ariza, Manuel A Urena, and Elidia B Blázquez. "Digital map conflation: a review of the process and a proposal for classification". In: *International Journal of Geographical Information Science* 25.9 (2011), pp. 1439–1466. DOI: 10.1080/13658816.2010.519707.

[102] Alan Saalfeld. "Conflation automated map compilation". In: *International Journal of Geographical Information System* 2.3 (1988), pp. 217–228. DOI: 10.1080/02693798808927897.

[103] Paul Salvini and Eric J Miller. "ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems". In: *Networks and Spatial Economics* 5.2 (2005), pp. 217–234. DOI: 10.1007/s11067-005-2630-5.

[104] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997. Chap. 9. DOI: 10.1201/9781439821862.

[105] D Schrank, B Eisele, T Lomax, and J Bak. *Appendix A: Methodology for the 2015 urban mobility scorecard*. Tech. rep. Technical report, Texas Transportation Institute, Texas A&M University, 2015. URL: https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-scorecard-2015-appx-a.pdf.

[106] David Schrank, Bill Eisele, and Tim Lomax. "T.T.I.'s 2012 urban mobility report". In: *Texas A&M Transportation Institute. The Texas A&M University System* (2012), p. 4.

[107] Elham Sharifi, Masoud Hamedi, Ali Haghani, and Hadi Sadrsadat. "Analysis of vehicle detection rate for bluetooth traffic sensors: A case study in maryland and delaware". In: 18th World Congress on Intelligent Transport Systems. 2011.

[108] w:en:Radim Baca Skinke. *Simple example of an R-tree for 2D rectangles*. URL: https://en.wikipedia.org/wiki/R-tree#/media/File:R-tree.svg.

[109] A Sklar. "Fonctions de répartition à n dimensions et leurs marges." In: *Publications de l'Institut de Statistique de l'Université de Paris* (1959).

[110] Brian L Smith and Michael J Demetsky. "Traffic flow forecasting: comparison of modeling approaches". In: *Journal of transportation engineering* 123.4 (1997), pp. 261–266. DOI: 10.1061/(asce)0733-947x(1997)123:4(261).

[111] Brian L Smith, Billy M Williams, and R Keith Oswald. "Comparison of parametric and nonparametric models for traffic flow forecasting". In: *Transportation Research Part C: Emerging Technologies* 10.4 (2002), pp. 303–321. DOI: 10.1016/s0968-090x(02)00009-8.

[112] Michael S Smith. "Bayesian approaches to copula modelling". In: *Bayesian Theory and Applications*. Ed. by Paul Damien, Petros Dellaportas, Nicholas G Polson, and David A Stephens. OUP Oxford, 2011, pp. 336–358. DOI: 10.2139/ssrn.1974297.

[113] Michael S Smith and Mohamad A Khaled. "Estimation of copula models with discrete margins via Bayesian data augmentation". In: *Journal of the American Statistical Association* 107.497 (2012), pp. 290–303. DOI: 10.2139/ssrn.1937983.

[114] TOPS Lab. *NPMRDS Travel Time Reliability*. 2014. URL: http://www.arcgis.com/home/item.html?id=7089b0b5870e4505a2f9f175c157563c.

[115] Alexandre Torday. "Simulation-based decision support system for real time traffic management". In: 89th Annual Meeting of the Transportation Research Board. 10-2120. Washington, DC, 2010.

[116] Transportation Research Board and National Academies of Sciences, Engineering, and Medicine. "Cost-effective performance measures for travel time delay, variation, and reliability". In: (2008). DOI: 10.17226/14167. URL: https://www.nap.edu/catalog/14167/cost-effective-performance-measures-for-travel-time-delay-variation-and-reliability.

[117] Dimitrios I Tselentis, Eleni I Vlahogianni, and Matthew G Karlaftis. "Improving short-term traffic forecasts: to combine models or not to combine?" In: *IET Intelligent Transport Systems* 9.2 (2014), pp. 193–201. DOI: 10.1049/iet-its.2013.0191.

[118] JWC Van Lint. "Online learning solutions for freeway travel time prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 9.1 (2008), pp. 38–47. DOI: 10.1109/tits.2007.915649.

[119] JWC Van Lint and CPIJ Van Hinsbergen. "Short-term traffic and travel time prediction models". In: *Artificial Intelligence Applications to Critical Transportation Issues* 22 (2012), pp. 22–41.

[120] Thaddeus Vincenty. "Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations". In: *Survey review* 23.176 (1975), pp. 88–93. DOI: 10.1179/sre.1975.23.176.88.

[121] Eleni I Vlahogianni, John C Golias, and Matthew G Karlaftis. "Short-term traffic forecasting: Overview of objectives and methods". In: *Transport reviews* 24.5 (2004), pp. 533–557. DOI: 10.1080/0144164042000195072.

[122] Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. "Short-term traffic forecasting: Where we are and where we're going". In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 3–19. DOI: 10.1016/j.trc.2014.01.005.

[123] David Voas and Paul Williamson. "An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata". In: *Population, Space and Place* 6.5 (2000), pp. 349–366. DOI: 10.1002/1099-1220(200009/10)6:5<349::aid-ijpg196>3.0.co;2-5.

[124] P Vythoulkas. "Alternative approaches to short term traffic forecasting for use in driver information systems". In: *Transportation and traffic theory* 12 (1993), pp. 485–506.

[125] Volker Walter and Dieter Fritsch. "Matching spatial data sets: a statistical approach". In: *International Journal of geographical information science* 13.5 (1999), pp. 445–473. DOI: 10.1080/136588199241157.

[126] Yibing Wang, Markos Papageorgiou, and Albert Messmer. "Real-time freeway traffic state estimation based on extended Kalman filter: Adaptive capabilities and real data testing". In: *Transportation Research Part A: Policy and Practice* 42.10 (2008), pp. 1340–1358. DOI: 10.1016/j.tra.2008.06.001.

[127] J Mark Ware and Christopher B Jones. "Matching and aligning features in overlayed coverages". In: *Proceedings of the 6th ACM international symposium on Advances in geographic information systems*. ACM. 1998, pp. 28–33. DOI: 10.1145/288692.288699.

[128] Joe Whittaker, Simon Garside, and Karel Lindveld. "Tracking and predicting a network traffic process". In: *International Journal of Forecasting* 13.1 (1997), pp. 51–61. DOI: 10.1016/s0169-2070(96)00700-5.

[129] Billy M Williams and Lester A Hoel. "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results". In: *Journal of transportation engineering* 129.6 (2003), pp. 664–672. DOI: 10.1061/(asce)0733-947x(2003)129:6(664).

[130] Billy Williams, Priya Durvasula, and Donald Brown. "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models". In: *Transportation Research Record: Journal of the Transportation Research Board* 1644 (1998), pp. 132–141. DOI: 10.3141/1644-14.

[131] Paul Williamson, Mark Birkin, and Phil H Rees. "The estimation of population microdata by using data from small area statistics and samples of anonymised records". In: *Environment and Planning A* 30.5 (1998), pp. 785–816. DOI: 10.1068/a300785.

[132] Bisheng Yang, Yunfei Zhang, and Xuechen Luan. "A probabilistic relaxation approach for matching road networks". In: *International Journal of Geographical Information Science* 27.2 (2013), pp. 319–338. DOI: 10.1080/13658816.2012.683486.

[133] Fan Yang, Zhaozheng Yin, Henry Liu, and Bin Ran. "Online recursive algorithm for short-term traffic prediction". In: *Transportation Research Record: Journal of the Transportation Research Board* 1879 (2004), pp. 1–8. DOI: 10.3141/1879-01.

[134] Xin Ye, Karthik Konduri, Ram M Pendyala, Bhargava Sana, and Paul Waddell. "A methodology to match distributions of both household and person attributes in the generation of synthetic populations". In: 88th Annual Meeting of the Transportation Research Board. Washington, DC, 2009.

[135] Jiyoun Yeon, Lily Elefteriadou, and Siriphong Lawphongpanich. "Travel time estimation on a freeway using Discrete Time Markov Chains". In: *Transportation Research Part B: Methodological* 42.4 (2008), pp. 325–338. DOI: 10.1016/j.trb.2007.08.005.

[136] Shuxin Yuan and Chuang Tao. "Development of conflation components". In: *Proceedings of Geoinformatics, Ann Arbor* (1999), pp. 1–13.

[137] Yanru Zhang. "Uncertainity Associated with Travel Time Prediction: Advanced Volatility Approaches and Ensemble Methods". Dissertation. University of Maryland, 2015.

[138] Yanru Zhang and Ali Haghani. "A gradient boosting method to improve travel time prediction". In: *Transportation Research Part C: Emerging Technologies* 58 (2015), pp. 308–324. DOI: 10.1016/j.trc.2015.02.019.

[139] Yanru Zhang, Yunlong Zhang, and Ali Haghani. "A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model". In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 65–78. DOI: 10.1016/j.trc.2013.11.011.

[140] Yunlong Zhang. "Special issue on short-term traffic flow forecasting". In: *Transportation research. Part C, Emerging technologies* 43 (2014), pp. 1–2. DOI: 10.1016/j.trc.2014.05.009.