EMERGENCY MEDICAL SERVICES DEMAND FORECASTING: MODERN MACHINE LEARNING APPROACHES FOR PRODUCING SHORT-TERM SPATIOTEMPORAL ESTIMATIONS

by

Richard Justin Martin

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing & Information Systems

Charlotte

2019

Approved by:

_____
Dr. Cem Saydam

_____
Dr. Mohamed Shehab

_____
Dr. Yaorong Ge

_____
Dr. Ertunga Ozelkan

ABSTRACT

RICHARD JUSTIN MARTIN. Emergency Medical Services Demand Forecasting: Modern Machine Learning Approaches for Producing Short-Term Spatiotemporal Estimations. (Under the direction of DR. CEM SAYDAM)

Emergency medical services (EMS); commonly referred to as ambulance, paramedic or pre-hospital emergency services, are a critical component in the delivery of urgent medical care to communities. EMS agencies, the organizations responsible for providing out-of-hospital acute medical care to the population of a specific service area, are confronted with the evolving task of effectively allocating the ambulances and medical personnel required to provide sufficient geographic coverage while minimizing response times to high-priority call requests. To meet this challenge, EMS practitioners and researchers have investigated the effectiveness of using various forecasting techniques for predicting future call volumes and demand densities. In this study, a forecasting methodology is proposed for producing spatiotemporal call volume predictions at a degree of granularity in time and space that is practical and actionable. A series of daily, hourly, and spatially distributed hourly call volume predictions are generated using a multi-layer perceptron (MLP) artificial neural network model following feature selection using an ensemble-based decision tree model. For spatially distributed predictions, K-Means clustering is applied to produce heterogeneous spatial clusters based on call location and associated call volume densities. The predictive performance of the MLP model is benchmarked against both a selection of traditional time-series forecasting techniques and a common industry method. Results show that MLP models outperform time-series and industry forecasting methods, particularly at finer levels of spatial granularity where the need for more accurate call volumes forecasts is more essential.

DEDICATION

I would like to dedicate this work to my mother and father, Carol Lynn and Rick Martin, who have always supported me in all of my endeavors and provided me with the opportunities to grow along the way; to my younger brother Brandon whose unwavering drive to succeed motivates me to keep pace every step of the way; and to my loving wife Brittany who never fails to spark my curiosity and challenges me to be better in every way.

ACKNOWLEDGEMENTS

First and foremost, I would like to recognize my dedicated and trusted advisor, Dr. Cem Saydam. I cannot thank him enough for his time, mentorship, and steadfast commitment to my academic career and success. I am proud to be his Last Mohican. I would also like to recognize the members of my dissertation committee Dr. Mohamed Shehab, Dr. Yaorong Ge, and Dr. Ertunga Ozelkan for their time and support. Finally, I would like to recognize the wonderful team members of MEDIC Mecklenburg EMS Agency who have supported this project from the beginning with their intellectual contributions and data. Thank you to Joe Penner, Dr. Jon Studnek, Allison Infinger, Jessica West, Chris Stephens, and all of the supporting Quality Improvement and Communications team members. I want to extend a special thanks to Jessica West for her diligent work in gathering and delivering the data used in this study, without such data this project would not be possible.

TABLE OF CONTENTS

LIST OF TABLES ..... **VIII**
LIST OF FIGURES ..... **IX**
CHAPTER 1: INTRODUCTION **1**
1.1 PROBLEM DOMAIN 1
1.2 EMS DEMAND FORECASTING 2
1.3 STUDY MOTIVATION AND EXPECTED CONTRIBUTION 3
1.4 STUDY OUTLINE 5
CHAPTER 2: LITERATURE REVIEW **6**
2.1 EMS DEMAND FORECASTING MODELS 6
2.2 COMPARISON OF METHODS 25
2.3 ALTERNATIVE MODELS AND METHODS 29
CHAPTER 3: DATA ANALYSIS **31**
3.1 DATA SOURCE 31
3.2 DATA STATISTICS AND PRE-PROCESSING 31
3.3 TEMPORAL FEATURE DECOMPOSITION AND ANALYSIS 35
CHAPTER 4: DAILY CALL VOLUME FORECASTS **40**
4.1 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE METHOD 40
4.2 THE MEDIC DAILY FORECASTING METHOD 43
4.3 DAY OF WEEK MOVING AVERAGE METHOD 44
4.4 HOLT-WINTERS TRIPLE EXPONENTIAL SMOOTHING METHOD 45
4.5 ARTIFICIAL NEURAL NETWORK METHOD 46
4.6 DAILY FEATURE SELECTION 48
4.7 ANN DEVELOPMENT FOR DAILY FORECASTS 53
4.8 DAILY CALL VOLUME FORECAST RESULTS 55
CHAPTER 5: HOURLY CALL VOLUME FORECASTS **57**
5.1 THE MEDIC HOURLY FORECASTING METHOD 57
5.2 HOURLY DAY OF WEEK MOVING AVERAGE METHOD 58
5.3 HOURLY FEATURE SELECTION 59
5.4 ANN DEVELOPMENT FOR HOURLY FORECASTS 60
5.5 HOURLY CALL VOLUME FORECAST RESULTS 61
CHAPTER 6: SPATIOTEMPORAL CALL VOLUME FORECASTS **63**
6.1 SPATIAL CLUSTERING 63
6.2 SPATIALLY DISTRIBUTED HOURLY CALL VOLUME FORECASTS 69
6.3 SPATIALLY DISTRIBUTED HOURLY FEATURE SELECTION 70
6.4 ANN DEVELOPMENT FOR SPATIALLY DIST. HOURLY FORECASTS 71

6.5 SPATIALLY DISTRIBUTED HOURLY FORECAST RESULTS 72
CHAPTER 7: SUMMARY AND CONCLUSIONS 74
7.1 DISCUSSION 74
7.2 STUDY LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH 77
REFERENCES 79
APPENDIX A: SUB-DAILY TIME FRAME CALL VOLUME FORECASTS 82
A.1 TEMPORAL CLUSTERING 82
A.2 SPATIALLY DISTRIBUTED TIME FRAME CALL VOLUME FORECASTS 85
A.3 SPATIALLY DISTRIBUTED TIME FRAME FEATURE SELECTION 85
A.4 ANN DEVELOPMENT FOR SPATIALLY DIST. TIME FRAME FORECASTS 86
A.5 SPATIALLY DISTRIBUTED TIME FRAME FORECAST RESULTS 86
APPENDIX B: GEOSPATIAL ANALYSIS 88

LIST OF TABLES

Table 1: EMS Demand Forecasting Study Comparison Matrix 26
Table 2: Annual Call Volume and Mecklenburg County Population Estimates 33
Table 3: MEDIC EMS Call Response Code & Priority Matrix 34
Table 4: Dispatch Priority Call Percentages 35
Table 5: Temporal Feature Labeling Schema 36
Table 6: Daily Call Volume Prediction Features Ranked Results 52
Table 7: Hourly Call Volume Prediction Features Ranked Results 60
Table 8: Total Call Volume Counts Per Cluster 67
Table 9: Spatial Hourly Call Volume Prediction Features Ranked Results 71
Table 10: Per Cluster MAPE/MAD Results by Method (7 Clusters) 73
Table 11: Per Cluster MAPE/MAD Results by Method (8 Clusters) 73
Table 12: Per Cluster MAPE/MAD Results by Method (9 Clusters) 73
Table 13: Per Cluster MAPE/MAD Results by Time Frame Method (9 Clusters) 86

LIST OF FIGURES

Figure 1: Total Call Volume by Year in Mecklenburg County, NC 32
Figure 2: Total P1+P2 Call Volume by Month (2010 – 2017) 36
Figure 3: Total P1+P2 Call Volume by Month (2014 – 2017) 37
Figure 4: Total P1+P2 Call Volume by Day of Week (2010 – 2017) 38
Figure 5: Average P1+P2 Call Volume by Day of Week (2010 – 2017) 39
Figure 6: P1+P2 Daily Call Volumes Rolling Mean and Std. Dev. (2010-2017) 41
Figure 7: Single Hidden Layer MLP Model 47
Figure 8: DOWMA Method MAPE Values (Periods 1-365) 56
Figure 9: H-DOWMA Method MAPE Values (Periods 1-350) 62
Figure 10: P1+P2 Spatial Clusters Elbow Plot 66
Figure 11: P1+P2 Calls K-Mean Spatial Clusters; K=7 (2010-2017) 68
Figure 12: P1+P2 Calls K-Mean Spatial Clusters; K=8 (2010-2017) 68
Figure 13: P1+P2 Calls K-Mean Spatial Clusters; K=9 (2010-2017) 69
Figure 14: Percent of Total Call Volume and MAPE Values by Cluster 75
Figure 15: MLP vs. MHF Absolute Error Frequency Distributions by Cluster 76
Figure 16: Avg. P1+P2 Avg. Hourly Call Volume Time Clusters (2010-2017) 83
Figure 17: P1+P2 Avg. Hourly Call Volume Time Clusters Elbow Plot 84
Figure 18: P1+P2 TCC-A Sample Time Clusters by Day of Week (2010-2017) 84

CHAPTER 1: INTRODUCTION

1.1 PROBLEM DOMAIN

Emergency medical services (EMS); commonly referred to as ambulance, paramedic or pre-hospital emergency services, are a critical component in the delivery of urgent medical care to communities. EMS agencies are organizations charged with the responsibility of providing out-of-hospital acute medical care to the population of a defined geographic region such as a county, city or local municipality. Depending on the local health care infrastructure, these agencies may be owned and operated by local governments, health care systems, or private organizations. EMS agencies routinely provide transportation to local clinical care facilities, such as hospitals and emergency departments, for patients who are unable to transport themselves due to the nature of their condition or circumstances. The primary goal of any EMS agency is to minimize their response times to high priority emergency call requests and lessen the rate of mortality and morbidity [1]. To adequately serve the population of their service regions, EMS managers and dispatchers continuously study the distribution of incoming call requests (demand) and establish resource deployment plans specifying the number of ambulances and emergency response personnel required for future periods. These deployment plans, derived from historical demand data and forecasts, determine the daily personnel work shift schedules. Throughout the course of a day, shifts may also be staggered and/or overlapped to account for hourly fluctuations in call volume demand and anticipated personnel fatigue [2]. While fundamentally deployment decisions are data and forecast driven, EMS dispatchers frequently describe the deployment planning process, and associated redeployment decisions, as an art based on the experiences and intuition of the individual [2].

1.2 EMS DEMAND FORECASTING

By their very nature, EMS systems are extraordinarily complex. The demand for ambulances is dynamic and is known to fluctuate spatially and temporally based on the time of day and day of the week [3]. EMS managers and dispatchers are faced with the evolving task of deploying ambulances and personnel required to provide adequate coverage for a defined geographic service area given limited resources. Dispatchers continuously redistribute (redeploy) their fleet of ambulances to different locations, often referred to as posts, throughout the day to compensate for spatiotemporal demand fluctuations. However, the scope of these adjustments is restricted by the pre-determined shift staffing plans for a given period. In instances where the incoming call volume exceeds the number of available ambulances, resources are allocated strictly based on call priority which is determined by observed patient severity. While higher priority calls take precedence, less severe call requests remain in queue until such time when sufficient resources become available [2]. Industry and academic researchers have conducted numerous studies focused on developing novel deployment strategies, and associated staffing plans, in an effort to reduce response time while maximizing service coverage [4]. These deployment models, typically developed based on historical data, are ultimately dependent on detailed 9-1-1 emergency call demand forecasts to serve as inputs. [3, 5, 6]. In a recent survey study, Aringhieri et al. [1] conducted a broad literature review of Emergency Medical Services related research and grouped papers into one of the following seven categories; *"1. Ambulance Location Problems, 2. Ambulance Relocation Problems, 3. Ambulance Dispatching and Routing Policies, 4. Interplay with other emergency health care delivery systems, 5. Evaluation and validation of EMS systems performance, 6.*

Forecasting, and, 7. Workforce Management". Further exploring the papers that fell into the forecasting category, the authors aggregated the studies into groups of papers that focused on forecasting EMS call demand, response travel time, and ambulance crew workloads. By investigating and cataloging a wide range of literature Aringhieri et al. were the first to clearly draw distinctions between the various fields of EMS research. Despite the critical importance of producing accurate forecasts, significantly fewer studies have been focused on establishing more sophisticated call forecasting approaches to improve the predictive models used for EMS demand planning.

1.3 STUDY MOTIVATION AND EXPECTED CONTRIBUTION

Regardless of past research efforts in the EMS community, the prevailing industry trend remains to use simple time-series forecasting methods, such as moving averages and proprietary formulas built into computer aided dispatch software, to predict future call volume and demand densities. While these types of forecasts are valuable for strategic and tactical capacity planning at broader geographical scales over longer periods of time (i.e. monthly or weekly), they do not adequately support short-term operational decisions, such as daily and hourly deployments and redeployments [5]. Furthermore, their accuracy decreases significantly when aggregated at finer spatial scales. Reviewed in the forthcoming chapter, the current literature presents no formal benchmark for comparing forecast precision or accuracy within emergency medical services systems. This prompted Brown et al. [7] and Setzler et al. [5] to conduct unique studies assessing the performance and reliability of current industry practices against novel forecasting approaches. They concluded that the improvements of newer, more advanced, methods have been marginal

by comparison to simple time-series methods and suffer from an increased level of complexity and number of underlying assumptions. This suggests that widespread industry adoption requires practical and elegant solutions that clearly outperform the current practices. Moreover, the majority of historic approaches have focused on time of the day as the sole dimension and generated forecast predictions for larger geographic areas such as entire counties or cities. Only in recent years have researchers began recognizing the importance of incorporating the temporal and spatial distribution of demand into their prediction models. The primary objective of this study is to establish a methodology for producing accurate spatiotemporal call volume forecasts that outperform current industry practices at a degree of granularity in time and space that is practical, actionable, and supports widespread adoption. A series of daily, hourly, and spatially distributed hourly call volume predictions are generated using a multi-layer perceptron (MLP) artificial neural network model following feature selection using an ensemble-based decision tree model. For spatially distributed predictions, K-Means clustering is applied to produce heterogeneous spatial clusters based on call location and associated call volume densities. The predictive performance of each MLP model is then benchmarked against both a selection of traditional time-series forecasting techniques and a common industry method. Results show that MLP models outperform time-series and industry forecasting methods, particularly at finer levels of spatial granularity where the need for more accurate call volumes forecasts is more essential. This study is being conducted under the endorsement of MEDIC, the Mecklenburg County EMS Agency. While MEDIC has been accepted in various capacities as an industry leader, agency directors at MEDIC continue to recognize and express the need for more accurate demand planning and prediction capabilities that

also incorporates the spatial distribution of demand. The results of this study are expected to improve upon current industry practices and progress the current state of the art presented in the literature.

1.4 STUDY OUTLINE

This study is organized in the following manner. In Chapter 2, a comprehensive review of the prominent research related to EMS demand forecasting is conducted and the various forecasting methods are compared based on their contribution to the field and effectiveness in application. Additionally, the principal strengths and weakness of the various models are summarized, and alternative forecasting methods and strategies are discussed. In Chapter 3, the data used in this study is reviewed, summarized, and pre-processed. In Chapters 4 and 5, non-spatially distributed daily and hourly call volume forecasting methods and comparative results are presented. Chapter 6 details the development and results of spatiotemporal call volume forecasting methods; including the implementation of a k-means clustering model as an alternative approach to fixed-grid geographic segmentation. Conclusions and suggestions for future studies are summarized in Chapter 7. Lastly, the effectiveness of an exploratory time frame clustering approach for producing sub-daily call volume predictions is examined in Appendix A and supplemental geospatial analysis work is included in Appendix B.

CHAPTER 2: LITERATURE REVIEW

2.1 EMS DEMAND FORECASTING MODELS

As noted by Aringhieri et al. [1], Channouf et al. [3], and McConnel and Wilson [8], some of the earliest works related to emergency medical services demand forecasting appeared in the 1970s. For instance in 1971, Hall [9] was one of the first to collect a sample of actual EMS call volume data in order to identify any latent trends or cycles in hourly, daily or weekly demand. While the paper was not focused specifically on forecasting, Hall formulated a systematic approach for evaluating the demand for ambulances in an urban setting. Using a one-month data set, provided by local emergency responder agencies in Detroit, Michigan, Hall performed a regression analysis that measured the relationship between the ambulance travel times and travel distances. Additionally, other independent variables were considered such as the type of emergency, traffic and weather conditions but were found to have little explanatory power given the available data [9]. Hall's research objective in analyzing this demand data was to develop a model of various operating policies that agencies could implement to determine the optimal number of ambulances to deploy in a geographic area. Despite the absence of a concrete forecasting model, Hall's work represents one of the earliest investigations focused on gaining a deeper understanding on the nature of EMS demand.

Aldrich et al. [10] conducted a separate study in 1971 that applied regression analysis to evaluate the nature of demand for emergency medical services over a one-year time period. The data used for their study consisted of ambulance trip records from the Los Angeles Central Receiving Hospital coupled with 1960 Los Angeles census data. The census data

enabled the authors to include socioeconomic characteristics such as age, sex, income, education level, marital status, and race as independent variables to estimate the number and type of medical emergency calls that would be received over the course of a year. As is common with census data, the information collected in 1960 was aggregated into geographic areas known as census tracts. Each of the individual ambulance call records provided for the study also included this census tract field, which enabled the researchers to effortlessly merge the datasets. Despite the fact that the data was already aggregated into geographic clusters Aldrich et al. designed their model to predict at the per capita level and not at the census tract level; citing that while the location is pertinent to operating policy, geographic units add complexity to the analysis [10]. With the advent of computer models and algorithms, this complexity limitation can be overcome by researchers seeking to add more meaning and accuracy to their models (refer to Setzler et al. [5]). The census tract information used also included land use variables that were factored into the model, such as the number of housing units and employment within the area. Similar to Hall [9], Aldrich et al. considered including weather as a factor contributing to emergency call demand but opted to exclude it from their final model. This decision was likely driven by the difficulty associated with collecting sufficient weather-related data to include in their analysis. It is notable to point out this omission, as we will see in later research, as several subsequent researchers postulate similar justifications for considering the inclusion of weather data but ultimately elect to exclude it. Aldrich et al. [10] concluded that "*the demand for public ambulances appears to be highly predictable, using a simple linear model employing socioeconomic variables, quality of service variables and land use variables.*" Given the demographic information available they found that areas with higher densities of children

and elderly people generate higher call volumes and that nonwhite and low-income families are more inclined to utilize public ambulance services [10]. Overall the major contribution of this study was the integration of socioeconomic and demographic data. Aldrich et al. were also the first in the EMS forecasting community to hypothesize a correlation between a population's level of access to health care and their use of emergency medicine; specifically stating that "*we expect that demand would be highest in tracts (areas) with a concentration of people of low socioeconomic status. These people may use the emergency system even in the absence of real emergencies because they generally do not have a regular physician*." A population's access to health care continues as a major theme today within emergency medical services research.

Siler [11] performed a study similar to that of Aldrich et al. [10] using comparable socioeconomic data sets. Both investigations carried out by Siler and Aldrich et al. used census information, land use data, ambulance call records, and multiple regression models to analyze the demand characteristics for emergency medical services over the course of a one-year period. However, Siler outlined several sharp distinctions between his model and the one developed by Aldrich et al. While Aldrich et al. used 1960 census tract information to study demand at the per capita level, Siler aggregated 1970 census records at the community level. Siler also used additional forms of socioeconomic variables, including non-linear forms, and concentrated on the relationships between the employment levels in residential and non-residential communities as determinants for the number of incidents requiring emergency medical services. Siler challenged the findings of Aldrich et al. stating that their models were over-specified and that the effects of several socioeconomic

variables were inconsistent from one regression model to the next within their study [11, 12]. Siler's final regression model contained the following four independent variables: *"(1) employment in an area as a proportion of the resident population, (2) the proportion of resident population that is both white and married, (3) housing units per residential area, and, (4) the ratio of white-collar to blue-collar employment among female residents"*. All four of the independent variables had a nonlinear relationship to the number of incidents. The employment in an area as a proportion of the resident population was found to have the most significant impact on demand for emergency medical services. This finding supported Siler's initial assertion that population density shifts during the day as residents leave their homes and travel to commercial areas throughout the city. Accordingly, Siler hypothesized that demand for ambulances would migrate with the population. Siler's model found that, similar to Aldrich et al., *"demand is higher for unmarried whites and for nonwhites regardless of marital status"*, that *"demand is higher the fewer occupants there are per housing unit"*, and finally that *"demand is greater the higher the proportion of blue-collar relative to white-collar job holders among employed female residents"* [11]. Serving as an extension of the initial investigation carried out by Aldrich et al., Siler contributed to studying the relationship between various socioeconomic measures and the demand for emergency medical services. Furthermore, Siler was the first to incorporate socioeconomic variables to represent the changes in population densities at the community level as people migrate from residential to commercial areas throughout the day.

Both studies conducted by Siler [11] and Aldrich et al. [10] used data restricted to the Los Angeles, California area, and the authors highlighted this as an important limitation in the

discussion of their results. This prompted Kvålseth and Deems [13] to conduct a similar investigation attempting to produce analogous results using data collected from the Atlanta, Georgia area. The time horizon for their study covered a one-month period and, like Siler and Aldrich et al., used ambulance calls records and 1970's census data merged by census tract. Following initial regressions runs, the authors found 18 exogenous variables to be significant in estimating the demand for ambulances. Comparable to Siler and Aldrich et al., the author's results showed that variables associated with demographics, economics and land-use could be used to explain variations in demand. Specifically, they found that increases in EMS demand corresponded to declining household income, increasing male unemployment, larger land acreage per capita, and the percent of non-white populations. Likewise, they concluded that demand for emergency medical services increases as the percent of the population under the age of 15 years and over the age of 65 years increases. While Kvålseth and Deems [13] did not cite the works of Siler specifically, overall their results were consistent with the general findings of both Aldrich et al.'s and Siler's investigations.

Kamenetzky et al. [12] completed a regression analysis study concentrated on establishing demand estimation models using comprehensive sets of ambulance call and census data. The authors claimed that previous investigations attempting to establish links between EMS demand and populations characteristics fell short by using incomplete datasets [12]. To perform their study, a series of step-wise multiple regression models were carried out using a combination of 1970's census data, population and employment estimates for 1979, and complete sets of 1979 call data provided by 82 EMS agencies throughout 200 civil

divisions in Southwestern Pennsylvania. This represented the largest and most geographically diverse collection of EMS data analyzed prior to 1982. In addition to developing models focused on estimating the number of emergent calls using historical call data and sociodemographic features, the researchers attempted to analyze the effect of provider characteristics on demand. This required dividing the agencies represented in the data into different service characteristic categories such as profit/non-profit, paid/volunteer, public/private, and community/municipal. However, preliminary tests indicated that the effect of provider characteristics on demand was not statistically significant when other factors were considered. Their final regression model produced reliable demand estimates for a given area using the following four sociodemographic indicators: *"(1) total resident population, (2) total employment in the area, (3) logarithm of the percentage of the population which is both white and married, and (4) the square of housing units per area resident"* [12]. These four variables, with the exception of employment, were taken from census data and are comparable to the variables used in the final models formulated by Kvålseth and Deems [13], Siler [11] and Aldrich et al. [10]. Kamenetzky et al. [12] were the first and only researchers to consider the effect of provider characteristics on demand. Their final model was also very unique compared to other approaches given that it incorporated estimations of unmet demand and differentiated between demand based on clinical categories.

Recognizing the fact that previous studies only leveraged regression approaches to analyze and forecast EMS demand, Baker and Fitzpatrick [14] decided to apply a Winters' exponential smoothing model to generate daily ambulance demand forecasts. Using data

provided by the South Carolina Department of Health and Environmental Control, the researchers selected four different counties in South Carolina, varying from urban to rural, to serve as their study area. The authors intentionally selected a diverse group of counties to overcome what they identified as a key limitation of previous studies focused exclusively on a single geographic area. They hypothesized that applying a Winter's exponential smoothing time-series model to produce short-term forecasts would be a more practical alternative for EMS managers over traditional multiple linear regression models that generate longer-term estimations and rely heavily on casual variables [14]. In an effort to optimize their exponential smoothing model, Baker and Fitzpatrick also developed goal and quadratic programming methods that adjusted the model parameter settings to minimize the mean squared error (MSE) and BIAS. The BIAS statistic they developed served as a gauge on the direction of the average forecast error. A positive BIAS indicated an over-estimation of average demand, while a negative BIAS represented under-estimation. Given that EMS resources are allocated and dispatched based on demand forecasts, the authors stated that under-estimation of demand is a more severe inaccuracy and parameters settings that produce over-estimated forecasts should be preferred [14]. The authors configured their model so the desired value and direction of the BIAS statistic could be adjusted by the decision maker (i.e. EMS dispatchers and managers). They compared the results of their multi-stage programming and forecasting approach against results produced by a single-objective Winters' exponential smoothing model and a multiple linear regression forecasting model. Overall, they found their approach produced more accurate forecasts for the various counties over competing models while also supporting the objectives of the decision makers surveyed [14]. Baker and Fitzpatrick were

the first researchers to explore the application of time series models to produce daily call volume predictions.

In 1998, Tandberg et al. [15] carried out a similar study exploring the performance of various time series models for generating ambulance forecasts. Specifically, the researchers applied four different models that included a model using raw observations from previous periods, a simple moving average model with data point periods ranging from one through five, a means with simple moving average (seasonal decomposition) model with data point periods ranging from one through five, and an autoregressive integrated moving (ARIMA) model [15]. Using a full year of call data (1994) provided by the Albuquerque Ambulance Service in Albuquerque, New Mexico, hourly call volume predictions were generated using all four models. The prediction accuracy of each model was then tested using records for the first 24 weeks of the following year (1995) and forecast performance was evaluated using the resulting R-squared values. After testing and evaluating a series of different model types and configurations, the authors found that the means with simple moving average (seasonal decomposition) model using a three-point moving average produced the most accurate forecast. Overall their final model explained 54.3% of the variation present in the 1995 test data [15]. In addition to generating forecasts for the volume of calls in a given hour (ambulance runs per hour), Tandberg et al. [15] also attempted to produce predictions for the average service times and call priorities for given calls but did not yield useable results. Their study represented the first attempt in the EMS forecasting community to produce hourly short-range call volume predictions.

Citing specific concerns over the rising senior citizen population, McConnel and Wilson [8] carried out a study examining the impacts of an aging society on emergency medical services demand. Consistent with preceding investigations, McConnel and Wilson collected one year's worth of ambulance calls records and census data from 1990 representing the population of Dallas, Texas. The primary objective of their study was to measure the demand disparities between different age and racial groups. Regarding methodology, McConnel and Wilson elected to use a variety of statistical tests to evaluate their hypotheses including multiple Chi-Squared tests and one instance of a Tukey's range test for multiple comparisons of proportions [8]. They began their analysis by calculating EMS utilization rates per thousand people based on two gender categories (male and female), four different categories representing population ethnicity (non-Hispanic white, African American, Hispanic, and other), and eight different age groups (i.e.; <5, 5-14, 15-24, 25-44, 45-64, 65-74, 75-84, 85+). As expected, and supported by previous research [10, 11, 13], McConnel and Wilson [8] determined that the demand for emergency medical services increases as the age of the population increases. Based on their findings, rates of utilization for those 85 years of age or older are 3.4 times higher compared to those falling in the (45-64) year old age group. Furthermore, results showed that African Americans were twice as likely to utilize emergency medical services compared to Non-Hispanic Whites and Hispanics and almost five times as likely compared to the other ethnicity categories [8]. This finding is also consistent with previous studies [10, 11, 13]. Lastly, McConnel and Wilson found overall EMS utilization rates for females to be noticeably higher when considering all incidents types, but lower for life-threatening situations, when compared to male utilization rates. While age and other demographic characteristics were

included as independent variables in several of the previous studies presented, McConnel and Wilson were the first to perform a study explicitly focused on measuring the demand disparities between various age groups drawing attention to the potential future impacts of an aging society on EMS systems and agencies.

Nearly a decade later in 2007, Channouf et al. [3] performed a study concentrated on generating daily and hourly EMS call volume forecasts by applying a variety time-series methods. Emphasizing the importance of accurately predicting demand, Channouf et al. highlight the fact that reliable forecasts serve as essential inputs into EMS planning models and accompanying staffing plans. The data used for their investigation was provided by the Calgary EMS System in Alberta, Canada and spanned a time-period of 50 months from 2000-2004. Consistent with previous studies, each ambulance call record contained the time of the incident, the geographic location where the call was placed from, and the initially assigned call priority [3]. Although available in the data, the call priority, and geographic location information were not used in this study. Rather Channouf et al. focused exclusively on time as the central variable and did not consider the spatial variation of demand. In preparing the data for analysis, the records were aggregated based on the number of calls occurring during each hour of the day. While previous studies, with the exception of Tandberg et al. [15], produced forecasts for broader time periods, such as months or years, Channouf et al. were the first to clearly state that the demand for emergency medical services varies significantly based on the time of the day and the day of the week [3]. Performing preliminary analysis of the data, Channouf et al. plotted the hourly call volume counts over a one-year, one-month, and one-week period to identify

any observable trends and seasonal components present in the data. From the one-year and one-month perspective a clear positive trend and seasonal behavior were visible, with demand reaching peak values during the months of July and December. One would assume these seasonal peaks are related to increased holiday travel, events, and activities common during these months. Channouf et al. attributed the positive upward trend as likely being caused by urban population growth and, as underlined by McConnel and Wilson [8], the advancement of the aging society [3]. Visualizing the hourly call volumes at the one-week time-scale uncovered an oscillation demand pattern mode, with demand reaching its highest values between the hours of 10:00am and 8:00pm Sunday-Thursday and spanning into the late night/early morning on Friday and Saturday [3]. While these results are relatively intuitive based on our understanding of human behavior in urban areas, they further stress the importance of understanding and predicting demand at a more granular level. Diving into methodology, Channouf et al. evaluated five different time series models. Three for forecasting daily call volumes and two for forecasting hourly call volumes. The three models employed to forecast daily calls included (1) a standard regression model, (2) a regression model configured with correlated residuals and (3) a doubly-seasonal ARIMA approach. Given the results, Channouf et al. found that the second model outperformed the other approaches when forecasting 1-2 days into the future with the standard regression model performing comparably at the 14-day mark. The two models used for forecasting hourly call volumes as described by Channouf et al. included (1) a model *"built around the conditional distribution of hourly volumes, conditional on the daily volume"* and (2) a model *"that fits a time-series model directly to the hourly data"* [3]. The authors concluded that both approaches were effective in producing a reasonably accurate short-term (12-24

hrs.) forecast. As previously noted, the major contribution of this investigation was the development of novel models that produced relatively accurate forecasts at daily and hourly time scales.

Brown et al. [7] performed a study to measure the accuracy of forecasts using a popular demand pattern analysis method initially introduced as part of the System Status Management framework. Introduced in 1983 by John Stout, an EMS industry leader, System Status Management (SSM) is an operations methodology currently used by numerous EMS agencies. SSM has also been incorporated into several popular Computer Aided Dispatch (CAD) software systems commonly employed by EMS agencies for dispatching ambulances. Broadly speaking, what Brown et al. refer to as "demand pattern analysis" is an industry approach where historical call data is analyzed to identify typical peak demand times. More specifically, SSM performs demand pattern analysis by using 20 weeks of EMS call volume data to calculate three key hourly forecast values; *"1. Average Peak Demand (AP), 2. Smoothed Average Peak Demand (SAP), and 3. The 90th Percentile for Ranked Demand (90R)"*. Brown et al. cite that prior to their publication no scientific studies had been conducted to evaluate the accuracy of forecasts produced using the System Status Management methodology [7]. To complete their study, Brown et al. collected 73 weeks of hourly call volume data from seven different EMS agencies serving diverse populations throughout the central United States. The first 20 weeks of call volume records, starting on April 12, 2004, served as the historical data set for estimation using the three SSM demand pattern analysis methods; AP, SAP, and 90R. The 21st week of data was omitted, acting as a buffer period, and the last 52 weeks of data were used to test the

accuracy of each method. The call volume forecasts produced by means of the three methods were compared against the actual call volumes for each hourly period over the 52-weeks. Descriptive statistics were developed to measure the accuracy of each forecasting method, including the number of instances where demand was underestimated and overestimated. There results showed that the 90th Percentile for Ranked Demand (90R) forecasting approach performed the best with an approximate 19% estimation accuracy. While the Average Peak Demand and Smoothed Average Peak Demand forecasts resulted in 13% and 10% estimation accuracy respectively [7] . Brown et al. noted that while the estimation accuracy for the exact number of calls predicted was found to be relatively low using these methods, call demand was correctly estimated or overestimated a combined 93-96% of the time. Consequently, only 4-7% of call instances would occur during a period where an insufficient number of ambulances were dispatched based on the forecast estimations. This relates to the sentiment expressed by Baker and Fitzpatrick [14] who pointed out the severity of underestimating demand versus the less concerning miscalculation of overestimation. Given this fact, Brown et al. concluded the SSM demand pattern analysis approach to be a "reasonable predictor for hourly ambulance staffing patterns" resulting mainly in an equitable overestimation of demand. This study represented the first academic exploration into the prolific System Status Management industry practice used for demand planning and analysis.

In 2009, Setzler et al. [5] performed a significantly novel study aimed at producing EMS call volume forecasts at various spatiotemporal granularities using artificial neural networks. The data used to conduct their investigation consisted of emergency calls

dispatched between 2002 and 2004 by MEDIC, an EMS agency responsible for serving the populace of Mecklenburg County, North Carolina. While Channouf et al. were the first to formally state that the demand for ambulances changes significantly based on the time of the day and the day of the week [3], Setzler et al. expanded on this observation by incorporating a spatial component of demand into their forecasting methodology. Individually each call record contained information related to the time and duration of the call, the call priority, and latitude-longitude coordinates identifying the calls location. This enabled the researchers to aggregate the call data at various gradations of time and space. Specifically, they grouped the call volumes into 1-hour and 3-hour time range buckets, and geographically into 2-mile x 2-mile and 4-mile x 4-mile square mile grid blocks; creating a total of four different model configurations at different levels of specificity. It is important to note that the researchers selected these different configurations arbitrarily, and not based on any quantitative measurements or observations. Using the prominent geographic information system (GIS) platform ArcGIS, 2-mile x 2-mile and 4-mile x 4-mile fishnet grid layouts were layered over a map of Mecklenburg Country creating a total of 168 and 40 grid blocks, respectively. Citing that many of the previous studies leveraged traditional casual forecasting and time series approaches, Setzler et al. hypothesized that artificial neural networks (ANN) were a viable alternative as they do not require specific assumptions about the data or error terms, are able to adapt to complex data sets and patterns, and have the ability to learn and model both linear and non-linear relationships [5]. They designed their neural network based on the spatial grid layout where (n) represented the total number of grid blocks, either 168 or 40. They also used four different input features, (H) which represented the hourly time bucket (24 or 8 in total), (S) for the

season of the year (1-4), (D) for the day of the week (1-7), and (M) representing the month of the year (1-12). Each grid block (n) had 4 input nodes into the ANN correlating to each of the four input variables (H, S, D, M). The output from the ANN represented the call volume forecasts for a given time-bucket for each of the (n) grid-block locations, for a total of 168 or 40 forecast values depending on the configuration. To benchmark the performance of their model, Setzler et al. compared the forecasts produced by their ANN against forecasts for each grid-block and time-bucket produced using a method applied by MEDIC and others in the EMS industry, a unique adaptation of a 20-point moving average [5]. Their results showed that the ANN approach moderately outperformed the moving average forecasts at the 4 x 4 mile 1-hour and 3-hour granularity levels. However, at the 2 x 2 mile 1-hour scale the moving average forecasts were more accurate on average over the ANN approach and exhibited no statistically significant difference at the 2 x 2 mile 3-hour level. The authors ultimately concluded that the performance and simplicity of the moving average method currently in use by MEDIC and other EMS agencies suggest no reasonable justification for implementing an ANN for demand forecasting. Despite their results, Setzler et al. [5] were pioneers in exploring the application of ANN, a modern machine learning technique, for forecasting emergency medical services demand.

Chen et al. [16] completed a complementary investigation comparing the performance of ANN, moving average, linear regression, and support vector regression models to predict daily and sub-daily EMS call volumes for three districts within New Taipei City, Taiwan. Closely citing the works of Setzler et al. [5], the researchers further emphasized the importance of incorporating the spatial component of demand for producing accurate

forecasts and effectively allocating EMS resources. In formulating their approach Chen et al. established a framework for practitioners to produce forecasts using a variety of input features and a collection of forecasting models. The authors characterize their framework as being broken up into two distinct components, (1) geographic information system (GIS) and (2) data analytics. The EMS call data is first pre-processed by the GIS which plots the call locations and divides the geographic area spatially into grids defined by the user. The smaller the grids the finer the spatial granularity. The calls are then labeled and grouped together based on grid location. It's important to note that this approach was not at all novel and closely mirrors the methodology originally conceived by Setzler et al. [5]. Chen et al. also used input features similar to Setzler et al. (year, season, month, day, day of the week and time bucket). To differentiate themselves slightly, they added categorical variables for calls occurring on the weekend and during rush hour times. After pre-processing, the data was separated into training, validation and testing datasets and fed into the four different models (ANN, MA, SR, and SVR). When each model was executed, cross validation was performed to identify the optimal combination of input features and model parameters that minimized the root mean squared error (RMSE) of the forecasts [16]. The authors identified this initial execution and cross-validation process as Phase 1. In Phase 2 the model yielding the lowest RMSE within a specific geographic grid-block is selected as the optional model for forecasting calls within that area [16]. Once the forecasts are produced for each respective grid-block, the results are returned to the GIS for spatial visualization. In assessing their results, the authors stated that their models produced acceptable daily call volume forecast accuracy (23.01% lowest MAPE in a single 2km by 2km spatial grid) for each of the three districts in New Taipei City. The forecast error values for sub-daily

(3-hour time bucket) volume estimations were much higher, with significantly lower overall accuracy. However, the primary contribution of their study to the EMS field was not their results. While this "two-phased" model selection approach is not at all unique within the data analytics and machine learning community, it is novel within EMS forecasting research. Chen et al. [16] established a spatiotemporal forecasting framework that enables applying different forecasting models and parameters to different geographic grids based on performance. Furthermore, their framework challenges the idea of a global optima with regards to a forecasting methodology and implements a local optima algorithm search strategy.

Vile et al. [6] completed a study in 2012 in which they produced ambulance demand predictions using a novel time series forecasting technique known as Singular Spectrum Analysis (SSA). SSA serves as a more flexible nonparametric estimation method, over traditional time series approaches, requiring less assumptions regarding the underlying structure and distribution of the data. Referencing previous studies that applied time series methods, Vile et al. hypothesized that SSA could provide more accurate forecasts of EMS call demand. After collecting several years of ambulance call data provided by the Welsh Ambulance Service Trust in Wales, the researchers generated daily call volume forecasts using a combination of SSA, ARIMA and a Holts-Winters (HW) exponential smoothing models [6]. They then compared the overall accuracy of the three methods using the root mean squared error (RMSE) metric. The researchers found that SSA was capable of producing short-term forecasts (7-14 days) with accuracies equivalent to ARIMA and HW. Additionally, SSA generated forecasts with greater accuracy over longer periods of time

(21-24 days). Vile et al. [6] state that in practice SSA is not only a comparable, or in some cases superior, forecast methodology, they argue that SSA is overall a more flexible approach capable of recognizing cycles and patterns in demand. The authors investigation represents the first attempt to apply the novel Singular Spectrum Analysis approach to forecasting EMS demand. An additional layer of detail could be added to future models by incorporating the spatial component of demand. If SSA was capable of producing forecasts of similar accuracies at finer temporal and spatial granularities it could potentially serve as a viable alternative to traditional time-series models in practice.

As previously noted, several researchers have considered including weather related variables in their demand forecast models but ultimately decided to exclude them. The relationships between weather conditions and the effects on human health have been deeply studied throughout various fields [17-20]. Within EMS research, Wong and Lai [21] and McLay et al. [22] performed novel studies analyzing the effects of weather on historical ambulance services demand. In 2014, Wong and Lai [23] carried out a follow-up study focused entirely on developing short-range forecasts using weather factors as causal predictors. By exploring three years of daily ambulance call data provided by the Hong Kong Hospital Authority, Wong and Lai generated daily call demand forecasts using autoregressive integrated moving average (ARIMA) time series models. They started by comparing the performance for three different model configurations. This included a base model using only historical call volume data with no weather variables, a second model that incorporated historic actuals for average temperature, and a third that included historic actuals for average temperature and relative humidity. Following preliminary runs, they

determined that average temperature had the most significant impact on improving the overall forecast accuracy, while the addition of relative humidity increased the mean absolute percent error (MAPE). Given that actual average temperatures are not useful for producing future demand predictions, Wong and Lai replaced the historic actuals from previous periods with forecasted temperatures for future periods. Their final model produced a weeks' worth of daily call demand forecasts with reasonable accuracy using historical call volumes and forecasted average temperatures. The authors cited that one of the primary deficiencies of their work was the fact that not all EMS demand is related to weather [23]. Therefore, the accuracy of their forecasts was reduced by demand that is independent of weather influences. To further improve their model, they recommended making distinctions between the different types of demand such as calls resulting from incidents possibly caused by weather versus other injuries and illnesses. This is commonly referred to in the industry as the dispatch determinate (Ex. traffic accident, overdose, chest pain, breathing problems, etc.). The study conducted by Wong and Lai [23] was the first attempt to incorporate weather related data into forecasting models to improve overall prediction accuracy. Previous researchers postulated the benefits of this inclusion but strayed away from the idea due to the difficulty associated with gathering the data. Given the detailed weather forecast data that is readily available today, collecting the necessary data required for analysis is no longer a significant limitation.

Zhou et al. [24] completed a study in 2014 where they developed a time-varying Gaussian mixture model to predict ambulance demand density. Serving as a probability distribution model, a Gaussian mixture model (GMM) approach assumes that data points are derived

from a collection of Gaussian distributions. GMM is often applied to clustering and density estimation problems and, in some forms, serves as an unsupervised machine learning technique. For EMS forecasting, the authors focused on demand density and used one month of call data provided by Toronto EMS for training their model and two separate months of call data for testing. Demand density was calculated spatially by forecasting call volumes for two-hour intervals within a group of 1km by 1km grid spaces. Those volume values were then normalized by the total demand volume forecasted for a given time period. The forecasted demand density values produced using the Gaussian mixture model were then compared to forecasted densities generated by variations of the moving average (MA) method. The authors found their proposed mixture model outperformed the MA method yielding higher statistical accuracy overall. The study conducted by Zhou et al. [24] provides a promising direction for future EMS demand forecasting research. Given the availability of richer datasets, various probability distribution-based models such as GMM may be capable of producing more generalized estimations of call demand at finer spatiotemporal granularities.

2.2 COMPARISON OF METHODS

In each of the preceding studies, the investigators reviewed the literature that most closely correlated to their hypotheses, research objectives, and methodologies. However, seldom did the authors attempt to categorize the historical approaches based on their unique methodology or usefulness to the emergency medical services field. Channouf et al. [2] suggested that EMS demand forecasting studies can be separated into two distinct groups *" (1) models of the spatial distribution of demand, as a function of demographic variables*

and (2) models of how demand evolves over time" [2]. Using that categorization as a guideline, the earlier works of Aldrich et al. [8], Siler [11], Kvålseth and Deems [10], Kamenetzky et al. [12] and McConnel and Wilson [6] that primarily used socioeconomic, demographic, and land use variables to explain the nature of EMS demand would fall into the first group. While the works of Hall [7], Baker and Fitzpatrick [14], Tandberg et al. [15], Channouf et al. [2], Brown et al. [11], Setzler et al. [5], Vile et at. [6], Wong and Lai [23], Zhou et al. [24], and Chen et al. [16] would be considered members of group two. Table 1 summarizes all the studies surveyed in this review and identifies several important factors to consider when comparing the usefulness of the various models that have been developed for forecasting EMS demand.

Table 1: EMS Demand Forecasting Study Comparison Matrix

Study	Year	Data	Time Scope	Geographic Scope	Dimension	Methodology	Category	Planning Horizon
Hall [8]	1971	Call	Monthly	Detroit, MI	Temporal	Regression	Causal	Tactical
Aldrich et al. [9]	1971	Call & Census	Yearly	Los Angeles, CA	Temporal	Regression	Causal	Strategic
Siler [10]	1975	Call & Census	Yearly	Los Angeles, CA	Temporal	Regression	Causal	Strategic
Kvlseth & Deems [12]	1979	Call & Census	Monthly	Atlanta, GA	Temporal	Regression	Causal	Tactical
Kamenetzky et al. [11]	1982	Call & Census	Yearly	Pennsylvania Counties	Temporal	Regression	Causal	Strategic
Baker & Fitzpatrick [13]	1986	Call	Daily	South Carolina Counties	Temporal	Exponential Smoothing	Time Series	Operational
Tandberg et al. [14]	1998	Call	Hourly	Albuquerque, NM	Temporal	Raw, Moving Average, ARIMA	Time Series	Operational
McConnel & Wilson [7]	1998	Call & Census	Yearly	Dallas, TX	Temporal	Chi-Squared & Tukey's Range	Statistical Testing	Strategic
Channouf et al. [2]	2007	Call	Daily/Hourly	Alberta, Canada	Temporal	Regression & ARIMA	Time Series	Operational
Brown et al. [15]	2007	Call	Hourly	Various (USA)	Temporal	SSM Industry Method	Time Series	Operational
Setzler et al. [5]	2009	Call	Hourly	Mecklenburg County, NC	Spatiotemporal	Artificial Neural Networks (ANN)	Machine Learning/Time Series	Operational
Vile et al. [6]	2012	Call	Daily	Wales, UK	Temporal	Singular Spectrum Analysis	Time Series	Operational
Wong & Lai [23]	2014	Call & Weather	Daily	Hong Kong	Temporal	ARIMA	Time Series	Operational
Zhou et al [24]	2014	Call	Hourly	Toronto, Canada	Spatiotemporal	Gaussian Mixture Model	Probabilistic Model	Operational
Chen et al. [16]	2016	Call	Daily/Sub-Daily	New Taipei City, Taiwan	Spatiotemporal	ANN, MA, Regression, SVR	Machine Learning/Time Series	Operational

The primary data source for each of these studies has been ambulance call records aggregated at various scales of time granularity. In each instance, the data used was restricted to a specific geographic area or, in the case of Kamenetzky et al. [12], Baker and Fitzpatrick [14] and Brown et al. [7], a limited sampling of locations. The recurring use of regression analysis in this problem space has proven to be an effective system for identifying explanatory components of demand, and most of the researchers making use of those techniques yielded comparable results from location to location. However, many of

the researchers also highlighted the fact that the geographic scope of their individual studies served as an underlying limitation of their results as they relate exclusively to the populace of a specific region.

The studies conducted by Baker and Fitzpatrick [14], Tandberg et al. [15], Channouf et al. [2], Brown et al. [11], Setzler et al. [5], Vile et at. [6], Wong and Lai [23], Zhou et al. [24], and Chen et al. [16] represented a significant paradigm shift in research trends, from models aimed at identifying and explaining the causal factors of demand, to models capable of predicting demand with reasonable accuracy. In terms of demand planning, this equates to a transition from models useful for establishing long-term or medium-term strategic and tactical policies based on population characteristics to models that aid in developing short-term operational plans. For EMS dispatchers and managers, a robust understanding of the nature of demand over longer periods of time (i.e. monthly or yearly) supports their decisions related to ambulance fleet size and system capacity; while short-range forecasts assist with producing weekly, daily and hourly workforce schedules and deployment plans. For EMS agencies, the need for accurate short-term forecasting methods that can be incorporated into their daily dispatching processes are critical to the delivery of responsive patient care. Brown et al. [7] underscored this point by evaluating the effectiveness of the current SSM industry practice. While they found the demand pattern analysis forecasting methods associated with SSM to be adequate at estimated or overestimating demand, they failed to consider the allocation and distribution of resources dispatched spatially. The work completed by Baker and Fitzpatrick [14] set the course for much of the future research by shifting the focus away from casual regression models that analyzed demand over longer

periods to time-series based forecasting models capable of producing daily predictions with greater accuracy. Additionally, Baker and Fitzpatrick [14] emphasizes the importance of developing practical models that could be implemented by EMS agencies and staff. The works of Tandberg et al. [15] and Channouf et al. [2] progressed the field further by exploring the effectiveness of relatively simple time series models for producing hourly call volume forecast.

The various temporal models developed by researchers have improved our ability to predict operational volume at the daily or hourly level for entire EMS service areas. However, accurate estimations that incorporate the spatial distribution of demand are vital to determining ambulance dispatch locations and establishing fluid redeployment plans based on spatial demand fluctuations [5, 24]. Concentrating on varying the levels of forecast granularity, Setzler et al. [5] were the first to add a spatial component of demand in their forecasting methodology. Ultimately their novel spatiotemporal approach set the trajectory for all future research efforts. The use of machine learning models, such as artificial neural networks, overcomes the dimension complexity limitation discussed by Aldrich et al. [10] when attempting to produce call forecasts values for an array of geographic locations. Chen et al. [16] further reinforced this idea by developing a framework for generating spatiotemporal forecasts using a combination of artificial neural networks, support vector regression and traditional time-series approaches. One of the unique challenges associated with predicting ambulance demand in both time and space is the sparse nature of demand instances at finer granularities. This results in a complex, non-normal, zero-inflated data distribution that is especially difficult for traditional regression and time-series approaches

to model accurately. Both Setzler et al. [5] and Chen et al. [16] encountered this challenge while evaluating the results of their forecast estimations. At finer granularities, fewer call instances exist at each time interval and grid location for the model to train against. With a larger number of zero demand (no call) instances occurring in a given geographic area, the resulting models consequently produced zero call volume forecasts for the majority of locations and time intervals. Given the number of dimension combinations with zero actuals, this also results in artificially low error values when the forecasts are compared against the actuals.

2.3 ALTERNATIVE MODELS AND METHODS

ANNs have been presented as effective alternatives to traditional time series forecasting methods across a wide range of applications. In 1993, Tang and Fishwick [25] compared the performance of neural networks against the Box-Jenkins method across sixteen time series datasets of varying complexity. Their results indicated that for long-term forecasts, neural networks consistently outperformed the Box-Jenkins model. Similarly, in 1996, Hill et al. [26] produced time series forecasts using neural networks and six prominent time series methods. They found that across monthly and quarterly time series, neural networks significantly outperformed the traditional methods. Since the publication of those early studies, the applications and production implementations of ANNs have dramatically evolved and expanded. Exploring the application of more complex models such as deep neural networks has the potential to significantly advance EMS call volume estimation capabilities at finer spatiotemporal granularities.

Data mining techniques, such as cluster analysis, could also be applied to improve forecasters understanding of demand distributions and aid in model feature selection and grouping. Prospective models must also continue to incorporate detailed spatial demand information when formulating predictions. More recent investigations such as those performed by Vile et al. [6] and Wong and Lai [23] that continue to focus on time as the sole dimension are comparatively antiquated. Aringhieri et al. [3] suggested that *"based on the spatial mode, data mining might allow the development of a prediction system for emergency demand, i.e., to identify the most likely region from where the next emergency request could arrive"*. This may also require integrating those socioeconomic, demographic, and land use attributes identified in previous studies as causal variables. Additional features such as traffic patterns, weather, events, and the dynamics of daily population shifts could also improve the richness and accuracy of future models [5, 16]. Probability distribution based models, such as those explored by Zhou et al. [24], as well as various Bayesian learning methodologies are also encouraging approaches for producing meaningful spatial estimations of demand. Future researchers might also consider incorporating features related to individual emergency call instances, such as the dispatch determinate and assigned call priority to add an additional layer of meaning to their projections. Lastly, it is essential that forthcoming investigations overcome the challenges associated with the zero-inflated demand distributions that are created at various spatiotemporal granularities.

CHAPTER 3: DATA ANALYSIS

3.1 DATA SOURCE

The data used in this study was provided by MEDIC, a private EMS Agency responsible for serving Mecklenburg County North Carolina, and contains all 911 emergency medical call occurrences recorded from January 1st, 2010 to December 31st, 2017. Each call record contains the call date/time, call location coordinates, responding ambulance identifier, assigned call priority, patient problem description, call response outcome (transferred, refused care, false call, etc.), and additional time stamps such as the time in route and arrival time. Emergency medical calls (or incidents) are uniquely identified using a field labeled "master incident number". While there is a one-to-one relationship between an individual call received by MEDIC and a single master incident number, the same master incident number may be logged for multiple records. This occurs in cases where multiple ambulances respond to a single call, or in situations where one or many ambulances serve multiple patients on a single call. For the purposes of this study, duplicate master incident number records will be removed from the dataset to provide a more accurate representation of true call demand instances and to remain consistent with the approaches of previous call demand forecasting studies.

3.2 DATA STATISTICS AND PRE-PROCESSING

The original dataset contained 907,883 records, representing every call MEDIC responded to over an eight-year period (January 1st, 2010 to December 31st, 2017). A total of 46,804 records with duplicate master incident numbers were removed using a custom Python script, leaving 861,079 unique call records. Meaning, over an eight-year period only 5.16%

of all calls responded to by MEDIC required multiple ambulances and/or served multiple patients on a single call. A more detailed analysis of these relatively infrequent calls could be conducted using the same dataset in a future study. Figure 1 displays the total call volume by year. Total call volume is defined as the total number of EMS calls received in a single period (i.e. year, month, week, day, hour, etc.).

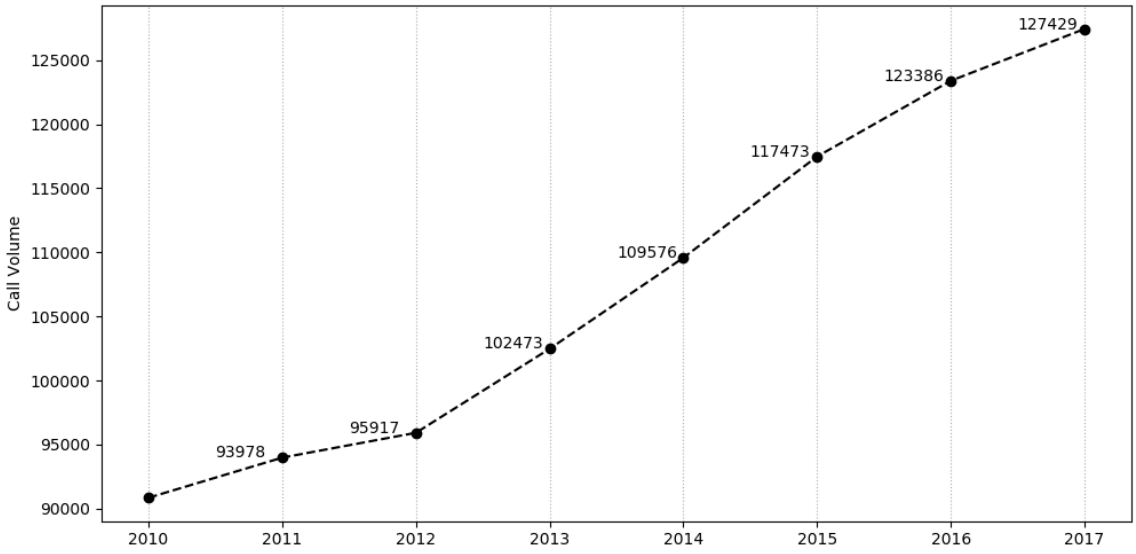

Figure 1: Total Call Volume by Year in Mecklenburg County, NC

As illustrated in Figure 1, the total call volume annually in Mecklenburg County has increased each year with an average increase of 4.97% over eight years. Mecklenburg County has also experienced a relatively proportional increase in the total population as shown in Table 2 [27]. Performing a simple linear regression between annual total call volume and the estimated population in Mecklenburg County each year yielded an adjusted R^2 value of 97.02%. While relative changes in population have little impact on daily and hourly demand, it's important to underscore this relationship as it contributes to the long-term trend of call volumes in the region.

Table 2: Annual Call Volume and Mecklenburg County Population Estimates

Year	Total Call Volume	Volume Percent Increase	Mecklenburg County Pop. Est.	Pop. Percent Increase
2010	90,847	-	923,326	-
2011	93,978	3.45%	945,113	2.36%
2012	95,917	2.06%	968,367	2.46%
2013	102,473	6.84%	991,322	2.37%
2014	109,576	6.93%	1,010,878	1.97%
2015	117,473	7.21%	1,033,466	2.23%
2016	123,386	5.03%	1,054,835	2.07%
2017	127,429	3.28%	1,076,837	2.09%

To produce accurate short-term spatiotemporal call volume forecasts the call date, time, location, and call priority are all critically important. As previously noted, the primary goal of EMS agencies is to minimize their response times to high priority calls [1]. Private agencies, such as MEDIC, contracted by local counties and municipalities are bound to specific response time service requirements based on assigned call priorities. EMS communication (911) operators fielding incoming call requests perform an initial patient triage by asking callers a series of structured questions to quickly determine patient severity and assign a call priority. Clinically, calls are coded using a series of phonetic response code words (Echo, Delta, Charlie, Bravo, Alpha, Omega) that distinguish between different patient severities and conditions. At the dispatch level these codes fall into three different dispatch priority categories; (P1) Priority 1: life-threatening, (P2) Priority 2: potentially life-threatening, and (P3) Priority 3: non-life threatening. For instance, if a patient is described as experiencing abdominal pain, or a headache, the call may be coded as a Charlie P2, i.e. potentially life threatening. While a patient presenting breathing problems, and/or sharp chest pain, may be coded as an Echo P1, i.e. life-threatening symptoms. To remain compliant with their service-level agreement in Mecklenburg County, MEDIC must respond to at least 90% of all P1 calls within 10 minutes 59 seconds, all P2 calls

within 12 minutes 59 seconds, and all P3 calls within 20 minutes. This response and dispatch priority information is summarized below in Table 3.

Table 3: MEDIC EMS Call Response Code & Priority Matrix

Dispatch Priority	Response Code Number	Response Code Word	Response Time Compliance
P1	9	Echo	10 min 59 sec
P1	1	Delta	10 min 59 sec
P2	2	Charlie	12 min 59 sec
P2	7	Bravo Hot	12 min 59 sec
P3	3	Bravo	20 min
P3	8	Bravo Cold	20 min
P3	4	Alpha	20 min

Every year MEDIC consistently meets, and typically exceeds, their contractual service-level response time requirements of 90% for all priority types. Between 2010-2017, MEDIC responded to an average of 96.98% of P1 calls, 97.76% of P2 calls, and 92.50% of P3 calls within response time compliance limits. They accomplished this milestone despite the fact that from 2010 to 2017 MEDIC's total transport volume increased by 47%, which is greater than three times the rate of Mecklenburg County's population growth [28]. While providing superior performance, MEDIC continually strives to reduce their response times to high priority calls. For the purposes of short-term (daily and hourly) demand planning, MEDIC indicated that call volume forecasts should focus exclusively on P1 and P2 calls. P1+P2 calls represent the most critical patients, require the fastest response times, and constitute the majority of all calls received (73.57%). Calls that do not require an immediate response are designated as non-emergent transport (NET), such as requests to transfer patients to and from clinical care facilities. NET calls also include pre-scheduled patient transports. A breakdown of the dispatch priority percentages for all call records are outlined in Table 4. Given the objective of concentrating solely on producing high priority

call predictions, 227,555 records representing all P3, NET, and other dispatch priority calls were dropped from the dataset. An additional 110 records that contained missing or incomplete fields were also dropped, leaving a total of 633,417 records in the final dataset.

Table 4: Dispatch Priority Call Percentages

Dispatch Priority	Description	# of Calls	% of Calls
P1	Life Threatening	160,912	18.69%
P2	Potentially Life Threatening	472,615	54.89%
P3	Non-Life Threatening	90,561	10.52%
NET/Other	Non-Emergent Transport	136,994	15.91%

3.3 TEMPORAL FEATURE DECOMPOSITION AND ANALYSIS

As previously noted, each call record in the dataset contains a variety of time stamps representing different events that occur throughout the call response process. Individual timestamps document the year, month, day, hour, minute and second of a given event. The first timestamp in each record is labeled “Response Date” and records the date and time of the incoming call request (i.e. demand occurrence). To conduct a detailed analysis and prepare the dataset for input into various forecasting models, the response date timestamp needs to be decomposed into a collective feature set. Using a custom Python script and a coding structure similar to Setzler et al. [5] and Chen et al. [16], the year, season, month, week, day, day of the week, and hour were extracted and labeled following the schema outlined in Table 5. Since forecasts will be generated at the daily, sub-daily, and hourly level, the recorded minutes and seconds values were ignored. Plotting the monthly call volume data for all P1 and P2 calls reveals an increasing trend and a distinct monthly seasonal pattern; as shown in Figures 2, and 3.

Table 5: Temporal Feature Labeling Schema

Feature	Coding Label
Year	2010 - 17
Season	1 - 4
Month	1 - 12 (Jan - Dec)
Week	1 - 52
Day	1 - 31
Day of Week	1 - 7 (Mon - Sun)
Hour	0 - 23

Recognizing that the drop in total monthly demand each February may relate to the fact that it is the shortest month of the year, the average daily volume for each month was calculate and compared from month to month and year to year. Between 2010 and 2017, February only had the lowest average daily demand in 2013 and 2014. In all other years, the lowest average daily demand occurred in December or January.

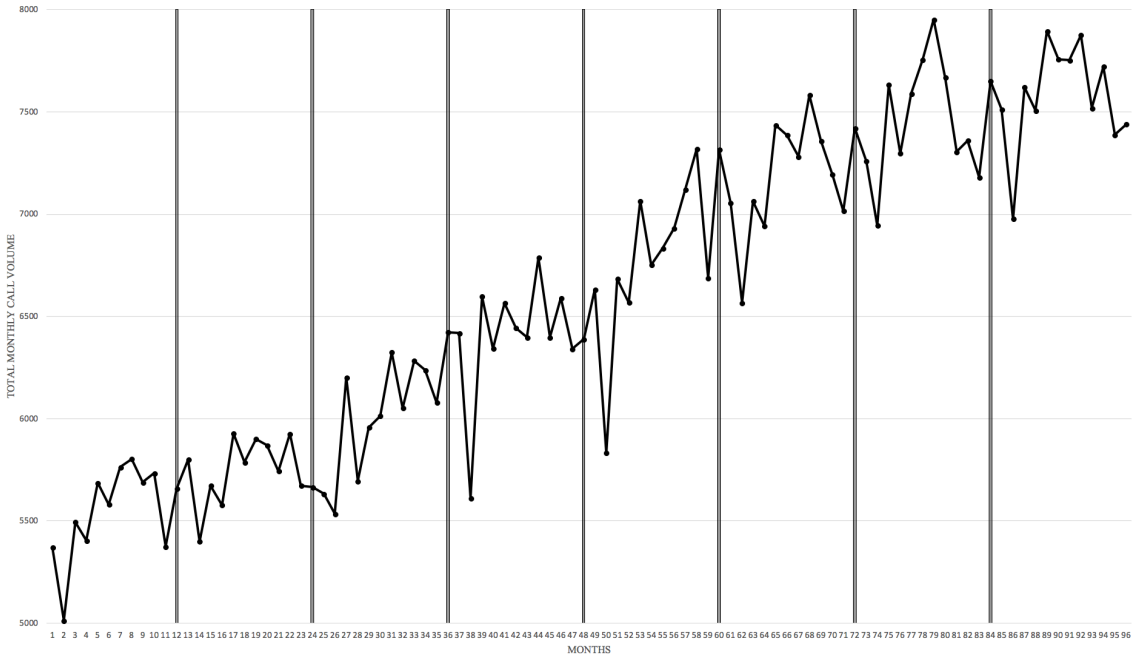

Figure 2: Total P1+P2 Call Volume by Month (2010 – 2017)

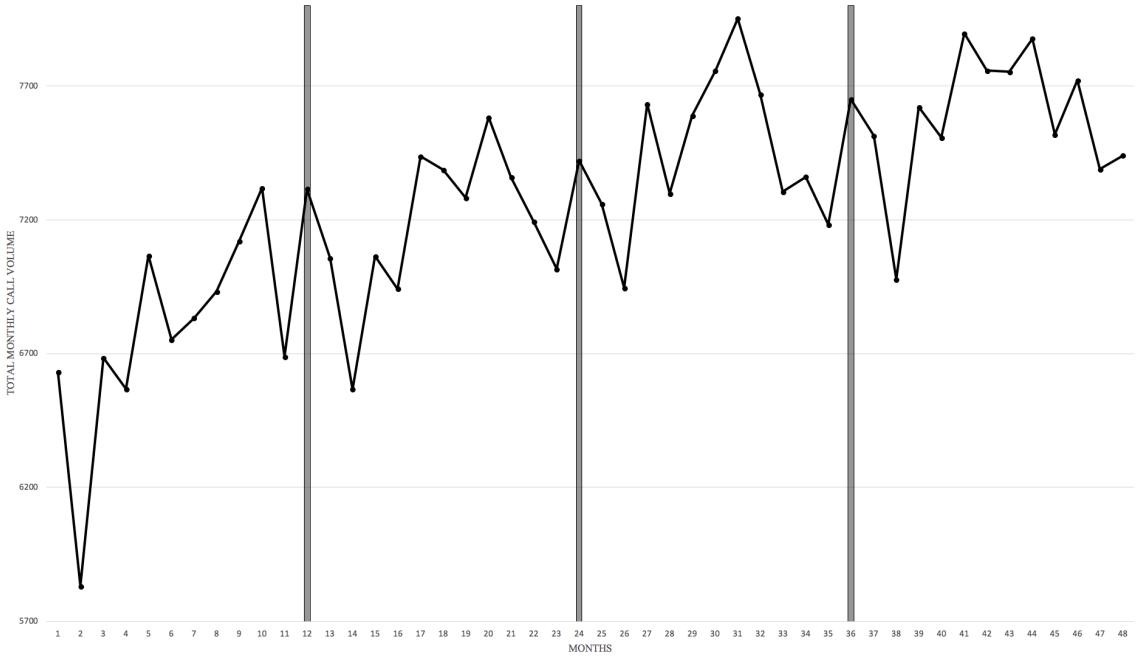

Figure 3: Total P1+P2 Call Volume by Month (2014 – 2017)

To prepare the data for further analysis and input into various forecasting models, two separate datasets were generated to represent the daily and hourly call volumes. Using custom Python scripts, the daily and hourly call volumes were aggregated separately by counting the number of call instances that occurred on individual days and within individual hours. In the hourly call volume data, a total of 199 instances of hours with no calls (i.e. zero call volume) were identified. To ensure a complete and accurate time series, zero call volume records for the missing 199-hour instances were generated using custom Python scripts. At the daily scale, a season component is present for the days of the week; as illustrated in Figures 4 and 5. Sunday is consistently the one day of the week with the lowest total call volume each year, while Saturday typically experiences the second lowest call frequency. Furthermore, totals are consistently higher during the week (Monday-Friday) reaching peak demand on either Monday, Tuesday or Friday. Anecdotally, this can be attributed to increased activity throughout the county as people travel to and from work,

school, and carry out routine weekly activities. Drilling down to the hourly level, the distinction between weekdays (Monday-Friday) and weekend days (Sunday/Saturday) is more pronounced. Figure 5 shows the average hourly call volume by the day of the week for all 2010-2017 call data. A clear oscillation pattern is visible at the hourly scale as call volumes decline in the early morning hours of the day to their lowest point at approximately 06:00 EST. Call volumes increase significantly between the hours of 06:00 and 12:00, remaining at near peak levels for eight to ten hours each day.

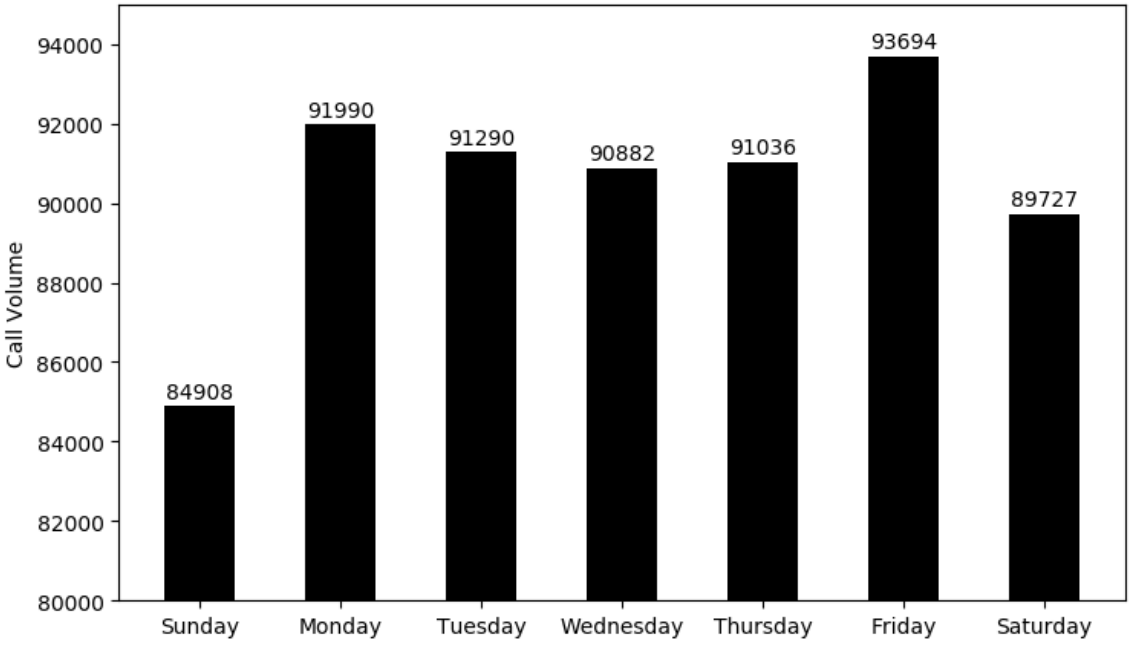

Figure 4: Total P1+P2 Call Volume by Day of Week (2010 – 2017)

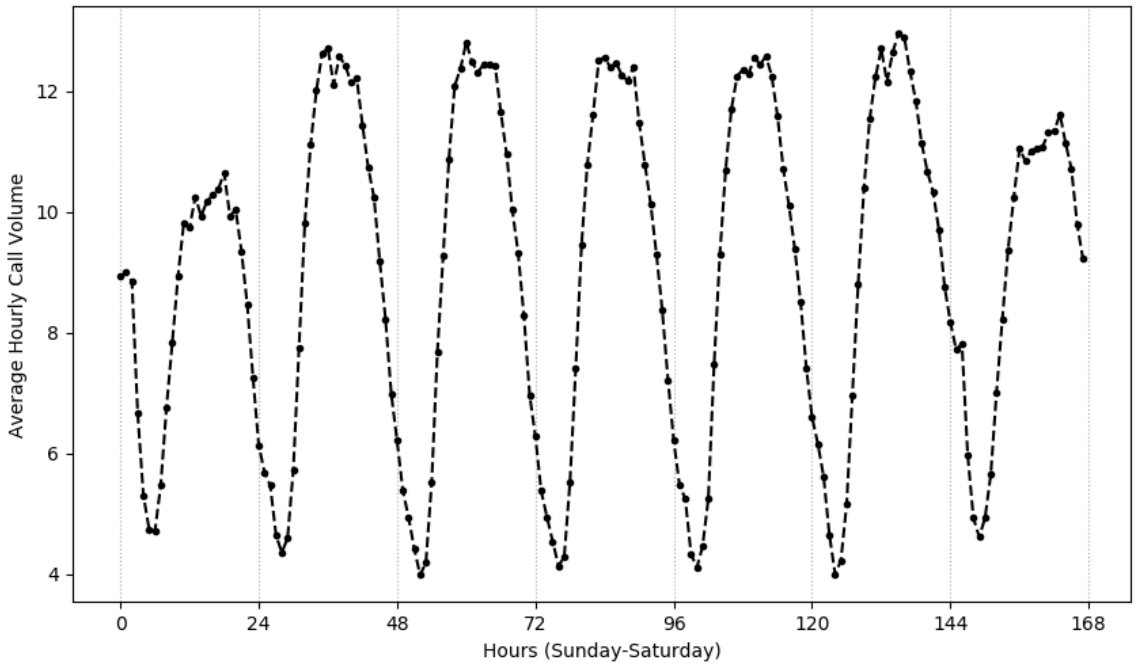

Figure 5: Average P1+P2 Call Volume by Day of Week (2010 – 2017)

While call volumes are overall consistently higher during the weekdays, volumes remain closer to their daily peak volume levels for more hours during the afternoons of the weekend days compared to the weekdays. The patterns and trends present in the MEDIC dataset are consistent with the behavior of a smaller dataset collected in Calgary, Alberta and used in the study conducted by Channouf et al. [3]. Their dataset also exhibited a clear long-term positive linear trend, monthly and daily seasonality. In the subsequent chapters, daily and hourly estimations are generated using machine learning models and a collection of traditional time-series methods which serve as benchmarks comparable to methods applied in previous studies. Hourly forecasts are also produced at finer spatial granularities to compare their performance against the capabilities of more advanced techniques and a range of machine learning models.

CHAPTER 4: DAILY CALL VOLUME FORECASTS

In this chapter, an assortment of time-series methods and a multi-layer perceptron (MLP) model are used to produce non-spatially distributed daily call volume estimations using the MEDIC dataset. An MLP model, loosely based on the original implementations by Setzler et al. [5] and Chen et al. [16], is developed over multiple iterations. An expanded version of the MLP developed in this chapter is later used in the forthcoming hourly forecasting chapters to produce and compare non-spatially distributed and spatiotemporal hourly call volume estimations.

4.1 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE METHOD

Between 2010 to 2017, the participating EMS agency received an average of 217 calls per day and approximately 9 calls per hour. When scaled down to the daily and hourly level, EMS call volume levels tend to be stationary and repetitive in nature. As such, the use of traditional time series forecasting techniques such as moving average models, variations of exponential smoothing models, and autoregressive integrated moving average (ARIMA) models have historically been capable of producing daily call volume predictions for broad geographic areas with reasonable accuracy [3, 15, 29]. ARIMA is one of the most widely used models for time-series analysis/forecasting and can be applied to both seasonal and non-seasonal datasets. A non-seasonal ARIMA model contains three fundamental parameters (p, d, q); where p represents the number of autoregressive (AR) terms, d serves as the number of non-seasonal differences, and q denotes the number of moving average terms. In a seasonal ARIMA model an additional set of (p, d, q) parameters are included to

account for the seasonal component of demand along with an (m) parameter representing the number periods in each season.

Prior to generating forecasts using $ARIMA_{(p,d,q)(p',d',q',m)}$, or some of the other traditional time-series techniques, it is important to first determine if a time series dataset is stationary (i.e. constant mean, variance, and covariance overtime). Stationarity is a fundamental pre-requisite for ARIMA to produce unbiased estimates. One visual technique that can be used to initially evaluate the stationary nature of a dataset is to plot the average and variance of observation values overtime. Figure 6 displays the daily call volume values using records from 2010-2017 plotted along with rolling 365-day mean and standard deviation values.

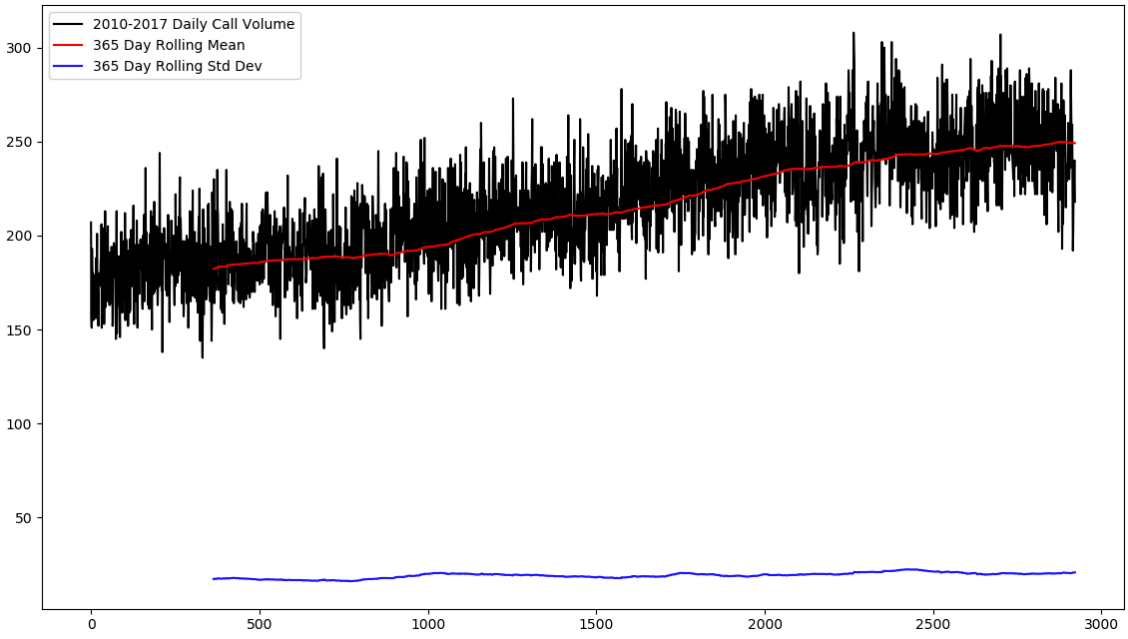

Figure 6: P1+P2 Daily Call Volumes Rolling Mean and Std. Dev. (2010-2017)

As previously identified in the data analysis chapter, the daily call volume values exhibit a long-term positive trend as illustrated by the increasing rolling mean, while fluctuations in the standard deviation values appear to be relatively minor. When a dataset is found to be

non-stationary as a result of a changing mean, the value of the differencing (d) parameter in a subsequent ARIMA model will be set to a positive integer value greater than zero. First or second order differencing is typically sufficient to transform a dataset with a variable mean from non-stationary to stationary [30].

To further evaluate the stationary nature of the time series, a common statistical technique known as the Augmented Dickey-Fuller (ADF) test is applied to the 2010-2017 daily dataset [30]. The null hypothesis in this test is that the time series can be forecasted by a unit root (time-dependent structure). The alternate hypothesis is that the time series cannot be represented by a unit root and therefore is stationary. With a resulting p-value of 0.23, well above the 0.05 critical value threshold, the dataset is found to be non-stationary. Therefore, prior to deploying ARIMA, the sources(s) of non-stationarity (i.e. trend in mean, variance, or covariance) need to be determined and the original non-stationary time series requires transformation into a stationary form. As illustrated in Figure 1, call volume experiences an upward trend during the 2010-2017 period. This underlying trend in mean is at least one source for non-stationarity. Hence, first-order differencing is applied to the time series and the ADF test is repeated. The test following differencing resulted in a practically zero p-value (5.0E-29), well below the 0.05 critical value threshold. This confirms that the source of non-stationarity was the trend in mean and that differencing is a sufficient step in transforming to a stationary time series. Since ARIMA carries out the differencing step innately, the original time series is used as input to ARIMA with the value of parameter d in $ARIMA_{(p,d,q)(p',d',q',m)}$ set to 1. To determine the optimal values for the parameters p and q, in the ARIMA model, an R implementation of the Auto-ARIMA

function is utilized [31]. Auto-ARIMA performs a series of hyper-parameter tuning searches to identify the best ARIMA model based on two key metrics, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both AIC and BIC are information criteria methods used to determine the model of best fit while also penalizing model complexity. Models yielding the lowest AIC/BIC values are preferred. For daily forecasts, the seasonality parameter (m) is set to seven, representing the days of the week. Using the daily call volume records from 2010-2016 as the training dataset, Auto-ARIMA determined the optimal model configuration to be a seasonal $ARIMA_{(3,1,3)(2,0,2,7)}$.

4.2 THE MEDIC DAILY FORECASTING METHOD

The performance of each daily forecasting model examined in this chapter is compared against a benchmark method presently used by practitioners at MEDIC. This benchmark method, identified in this study as the "MEDIC Daily Forecasting" (MDF) method, calculates daily call volume estimations. The equation for the MDF method is defined as:

$$F_{d,m} = \bar{A}_{d,m-12}\left[1 + \frac{\sum_{m=-1}^{-12} A_m - \sum_{m=-12}^{-23} A_m}{\sum_{m=-12}^{-23} A_m}\right] \quad (1)$$

Where the following notation applies:

F = forecasted call volume
A = actual call volume
d = day of the week (Mon, Tue, Wed, Thu, Fri, Sat, Sun)
m = month

The MDF equation acts as a trend smoothing function, scaling the average actual call volume of a specific day of the week (d) from month $(m-12)$ in the prior year by the total volume percent change over the previous 12 months. Each day of the week in an

individual month is weighted equally. The calculated (F_{dm}) value serves as the forecasted call volume for each instance of that (d) day of the week in a month (m). Representatives at MEDIC stated that historically the MDF method has resulted in a mean absolute percent error (MAPE) value of approximately 6%-15% over an annual planning period [32]. MAPE is an error metric commonly used by academics and EMS practitioners for evaluating the performance of various forecasting models. As such, MAPE will serve as the primary error indicator throughout this study. The equation for MAPE is defined as:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t} \quad (2)$$

where the following notation applies:

F = forecasted call volume
A = actual call volume
t = time period
n = number of periods

4.3 DAY OF WEEK MOVING AVERAGE METHOD

Numerous studies have examined the performance of moving average (MA) methods using various configurations and number of observation periods to generate daily and hourly call volume predictions. While this approach is rather naïve, MA has proven to be an effective method for time-series forecasting given stationary observations. As examined in the previous chapter, and other related studies, significant seasonal patterns exist within EMS call data in relation to the day of the week. Consequently, researchers and industry practitioners have developed a wide variety of time-series forecasting techniques, such as the MDF method, that formulate predictions using call volumes for specific days of the week [5, 7, 16]. To evaluate the performance of a similar method, the following moving average equation is proposed:

$$F_{d,w} = \frac{A_{d,w-1} + A_{d,w-2} + A_{d,w-3} + \cdots + A_{d,w-n}}{n} \quad (3)$$

where the following notation applies:

F = forecasted call volume
A = actual call volume
d = day of the week (Mon, Tue, Wed, Thu, Fri, Sat, Sun)
w = week
n = number of periods

This variation of the moving average technique branded the "Day of Week Moving Average" (DOWMA) method, calculates the daily call volume for a given day of the week using a collection of actuals representing call volume totals from the same day of the week in (n) previous weeks. The number of periods (n) are adjusted or determined with the goal of minimize the MAPE value.

4.4 HOLT-WINTERS TRIPLE EXPONENTIAL SMOOTHING METHOD

The Holt-Winters Triple Exponential Smoothing method is a prominent forecasting technique that is effective at generating estimations given time series data exhibiting both trend and seasonality. Baker and Fitzpatrick [14] first used a variation of the Holt-Winters method, along with a parameter optimization model, to generate daily EMS call volume forecasts. To compare the performance of this method against the MDF benchmark method, and other approaches, the following multiplicate seasonal effects version of the Holt-Winters model is selected:

$$F_{t+n} = (E_t + nT_t)\, S_{t+n-p} \quad (4.1)$$

where:

$$E_t = \alpha\left(\frac{A_t}{S_{t-p}}\right) + (1-\alpha)(E_{t-1} + T_{t-1}) \quad (4.2)$$

$$T_t = \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1} \quad (4.3)$$

$$S_t = \gamma\left(\frac{A_t}{E_t}\right) + (1-\gamma)S_{t-p} \quad (4.4)$$

and the following notation applies:

F = forecasted call volume
A = actual call volume
E = expected base level
T = estimated trend value
S = estimated seasonal factor
p = the number of seasons in the time series (i.e. p=7 for weekly data)
n = time periods into the future
t = current time period
α, β, γ = smoothing parameters where $(0 \leq \alpha \leq 1,\ 0 \leq \beta \leq 1,\ 0 \leq \gamma \leq 1)$

The values of the smoothing parameters, notated as $\alpha, \beta, and\ \gamma$, in the model are optimized using the evolutionary search algorithm available in the 2019 version of Microsoft's Excel Solver platform. The objective function is set to minimize the MAPE of 2017 estimations while maintaining parameter values between 0 and 1. The optimal parameter values for daily predictions were determined to be 0.0351, 0.0100, and 0.0286 for $\alpha, \beta, and\ \gamma$ respectively.

4.5 ARTIFICIAL NEURAL NETWORK METHOD

In the two previous studies that explored ANNs as an approach for generating call volume estimations, both Setzler et al. and Chen et al. used a popular class of feed-forward ANNs known as a multi-layer perceptron model. MLP is categorized as a supervised machine learning algorithm that uses a technique known as backpropagation to train an estimation model given a training dataset containing *(x)* number of input features [5, 16]. MLP models

can be applied to both classification (discrete) and regression (continuous) problems. The structure of an MLP network consists of a series of layers, each containing a collection of nodes (neurons). Represented mathematically as a directed graph, the individual nodes at each layer are fully connected to the nodes of the following layer. Each network contains a single input layer, an output layer, and one or more hidden layers for processing inputs. An illustration of a generic MLP model with a single hidden layer is shown in Figure 7.

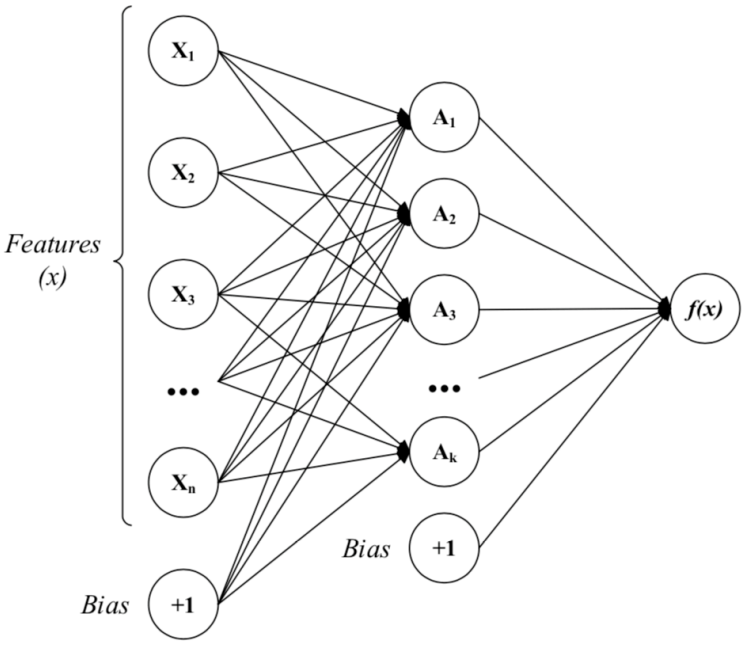

Figure 7: Single Hidden Layer MLP Model

The number of nodes in an input layer typically correlates with the number of features and frequently contains an additional bias node. While the number of hidden nodes and layers in a network is less derivative and depends on a variety of factors including the number of input features, the level of noise in the training data, the amount of training patterns, the activation function being applied, and the complexity of the function or classification task to be learned. In formulating their MLP models, Setzler et al. and Chen et al. referenced a prominent study by Zhang et al. [33] that reviewed best practices for forecasting with

artificial neural networks. In their survey, Zhang et al. discuss historical findings and techniques for determining the number of hidden layers and nodes in an ANN. They concluded that, for the majority of forecasting problems, one to two hidden layers are appropriate for producing accurate results. When determining the optimal number of hidden nodes, Zhang et al. noted that networks with fewer hidden nodes are typically preferred as they produce more generalized models, reduce the likelihood of overfitting the model based on the training data, and have lower complexity. As a general heuristic, the number of hidden nodes in a layer is typically initialized based on the approximate number of input features and then altered through a process of trail-and-error that seeks to minimize the models testing error. Hyperparameter tuning approaches such as grid-search and random-search are also effective strategies that can be applied to identify near-optimal configurations for the number of hidden layers and nodes.

4.6 DAILY FEATURE SELECTION

For any predictive model to perform well, the proper input data (predictors/features) need to be identified. Feature selection, also referred to as attribute or variable selection, is the process of systematically selecting a subset of attributes from a dataset that are determined to be the most relevant to a prediction problem. Hence, feature selection serves as an important pre-processing step for developing effective learning models and removing noisy data. Reducing the total number of features included in an estimation model has several advantages. Learning algorithms have a tendency to experience decreased prediction accuracy when the number of input features exceed an optimal level [34]. Additionally, removing variables that are determined to be statistically irrelevant, redundant, and/or

confounding reduces overall model complexity and the amount of computing time required. While feature selection can help alleviate data dimensionality issues commonly encountered in machine learning problems, feature selection itself is not dimensionality reduction. Both feature selection and dimensionality reduction techniques seek to reduce the total number of attributes present in a dataset. Dimensionality reduction accomplishes this task by creating new combinations of attributes, whereas feature selection methods aim to include or exclude attributes based on their original values and relevancy. Serving as an important pre-processing step in the applied machine learning process, feature selection can be accomplished using a variety of techniques and methods. One of the most common techniques is univariate feature selection, where the best input features are identified based on various univariate statistical tests such as chi-squared and F-tests. While these statistical techniques are simple to compute and compare, they are univariate in nature and are predominantly useful in situations where the input features are entirely independent from one another. Feature selection can also be achieved using multivariate models, such as support vector machines, that recursively eliminate individual input features and measure the impacts on estimation accuracy. While this approach can be very effective in application, the major limitation of multivariate (recursive) feature elimination is a rather arbitrary feature exclusion tactic that concentrates on maximizing the accuracy performance of a specific estimator while potentially discarding relevant and important input features.

In this study, a novel feature selection method known as Boruta is applied to identify all relevant features in EMS call volume datasets [35]. The literature [5, 16] previously

identified several temporal features (year, season, month, week, day, day of week) as candidate features for forecasting daily call volumes. In addition to these temporal features, weather attribute data collected from the National Oceanic and Atmospheric Administration (NOAA) including the historic average daily temperature, minimum daily temperature, maximum daily temperature, precipitation in inches, and binary variables indicating occurrences of daily rain and snow fall were added to the candidate feature set [23]. Lastly, based on the structure of the seasonal $ARIMA_{(3,1,3)(2,0,2,7)}$ model, distributed lag values of periods 1, 2, 3, 7 and 14 for the daily features (day, day of the week, daily avg/max/min temperatures, precipitation, rain, snow, and volume) were added to the candidate feature set. The inclusion of the lag variables for call volume in the MLP model results in rolling window predictions which are comparable to time-series forecasting methods (e.g. ARIMA, exponential smoothing, moving average).

After identifying all the candidate features for the daily predictions, Boruta (implemented in R) was used to determine which final features should be included in the model [35]. Boruta functions as a wrapper method leveraging ensemble-based learning algorithms, specifically variations of decision-tree ensembles. For each candidate feature, Boruta creates a *shadow* feature by shuffling the values of the original feature. For instance, if the value of feature *maxTemp* for day 1 is equal to 78 F and day 2 is equal to 83 F, the values of the *maxTemp* shadow feature are created by a randomly shuffling of 78 F and 83 F. This shadow feature maintains the same mean and standard deviation of the original feature, yet it should not have any significant correlation with the original variable or target variable (as it has been created by a random process). After creating shadow features for each

original feature, Boruta compares the predictive power of each original feature against the shadow features. If any of the original features have a lower predictive power compared to their corresponding shadow feature, the algorithm labels those features as “unimportant”. If the predictive power of an original feature is higher than that of its shadow feature, the algorithm will label it as “important”.

The relative importance (rank) of each feature is also returned, identifying which features have a greater importance over the top performing shadow features (random chance). This ranking information is captured by the algorithm as “hits” as additional iterations are performed. The process of recording “hits” is cyclical, with the algorithm examining if a given input feature performs better than random chance. The algorithm accomplishes this by comparing the number of times a feature performed better than its shadow feature using a binomial distribution. If a feature is determined to be important after several continuous hits over multiple iterations it is confirmed, removed from the dataset, and the next iteration is performed using the reduced dataset. If a feature fails to receive any hits after several iterations are conducted, it is rejected and removed from the dataset. Features that are neither confirmed nor rejected are kept as tentative and remain in additional iterations. After all of the features have been identified as either confirmed or rejected, or a pre-determined iteration limit is reached, the algorithm stops and returns the ranking results. The final output from the Boruta algorithm is a list of “important” ranked input features for use in subsequent learning models. After running Boruta using a random forest regressor, the features day, precipitation, rain, snow, and their corresponding lag variables

were rejected. The remaining candidate features were confirmed to be important for estimating daily EMS call volumes, ranked results are presented in Table 6.

Table 6: Daily Call Volume Prediction Features Ranked Results

Features	Mean Importance	Median Importance	Minimum Importance	Maximum Importance	Norm Hits	Decision
year	30.10	30.11	27.94	32.16	1.00	Confirmed
volumeLag1	25.12	25.31	22.61	27.98	1.00	Confirmed
volumeLag7	23.30	23.18	21.70	25.33	1.00	Confirmed
volumeLag3	23.27	23.33	21.04	24.70	1.00	Confirmed
volumeLag2	23.10	22.92	21.13	25.14	1.00	Confirmed
volumeLag14	21.87	21.85	19.85	23.70	1.00	Confirmed
dayOfWeekNumLag7	14.60	14.70	12.20	16.72	1.00	Confirmed
dayOfWeekNumLag14	14.56	14.76	12.26	15.72	1.00	Confirmed
dayOfWeekNum	14.48	14.35	12.60	15.69	1.00	Confirmed
maxTemp	12.57	12.47	7.89	16.44	1.00	Confirmed
avgTemp	11.78	11.75	8.36	14.02	1.00	Confirmed
minTemp	10.60	10.96	7.72	12.75	1.00	Confirmed
maxTempLag1	8.74	8.71	5.19	11.62	1.00	Confirmed
minTempLag1	8.45	8.59	5.65	10.67	1.00	Confirmed
maxTempLag14	8.36	8.16	6.78	10.53	1.00	Confirmed
maxTempLag3	8.35	8.30	6.07	10.31	1.00	Confirmed
weekNumber	8.30	8.44	5.76	10.84	1.00	Confirmed
avgTempLag1	8.29	8.36	4.55	10.31	1.00	Confirmed
avgTempLag3	8.12	8.22	4.91	11.65	1.00	Confirmed
maxTempLag2	7.86	7.89	5.45	9.71	1.00	Confirmed
avgTempLag2	7.62	7.52	5.71	10.17	1.00	Confirmed
maxTempLag7	7.51	7.41	5.77	9.63	1.00	Confirmed
avgTempLag7	7.36	7.27	5.62	10.35	1.00	Confirmed
minTempLag3	7.28	7.38	4.87	9.97	1.00	Confirmed
avgTempLag14	7.03	7.16	4.31	9.49	1.00	Confirmed
minTempLag7	7.03	6.98	4.80	9.22	1.00	Confirmed
minTempLag2	7.03	7.29	4.40	9.64	1.00	Confirmed
dayOfWeekNumLag1	6.88	6.81	5.27	9.40	1.00	Confirmed
minTempLag14	6.36	6.40	3.70	8.44	1.00	Confirmed
month	6.33	6.24	4.26	8.41	1.00	Confirmed
dayOfWeekNumLag2	5.14	5.06	3.05	7.00	1.00	Confirmed
dayOfWeekNumLag3	4.55	4.64	1.74	5.98	0.97	Confirmed
season	4.49	4.69	2.85	6.60	1.00	Confirmed
day	1.46	1.44	-0.45	3.70	0.21	Rejected
dayLag2	1.23	1.19	-0.92	4.63	0.11	Rejected
dayLag1	1.21	0.90	0.01	2.47	0.02	Rejected
dayLag3	1.20	1.16	-0.12	2.77	0.07	Rejected
dayLag7	1.05	1.02	-0.15	2.15	0.00	Rejected
precip	0.68	0.65	-1.32	2.66	0.04	Rejected
dayLag14	0.60	0.53	-1.41	2.88	0.11	Rejected
rain	0.45	0.32	-1.16	2.07	0.00	Rejected
precipLag2	0.42	0.14	-1.71	3.77	0.11	Rejected
rainLag3	0.20	0.33	-1.25	1.56	0.00	Rejected
rainLag1	0.04	0.07	-1.47	0.96	0.00	Rejected
rainLag14	0.03	0.00	-1.21	1.87	0.00	Rejected
rainLag7	0.00	0.22	-1.53	1.13	0.00	Rejected
snowLag3	-0.05	-0.21	-1.24	1.04	0.00	Rejected
precipLag1	-0.07	-0.25	-0.82	1.37	0.00	Rejected
precipLag3	-0.16	-0.39	-1.51	1.00	0.00	Rejected
precipLag14	-0.17	-0.33	-2.61	1.94	0.00	Rejected
precipLag7	-0.25	-0.43	-2.72	2.34	0.02	Rejected
rainLag2	-0.28	-0.62	-1.42	0.85	0.00	Rejected
snowLag2	-0.47	0.04	-2.10	0.64	0.00	Rejected
snowLag14	-0.87	-1.23	-2.69	0.94	0.00	Rejected
snowLag7	-0.91	-1.05	-1.78	1.08	0.00	Rejected
snow	-0.92	-1.00	-2.45	1.41	0.00	Rejected
snowLag1	-0.98	-1.34	-3.21	1.13	0.00	Rejected

4.7 ANN DEVELOPMENT FOR DAILY FORECASTS

This study utilized the Python-based machine learning library scikit-learn to develop and implement MLP regression models for generating call volume predictions. To prepare the data for model processing, records were divided into training, validation and testing datasets. Daily call volume records from 2010-2015 (2177 instances) are used for training, while records from 2016 (366 instances) are used for validation and 2017 records (365 instances) are reserved for testing. Following hyperparameter tuning and validation, records from 2010-2016 are used for training the final model that generate 2017 predictions. One notable disadvantage of MLP models is that network weights are initialized randomly. As a result, predictive performance levels naturally vary slightly from run to run. One strategy to overcome this issue is to cycle through numerous iterations during validation testing and select the model instance with the highest predictive performance for each model configuration. Therefore, a total of 1000 iterations were performed for each model configuration and the instance with the lowest MAPE across the validation dataset was recorded. Cross validation strategies, such as K-folds, that iteratively train models using slices of a given dataset ignore the temporal components inherent to time series data and consequently are not beneficial in this setting. As an expansion of earlier investigations, the initial MLP models tested were initialized based on the final parameter settings used by Setzler et al. and Chen et al. Each configured their MLP networks with the number of input nodes matching the number of features, a single hidden layer mirroring the number of nodes from the input layer, and a sole output node representing the volume estimation values. In reviewing best practices for forecasting with ANNs, Zhang et al. [33] evaluated historical findings and techniques for determining the

number of hidden layers and nodes to use in a network. They concluded that, for the majority of forecasting problems, one to two hidden layers are appropriate for producing accurate results. When determining the optimal number of hidden nodes, Zhang et al. noted that networks with fewer hidden nodes are typically preferred as they produce more generalized models, reduce the likelihood of overfitting the model based on the training data, and have lower complexity.

To commence MLP model tuning a set of initial hyperparameters were explored through an exhaustive grid search, with the resulting MAPE value serving as the key performance metric. The logistic sigmoid, hyperbolic tangent, and rectified linear unit activation functions were evaluated along with the "lbfgs", stochastic gradient descent, and stochastic gradient-based optimizer solver functions. Initial tests using the stochastic gradient descent and stochastic gradient-based optimizer solver function confirmed that an adaptive learning rate consistently produced better results. Lastly, the hidden node configuration strategy described by Zhang et al. was evaluated in the initial grid search using one to two hidden layers and hidden node sizes equal to 20, x, x+1, x-1, 2x, 50, 100, 150, 200, and 300 where (x) represents the number of input features. Using the complete set of relevant features identified during feature selection, multiple rounds of random and grid search found that models configured with the logistic sigmoid activation function, stochastic gradient-based optimizer solver function, a single hidden layer, and 100 hidden nodes outperformed alternative configurations.

4.8 DAILY CALL VOLUME FORECAST RESULTS

To compare the estimation performance across the assorted daily forecasting methods, the MAPE values were calculated using the daily call volume records from 2017. Mean absolute deviation (MAD) values were also calculated for each method to provide perspective on the average number of calls under or over forecasted. All of the forecasting methods resulted in considerably accurate daily call volume estimations. The MLP model resulted in the lowest overall MAPE of 5.91% with a MAD of 14.54. $ARIMA_{(3,1,3)(2,0,2,7)}$ performed the worst yielding a MAPE of 7.31% and a MAD of 17.42. The MDF method resulted in a MAPE of 6.66% and a MAD of 16.38, which is consistent with the historical forecasting error reported by MEDIC [32]. The Holt-Winters's triple exponential smoothing method using optimal parameter values resulted in a comparable MAPE value of 6.22% with a MAD of 15.20. Estimations using the DOWMA method were calculated for a range of period values from (1-365). It was found that as the number of periods increased the MAPE decreased until reaching a global minimum MAPE value of 6.16% at 56 periods with a MAD of 15.28. The range of MAPE values by means of increasing DOWMA periods are illustrated in Figure 8.

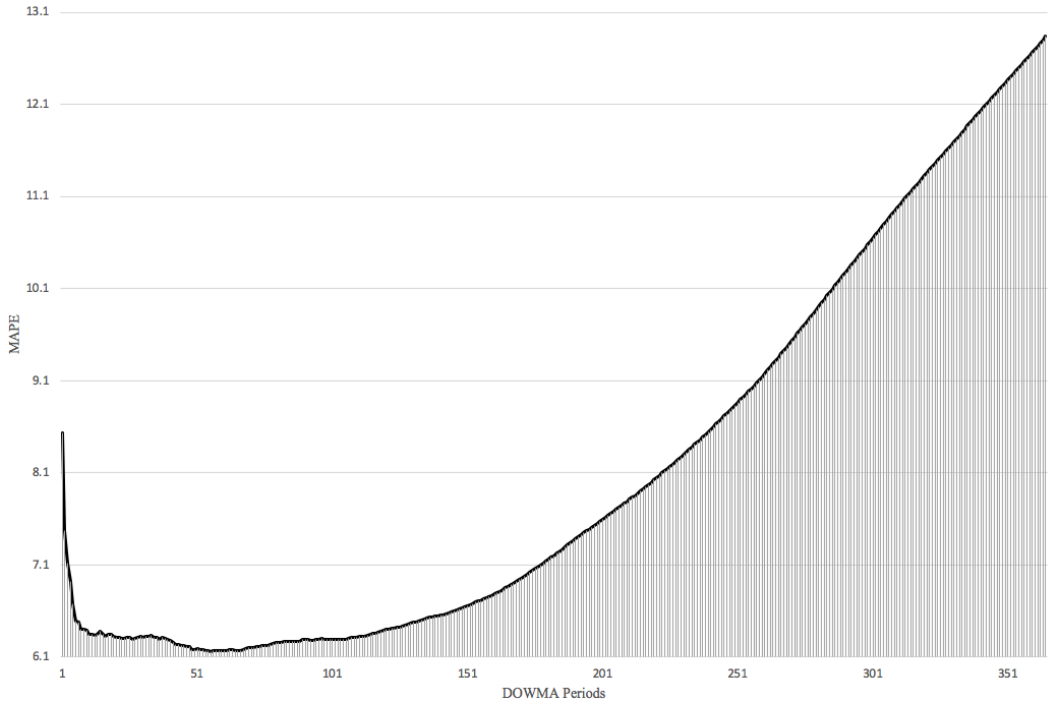

Figure 8: DOWMA Method MAPE Values (Periods 1-365)

These findings suggest that EMS agencies can easily generate accurate daily call volume forecasts, which are used for work force and vehicle scheduling, using any of the approaches evaluated in this study. However, albeit it is a more complicated method, there are several inherit advantages of using an MLP approach. ANNs do not require statistical assumptions regarding the distribution of data, for example, they do not require stationarity to produce unbiased estimates. Furthermore, following training, they produce predictions rapidly and are exceedingly reliable in a production environment.

CHAPTER 5: HOURLY CALL VOLUME FORECASTS

Hourly call volume forecasts serve as critical inputs into detailed work-force schedules (peak and off-peak demand periods). To generate hourly forecasts, an MLP model is further developed based on the performance of daily forecasting models and results are compared against a benchmark method currently used by practitioners at MEDIC to generate hourly call volume estimations. ARIMA and a variation of the DOWMA method for hourly estimations are also employed as models of comparison. To determine the stationary nature of the hourly time series, the ADF test was performed against the complete hourly call volume dataset, yielding a p-value of 4.1E-28, verifying that the series is stationary. Preparing the ARIMA model for hourly forecasts, the seasonality parameter (m) is set to twenty-four. Auto-ARIMA identified the optimal model configuration to be $ARIMA_{(1,1,1)(0,0,4,24)}$.

5.1 THE MEDIC HOURLY FORECASTING METHOD

Identified in this study as the "MEDIC Hourly Forecasting" (MHF) method, the MHF equation is defined as:

$$F_{h,d,w,y} = \frac{\Sigma_{i=0}^{4} \Sigma_{j=1}^{4} A_{h,d,w-j,y-i}}{20} \quad (6)$$

where the following notation applies:

F = forecasted call volume
A = actual call volume
h = hour
d = day of the week (Mon, Tue, Wed, Thu, Fri, Sat, Sun)
w = week
y = year

The MHF method is a unique adaptation of a traditional moving average formula that averages a total of 20 periods to produce estimations. The forecasted value for a given hour is calculated by averaging call volumes from the same hour of the day for the previous four day of the week time periods over the past five years. Citing the formula as a common industry approach, the MHF method was also used by Setzler et al. [5] as a benchmark method throughout their study. It is important to note that while Setzler et al. used the MHF method to produce and compare spatiotemporal call volume estimations, MEDIC does not spatially distribute their hourly forecasts using this method. Despite this fact, the MHF method is a suitable benchmark method for non-spatial and spatially distributed forecasts.

5.2 HOURLY DAY OF WEEK MOVING AVERAGE METHOD

Given that the DOWMA method produced the second lowest MAPE value for daily call volume estimations, the following modified version is proposed for calculating non-spatial and spatially distributed hourly estimations:

$$F_{h,d,w} = \frac{A_{h,d,w-1} + A_{h,d,w-2} + A_{h,d,w-3} + \cdots + A_{h,d,w-n}}{n} \quad (7)$$

where the following notation applies:

F = forecasted call volume
A = actual call volume
h = hour
d = day of the week (Mon, Tue, Wed, Thu, Fri, Sat, Sun)
w = week
n = number of periods

This variation of the DOWMA method, identified in this study as the H-DOWMA method, calculates the hourly call volume for a given hour by averaging the total call volume actuals from the same hour and day of the week over (n) previous week. As with the daily

DOWMA forecasts, the number of periods (n) are adjusted in order to minimize the resulting MAPE value.

5.3 HOURLY FEATURE SELECTION

Feature selection for hourly call volume MLP models was completed following the same process utilized for daily call volume models. In addition to the features previously identified (year, season, month, week, day, day of week, hour, average daily temperature, minimum daily temperature, maximum daily temperature, precipitation in inches, and binary variables indicating occurrences of daily rain and snow fall) lag values of hourly call volume for periods 1-7, 24, 48, 72, 96, 168 were added to the candidate feature set based on the optimal hourly $ARIMA_{(1,1,1)(0,0,4,24)}$ model. Using a random forest regressor, Boruta rejected the snow feature and confirmed all remaining features as relevant for producing hourly EMS call volume estimations. The complete ranked results are presented in Table 7.

Table 7: Hourly Call Volume Prediction Features Ranked Results

Features	Mean Importance	Median Importance	Minimum Importance	Maximum Importance	Norm Hits	Decision
year	63.80	63.53	57.02	67.62	1.00	Confirmed
dayOfWeekNum	57.81	57.71	53.02	63.47	1.00	Confirmed
hour	57.50	57.40	53.60	60.53	1.00	Confirmed
volumeLag1	50.74	51.02	47.36	53.66	1.00	Confirmed
volumeLag168	49.29	49.51	44.56	53.04	1.00	Confirmed
volumeLag2	47.42	47.61	43.70	51.14	1.00	Confirmed
volumeLag24	44.01	44.22	41.55	46.33	1.00	Confirmed
volumeLag3	39.90	39.73	36.59	43.31	1.00	Confirmed
volumeLag96	37.98	38.06	36.08	40.53	1.00	Confirmed
volumeLag48	36.74	36.37	34.43	40.27	1.00	Confirmed
volumeLag72	36.01	36.01	31.27	38.82	1.00	Confirmed
volumeLag7	31.89	31.98	28.74	34.00	1.00	Confirmed
volumeLag6	30.22	30.17	27.40	32.96	1.00	Confirmed
volumeLag4	29.99	30.09	27.60	32.39	1.00	Confirmed
maxTemp	28.83	28.63	17.97	33.44	1.00	Confirmed
volumeLag5	27.87	27.83	24.82	31.51	1.00	Confirmed
avgTemp	25.72	25.90	16.24	30.07	1.00	Confirmed
minTemp	24.69	24.97	19.96	29.58	1.00	Confirmed
weekNumber	20.75	21.57	7.01	25.86	1.00	Confirmed
month	16.01	16.59	5.73	20.62	1.00	Confirmed
season	12.16	12.08	9.97	14.60	1.00	Confirmed
precip	3.41	3.46	1.50	5.24	0.94	Confirmed
day	3.35	3.46	1.41	6.27	0.94	Confirmed
rain	2.99	3.00	0.92	5.48	0.81	Confirmed
snow	-2.57	-2.36	-3.87	-0.84	0.00	Rejected

5.4 ANN DEVELOPMENT FOR HOURLY FORECASTS

Based on the performance of the MLP model developed for generating daily call volume predications, a subsequent MLP model was initialized with comparable configurations for producing hourly call volume predictions. As before, the call volume records from 2010-2016 serve as the training/validation dataset and 2017 records are used for testing. Consistent with the grid-search strategy applied to daily MLP models, hourly MLP models having (x) number of input features were evaluated using hidden node sizes equal to 20, x, x+1, x-1, 2x, 50, 100, 150, 200, and 300 with one to two hidden layers. Multiple rounds of hyperparameter tuning was completed using the array of activation functions, solver

functions, and hidden node/layer combinations. Through grid-search the MLP model yielding the lowest MAPE utilized the hyperbolic tangent activation function, stochastic gradient descent solver function, and a single hidden layer containing 25 (x+1) hidden nodes.

5.5 HOURLY CALL VOLUME FORECAST RESULTS

To compare the estimation performance across all of the non-spatially distributed hourly forecasting methods, the MAPE values were calculated using the hourly call volume records from 2017. While in practice, the EMS operations planning horizon is significantly shorter (hourly, daily, weekly), testing against the complete annual dataset serves as a more generalized benchmark when comparing the adaptive nature of a machine learning method for time-series forecasting, i.e. MLP, versus rolling moving-average methods such as MHF and H-DOWMA. This strategy also ensures the seasonal adjustment (hourly, daily, weekly, monthly, etc.) capabilities of each method are thoroughly evaluated. ARIMA produced the worst results at the hourly scale with a MAPE of 41.37% and a MAD of 3.52, while the MHF method resulted in a MAPE of 27.14% and a MAD of 2.67. The MLP model yielded a MAPE value of 26.97% and a MAD of 2.64. Hourly call volume estimations using the H-DOWMA method were calculated based on the range of available data periods (1-350). The MAPE values decreased following a negative exponential pattern as the number of periods increased ultimately reaching a minimum of 26.22% at 330 periods. This pattern is illustrated in Figure 9.

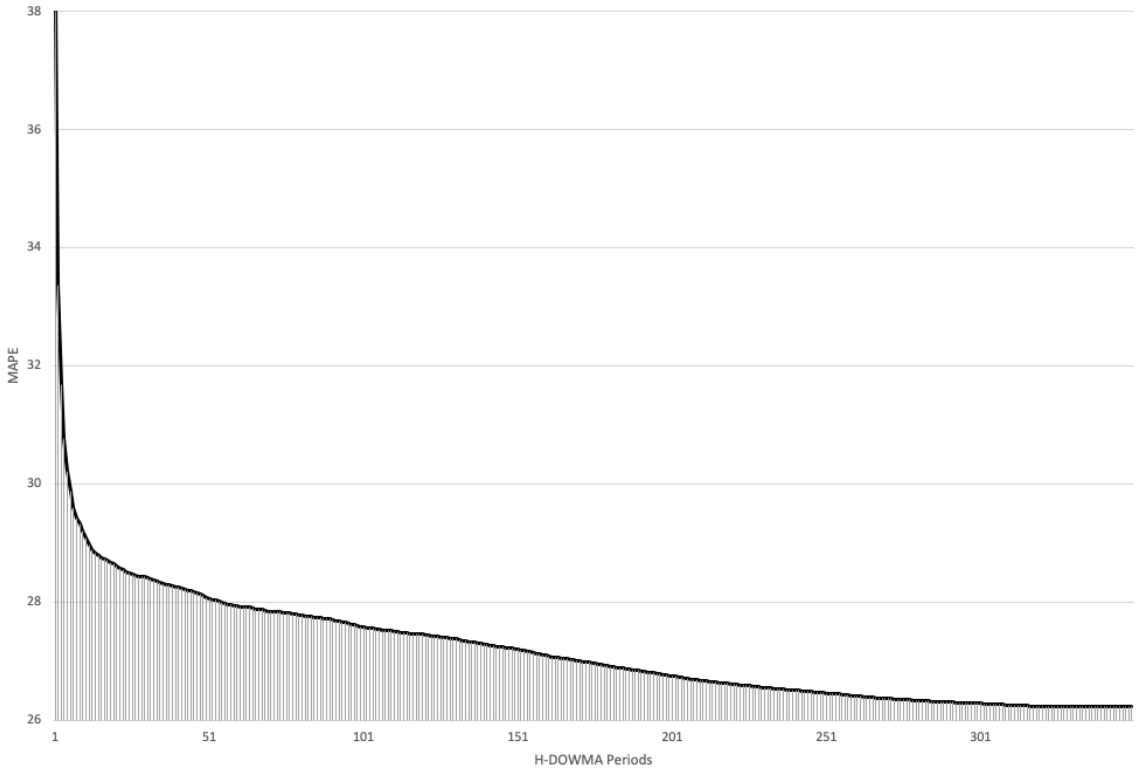

Figure 9: H-DOWMA Method MAPE Values (Periods 1-350)

CHAPTER 6: SPATIOTEMPORAL CALL VOLUME FORECASTS

Accurate hourly call volume forecasts that incorporate the spatial component of demand are critical to supporting real-time operational activities such as ambulance dispatch decisions and initiating dynamic redeployment strategies aimed at perpetually maximizing response coverage across a defined geographic service area. As previously noted, as calls are received, the latitude and longitude coordinates representing the originating call locations are recorded. This information is useful for conducting spatial demand analysis and identifying areas that regularly experience higher call volumes. To incorporate the spatial component of demand into temporal forecasting models, a K-Means clustering approach is implemented to separate calls into related groupings based on call location densities. Ancillary geospatial analysis work is presented in Appendix B.

6.1 SPATIAL CLUSTERING

As discussed in the literature review, Setzler et al. [5] were the first to incorporate the spatial component of demand and generate demand estimations at various spatiotemporal granularities. Chen et al. [16] and Zhou et al. [24] later conducted related spatiotemporal studies exploring the effectiveness of a variety of forecasting techniques. In each of the three investigations, the researchers developed fixed square-mile/kilometer grid block systems to segment geographic areas and formulate estimations. In doing so, Setzler et al. [5] and Chen et al. [16] encountered a key limitation of this approach; zero-inflated demand distributions that are created when scaled to finer degrees of spatial granularity. After presenting their results, Setzler et al. [5] stated that at a certain degree finer levels of specificity have little practical value and suggested future investigations focus on *"varying*

population densities in order to determine optimal or near-optimal geographic grid sizes and time intervals".

The relationship between population density and call volume densities has been recognized in numerous studies [3, 8, 10-13]. Employing even simple methods, such as linear regression, consistently reveals strong causal relations between call volumes and population counts. While fixed geographic grids provide a straightforward approach to spatial segmentation, the equal division of space does not account for the unequal distribution of populations and associated call demand. Alternatively, a general-purpose clustering algorithm, such as K-Means, can be applied as a data mining technique to produce geographically heterogeneous spatial clusters based on call location (latitude and longitude coordinate) and associated call volume densities.

K-Means is a robust clustering algorithm that divides a collection of N samples into K distinct clusters, each defined by the average (centroid) of the samples in the cluster. The objective of the K-Means algorithm is to minimize the within-cluster sum of squared (WCSS) error (Euclidean distance) between sample points and the cluster centroid [36]. Since data points are grouped based on proximity, K-Means is a common approach for creating spatial clusters using latitude and longitude coordinates. In the prolific geographic information systems platform ArcGIS, K-Means is one of the primary unsupervised machine learning methods used by the ArcMap grouping analysis tool to identify natural groupings in spatial datasets. Moore and Dixon [37] utilized the grouping analysis tool to cluster instances of tornados produced by Hurricane Ivan in 2004 based on their

latitude/longitude coordinates and relative time of touchdown. Naaman et al. [38] also explored a variation of K-Means for automatic organization of digital photos based on geographic coordinates and time-based event information.

The scikit-learn implementation of K-Means used in this study requires that the number of clusters *K* be specified and automatically initializes the clusters through the random selection of *K* number of samples that serve as the initial cluster centroids. The algorithm continues by looping between two steps. As described in the sci-kit learn K-Means documentation; "*The first step assigns each sample to its nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed, and the algorithm repeats these last two steps until the centroids do not move significantly based on a user-specified tolerance"* [39].

Establishing clusters of varying sizes based on call location volume results in groupings that are more representative of associated population density and reduces the likelihood of generating zero-inflated distributions. As suggested by Setzler et al., a practical level of granularity is required to produce operationally valuable demand estimations and support real-time dispatching decisions. Citing average travel times, distances, and the size of the study area, operations managers at the participating EMS agency stated that approximately seven to nine spatial clusters would serve as an optimal level of granularity for utilizing hourly call volume forecasts for real-time redeployment decisions [32]. This equates to three different K-Means model configurations (i.e. K = 7, K = 8, and K = 9). To further

evaluate the optimal number of clusters (i.e. the optimal level of granularity) a technique known as the elbow method is applied. The elbow method aids in identifying an optimal or near-optimal value for K by fitting the model with a range of values for K and calculating the average within-cluster sum of squared (WCSS) error for each cluster set. The WCSS error values for each model are then plotted on a line chart, which frequently resembles the shape of an arm if the data is properly clustered. The elbow, or inflection point on the curve, is an indicator of the optimal value, or range of values, of K. After applying the elbow method to our call location dataset, 5-10 clusters appear to be the optimal range based on the declining rate of change of the within clusters sum of squared error. This supports the anecdotal recommendation of the participating EMS agency that 7-9 spatial clusters serve as an optimal level of granularity for hourly call volume planning. The resulting elbow method chart is shown below in Figure 10.

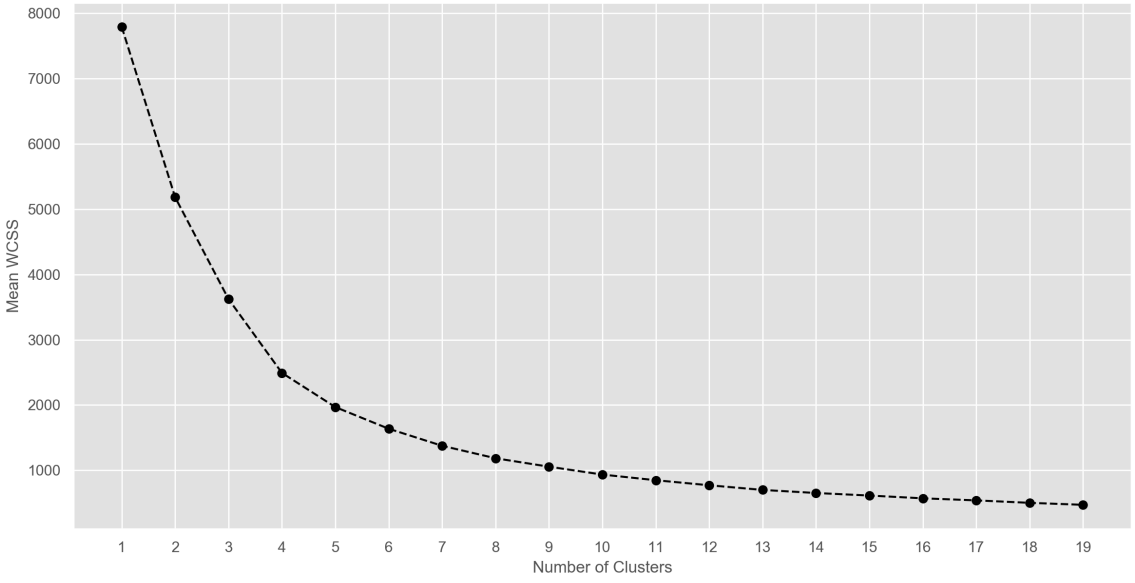

Figure 10: P1+P2 Spatial Clusters Elbow Plot

Using the original P1+P2 call records, the latitude and longitude coordinates for each call were extracted and processed through a K-Means clustering algorithm implemented in

scikit-learn. Upon convergence, a numeric label identifying the corresponding spatial cluster for each call was appended to each of the original call records. This process was repeated for values of K = 7, 8, and 9. The call volume counts per cluster are listed in Table 8. The resulting K-Mean clusters are visualized below in Figures 11, 12, and 13.

Table 8: Total Call Volume Counts Per Cluster

Cluster	K=7	% of total	K=8	% of total	K=9	% of total
0	95416	15%	102252	16%	65065	10%
1	164594	26%	50723	8%	74044	12%
2	89018	14%	87136	14%	49174	8%
3	58100	9%	62441	10%	147307	23%
4	37516	6%	82883	13%	37062	6%
5	72756	11%	34282	5%	45524	7%
6	116017	18%	51004	8%	77317	12%
7	-	-	162696	26%	113314	18%
8	-	-	-	-	24610	4%

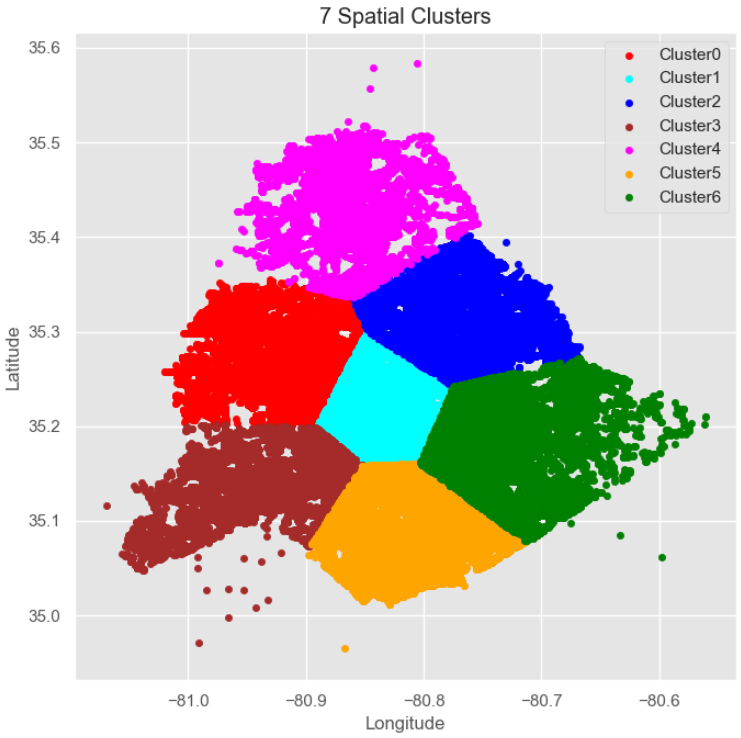

Figure 11: P1+P2 Calls K-Mean Spatial Clusters; K=7 (2010-2017)

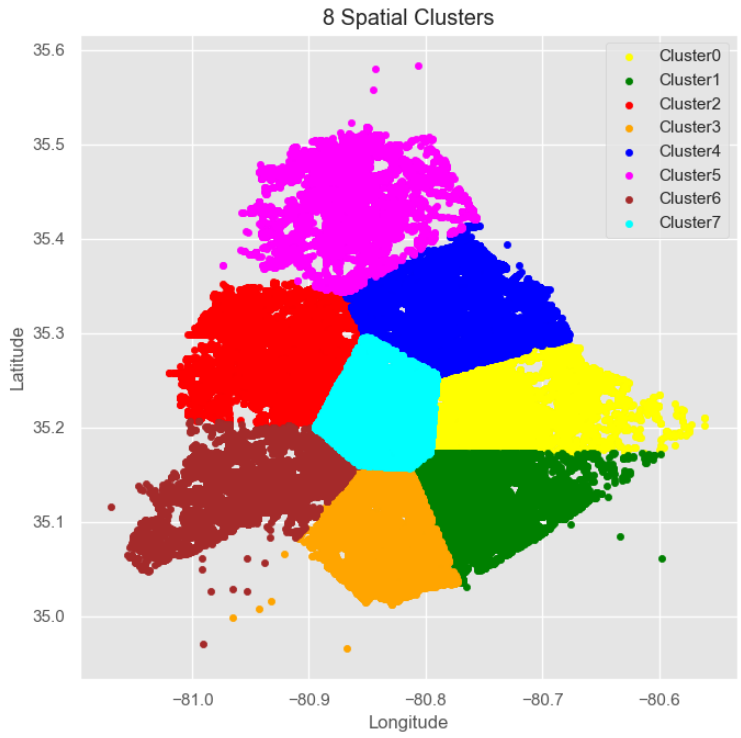

Figure 12: P1+P2 Calls K-Mean Spatial Clusters; K=8 (2010-2017)

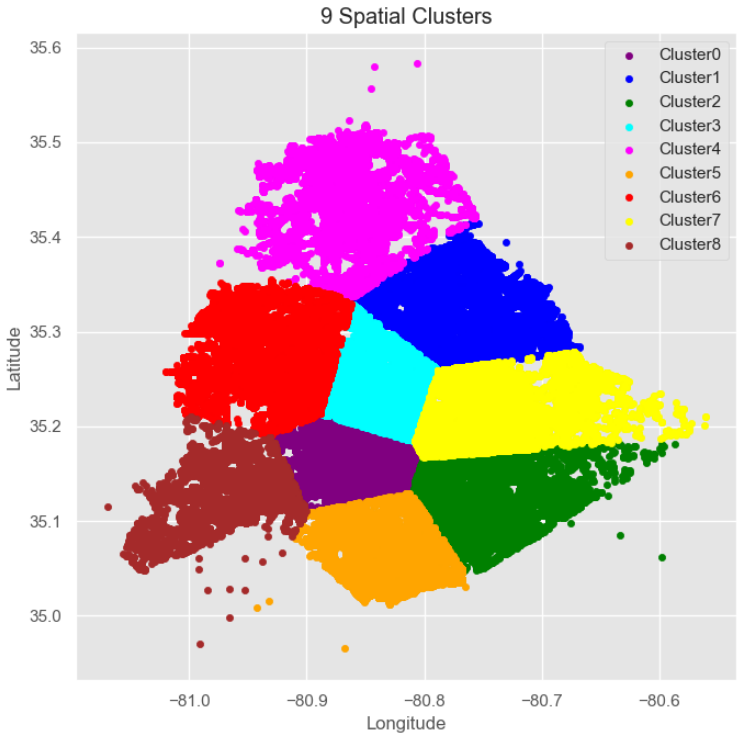

Figure 13: P1+P2 Calls K-Mean Spatial Clusters; K=9 (2010-2017)

6.2 SPATIALLY DISTRIBUTED HOURLY CALL VOLUME FORECASTS

As an extension of the non-spatial hourly call volume forecasts, the performance of an MLP model is compared against the MHF method, H-DOWMA method, and ARIMA for producing spatially distributed hourly call volume forecasts. In preparation for model processing, the hourly call volumes for the range of spatial cluster sets (K = 7, 8, 9) were aggregated separately by counting the number of call instances that occurred within each cluster during individual hours. This resulted in three spatially distributed hourly call volume datasets for 7, 8 and 9 spatial clusters respectively. To ensure a complete and accurate time series, zero call volume records for any missing cluster/hour instances were generated and inserted into to each dataset. The ADF test was again conducted to verify that each of the newly generated time series datasets are stationary prior to model processing. Processing each of the datasets individually, spatially distributed call volume

estimations using the hourly $ARIMA_{(1,1,1)(0,0,4,24)}$ model were produced. For the H-DOWMA and MHF methods, call volume prediction values for individual hours were generated by iteratively filtering the dataset using the corresponding day of the week, hour, and k-means spatial cluster label. This resulted in spatially disturbed hourly call volume forecasts by cluster using $ARIMA_{(1,1,1)(0,0,4,24)}$, H-DOWMA, and the MHF.

6.3 SPATIALLY DISTRIBUTED HOURLY FEATURE SELECTION

Feature selection for spatially distributed hourly call volume MLP models was carried out following the same process employed for hourly and daily call volume models. With the addition of the K-Means clustering label for each of the three spatial cluster sets (K = 7, 8, 9) the complete candidate feature set included: year, season, month, week, day, day of week, hour, average daily temperature, minimum daily temperature, maximum daily temperature, precipitation in inches, binary variables indicating occurrences of daily rain and snow fall, the K-Means cluster label, and lag values of hourly call volume per cluster for periods 1-7, 24, 48, 72, 96, 168. Processing each of the three datasets separately through a random forest regressor, Boruta consistently rejected the snow feature and confirmed all other features as relevant for estimating spatially distributed hourly EMS call volumes. The rank results are presented below in Table 9.

Table 9: Spatial Hourly Call Volume Prediction Features Ranked Results

Features	Mean Importance	Median Importance	Minimum Importance	Maximum Importance	Norm Hits	Decision
hour	156.13	156.53	149.79	161.67	1.00	Confirmed
kMeanCluster	140.36	140.93	133.71	143.47	1.00	Confirmed
maxTemp	56.88	56.20	47.35	65.29	1.00	Confirmed
year	55.13	54.70	52.43	58.59	1.00	Confirmed
weekNumber	51.41	51.31	43.68	58.76	1.00	Confirmed
avgTemp	47.90	47.48	44.70	51.94	1.00	Confirmed
volumeLag6	46.96	46.76	44.71	49.64	1.00	Confirmed
volumeLag2	46.33	45.93	43.81	50.43	1.00	Confirmed
volumeLag168	45.61	45.52	42.82	48.30	1.00	Confirmed
dayOfWeekNum	44.30	44.42	43.06	44.99	1.00	Confirmed
minTemp	43.92	42.90	38.22	52.58	1.00	Confirmed
volumeLag96	41.82	41.52	39.61	44.78	1.00	Confirmed
volumeLag24	40.96	40.99	37.91	43.54	1.00	Confirmed
volumeLag48	40.93	40.76	37.78	44.19	1.00	Confirmed
volumeLag72	40.80	40.71	38.78	43.58	1.00	Confirmed
volumeLag1	39.31	39.61	33.52	42.14	1.00	Confirmed
volumeLag4	35.35	35.91	32.93	36.77	1.00	Confirmed
volumeLag3	33.23	33.24	31.27	34.53	1.00	Confirmed
month	33.15	32.38	28.27	39.97	1.00	Confirmed
volumeLag5	27.36	27.39	25.89	28.24	1.00	Confirmed
volumeLag7	22.33	22.49	20.53	23.67	1.00	Confirmed
season	21.02	20.71	18.03	26.86	1.00	Confirmed
rain	11.23	11.80	8.06	13.01	1.00	Confirmed
precip	10.85	10.85	9.24	12.07	1.00	Confirmed
day	6.20	6.72	4.39	6.98	1.00	Confirmed
snow	-2.280433	-2.230527	-4.249357	-0.4524194	0	Rejected

6.4 ANN DEVELOPMENT FOR SPATIALLY DIST. HOURLY FORECASTS

MLP model development, hyperparameter tuning, and implementation for generating spatially distributed hourly call volume estimations was carried out following the same process used for producing hourly non-spatial call volume predictions. Each model instance was trained using the input features and hourly call volume target values from each dataset for all clusters, with the K-Means cluster label serving as the spatial identifier in each configuration K=7, 8, 9. Following multiple rounds of hyperparameter tuning, the top performing MLP model was a deep neural network configured with the hyperbolic

tangent activation function, the stochastic gradient descent solver function, and two hidden layers each containing 200 hidden nodes.

6.5 SPATIALLY DISTRIBUTED HOURLY FORECAST RESULTS

To compare the estimation performance across the MHF method, H-DOWMA, ARIMA, and MLP models, MAPE values were calculated separately for each cluster using 2017 records from each of the three spatially distributed hourly call volume datasets (K = 7, 8, 9). ARIMA consistently resulted in the highest MAPE values of 53.48%, 54.93%, and 51.55% for spatial cluster sets 7, 8, and 9 correspondingly. H-DOWMA also resulted in higher error for each spatial cluster set with MAPE values of 51.71%, 49.60%, and 47.86% for 7, 8 and 9 spatial clusters respectively. The performance of the MHF method improved slightly as the level of cluster granularity increased with MAPE values of 48.57% at 7 clusters, 47.59% at 8 clusters, and 46.13% at 9 clusters. The MLP models considerably outperformed the other methods with average MAPE values of 40.22%, 38.10%, and 36.76% across 7, 8, and 9 cluster sets respectively. A comparison of the MAPE and MAD values for ARIMA, the MHF method, and MLP at the per cluster level using 7, 8, and 9 clusters are provided in Tables 10, 11, and 12 respectively.

Table 10: Per Cluster MAPE/MAD Results by Method (7 Clusters)

Cluster	ARIMA MAPE	ARIMA MAD	MHF MAPE	MHF MAD	MLP MAPE	MLP MAD
0	45.96	1.03	51.94	1.01	43.50	1.03
1	66.79	1.51	49.66	1.32	45.31	1.37
2	45.61	1.02	49.26	0.99	43.57	0.99
3	49.10	0.79	46.33	0.81	37.33	0.82
4	58.17	0.72	41.68	0.65	30.61	0.64
5	48.18	0.86	49.02	0.86	36.94	0.87
6	60.54	1.15	52.07	1.08	44.31	1.09
Average =	53.48	1.01	48.57	0.96	40.22	0.97

Table 11: Per Cluster MAPE/MAD Results by Method (8 Clusters)

Cluster	ARIMA MAPE	ARIMA MAD	MHF MAPE	MHF MAD	MLP MAPE	MLP MAD
0	66.59	1.12	52.33	1.01	42.32	1.04
1	53.66	0.74	45.05	0.72	35.45	0.67
2	44.72	0.96	51.28	0.97	40.49	1.00
3	49.90	0.78	47.24	0.79	35.73	0.76
4	44.53	0.96	48.85	0.96	40.26	1.02
5	60.14	0.72	40.61	0.62	28.86	0.59
6	52.03	0.76	45.47	0.76	35.68	0.77
7	67.89	1.51	49.91	1.31	46.02	1.32
Average =	54.93	0.94	47.59	0.89	38.10	0.89

Table 12: Per Cluster MAPE/MAD Results by Method (9 Clusters)

Cluster	ARIMA MAPE	ARIMA MAD	MHF MAPE	MHF MAD	MLP MAPE	MLP MAD
0	48.41	0.79	48.03	0.81	38.04	0.78
1	44.88	0.88	48.14	0.90	39.56	0.92
2	54.57	0.74	44.79	0.71	34.89	0.65
3	51.89	1.33	49.65	1.26	45.72	1.30
4	58.66	0.72	41.56	0.65	31.39	0.61
5	55.55	0.73	43.78	0.69	32.07	0.67
6	45.69	0.87	51.35	0.91	43.24	0.94
7	63.75	1.14	52.14	1.05	44.36	1.07
8	40.52	0.58	35.71	0.53	21.57	0.46
Average =	51.55	0.86	46.13	0.83	36.76	0.82

CHAPTER 7: SUMMARY AND CONCLUSIONS

7.1 DISCUSSION

The objective of this study was to present a forecasting methodology that utilizes machine learning methods to generate daily, hourly, and spatiotemporal call volume estimations at a degree of granularity in space and time that is both practical and actionable. Forecasting model results demonstrate that, given systematic feature selection and hyperparameter model tuning, MLP models consistently produce more accurate predictions. For non-spatially distributed daily and hourly call volume predictions, traditional time-series and the MDF/MHF benchmark methods are shown to perform at marginally similar levels of predictive performance without the added complexity of machine learning methods. This means that when practitioners forecast daily or hourly call volumes for the entire county, they can employ either the MHF method or MLP without sacrificing predictive performance. These results are consistent with the findings of previous studies by Setzler et al. [5] and Chen et al. [16].

Conversely, when producing spatially distributed call volume forecasts for finer geographic areas, MLP models significantly outperform traditional time series methods across all spatial levels evaluated (7, 8, and 9 clusters). Comparing the performance of spatially distributed hourly forecasts generated using MLP versus the MHF benchmark method, the MLP model consistently yields lower estimation error in terms of MAPE and MAD. This finding has practical implications given the need practitioners have for accurate call volume predictions at more granular spatiotemporal levels. Additionally, MLP model predictions are found to be more accurate within clusters containing relatively lower call

volumes (clusters with 10% or less of the total county-wide call volume). Using the 9-cluster configuration for visual analysis, Figure 14 presents the MAPE values and percent differences between the MLP and MHF methods on a choropleth map shaded based on the percent of total call volume by cluster. The distinction in predictive performance levels between the MLP and MHF methods at the per cluster level is also apparent in terms of absolute deviation (error) when examining the error distribution per cluster (Figure 15).

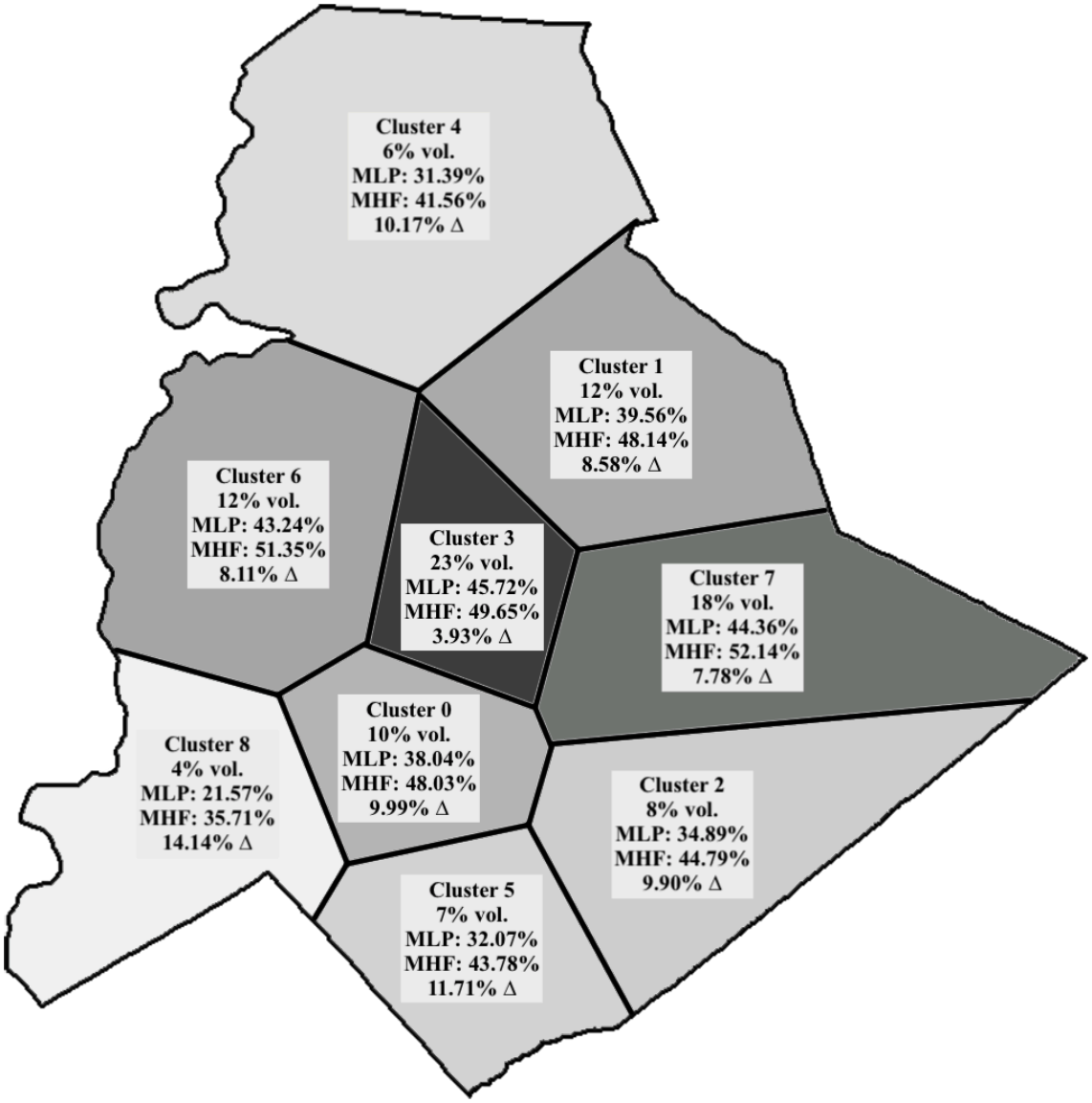

Figure 14: Percent of Total Call Volume and MAPE Values by Cluster.

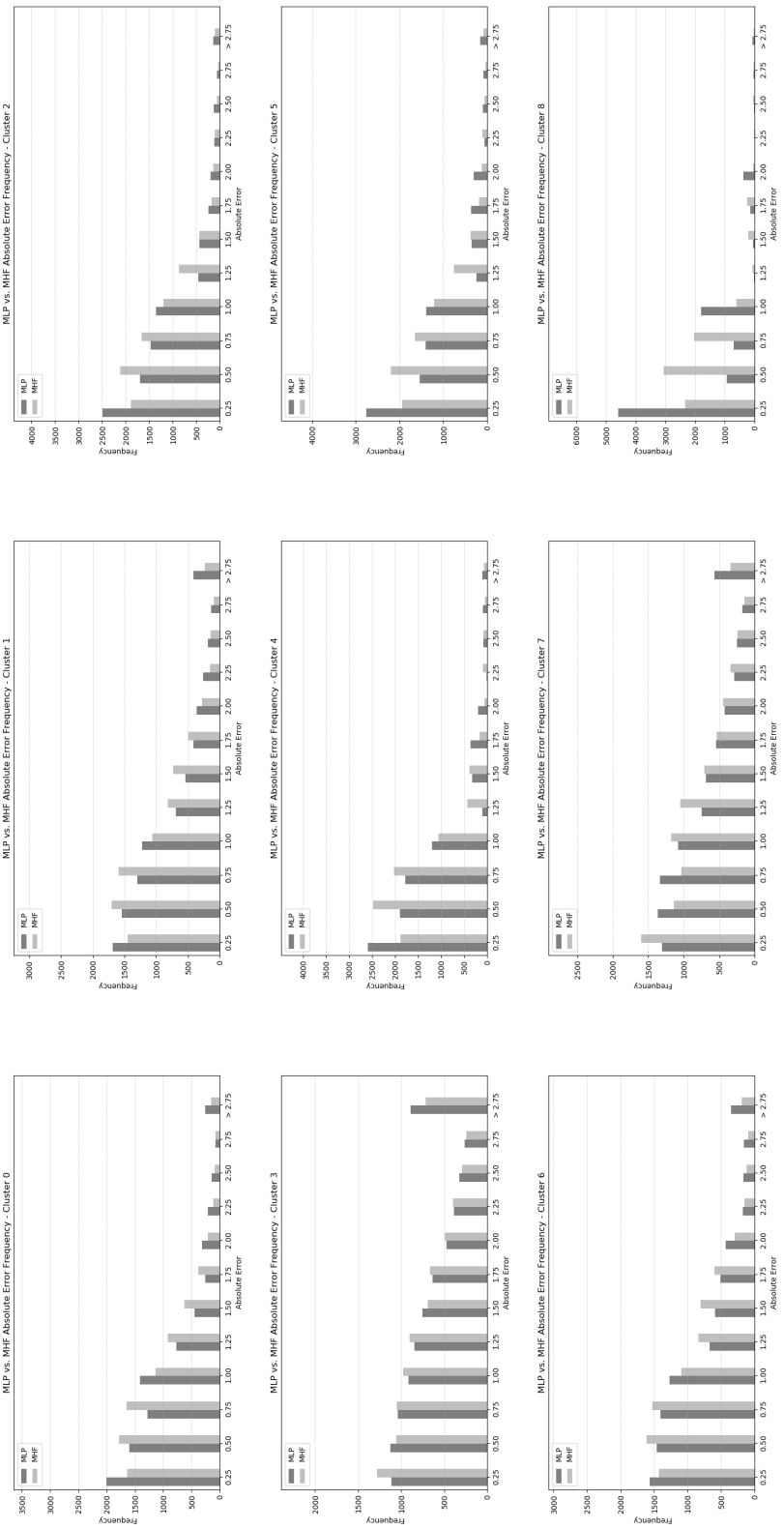

Figure 15: MLP vs. MHF Absolute Error Frequency Distributions by Cluster

Both approaches are shown to produce higher MAPE values in clusters containing higher (>10%) call volume (and associated population) densities (i.e. clusters 1, 3, 6, and 7). Geographically, these clusters represent urban centers within the county and include a greater concentration of mixed-use developments, multi-family housing units, and commercial buildings. These higher density clusters also encompass, or are adjacent to clusters containing, several hospitals with emergency departments (e.g., cluster 3). According to practitioners at the participating EMS agency, at any given time there are multiple ambulances located in and around the hospitals either transferring patients to the facility or completing paperwork. Hence, a major challenge of deployment planning is providing sufficient ambulance coverage across a mix of urban, suburban, and rural areas. Suburban and rural areas are exceptionally difficult for coverage planning; given lower call frequency and increased travel distances [32]. Therefore, a forecasting model, such as MLP, that produces more accurate spatiotemporal call volume estimations than the current industry method (MHF) is critical to providing adequate coverage and minimizing call response times. Particularly in lower call volume areas where the need for accurate predictions is greater.

7.2 STUDY LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

There are several limitations of this study that should be considered. First, the EMS call response data used in this study was provided by a single EMS agency serving Mecklenburg County, North Carolina. While geographically the country is comprised of a mixture of urban, suburban and rural areas, the effectiveness of the forecasting methodology presented in this paper should be investigated in similar and dissimilar EMS

service areas. Furthermore, it's important to note that the optimal level of spatial granularity will vary based on the service area size, geographical and topographical features, and population densities. The effectiveness of alternative clustering methods should be explored, and EMS practitioners should be consulted to aid in identifying practical levels of spatial granularity for use in subsequent investigations.

Future studies have the potential to further advance the performance of spatiotemporal call volume prediction models by incorporating additional explanatory variables related to population shifts and weather data (i.e. temperatures, rainfall, etc.) at the hourly level. Datasets related to traffic patterns and population densities distributed by time and space collected through services such as Google Maps™ and Waze™ could be introduced to future models. Given the strong correlation between population and call volume densities, such datasets are likely to further improve the predictive performance of forecasting models. Supplemental demographic and socioeconomic variables may also be beneficial. Lastly, researchers could explore the applications of alternative machine learning models and methods for generating EMS call volume predictions.

REFERENCES

1. Aringhieri, R., et al., *Emergency Medical Services and beyond: Addressing new challenges through a wide literature review.* Computers & Operations Research, 2016.
2. Stephens, C. and E. Mitchell, *MEDIC Communications & Dispatch Office Interview*, R.J. Martin, Editor. 2018.
3. Channouf, N., et al., *The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta.* Health Care Management Science, 2007. **10**(1): p. 25-45.
4. Rajagopalan, H.K., et al., *Ambulance Deployment and Shift Scheduling: An Integrated Approach.* Journal of Service Science and Management, 2011. **1**(4): p. 66-78.
5. Setzler, H., C. Saydam, and S. Park, *EMS call volume predictions: A comparative study.* Computers & Operations Research, 2009. **36**(6): p. 1843-1851.
6. Vile, J.L., et al., *Predicting ambulance demand using singular spectrum analysis.* Journal of the Operational Research Society, 2012. **63**(11): p. 1556-1565.
7. Brown, L.H., et al., *Are EMS Call Volume Predictions Based on Demand Pattern Analysis Accurate?* Prehospital Emergency Care, 2007. **11**(2): p. 199-203.
8. McConnel, C.E. and R.W. Wilson, *The demand for prehospital emergency services in an aging society.* Social Science & Medicine, 1998. **46**(8): p. 1027-1031.
9. Hall, W.K., *Management science approaches to the determination of urban ambulance requirements.* Socio-Economic Planning Sciences, 1971. **5**(5): p. 491-499.
10. Aldrich, C.A., J.C. Hisserich, and L.B. Lave, *An analysis of the demand for emergency ambulance service in an urban area.* American journal of public health, 1971. **61**(6): p. 1156-1169.
11. Siler, K.F., *Predicting demand for publicly dispatched ambulances in a metropolitan area.* Health Services Research, 1975. **10**(3): p. 254-263.
12. Kamenetzky, R.D., L.J. Shuman, and H. Wolfe, *Estimating need and demand for prehospital care.* Operations Research, 1982. **30**(6): p. 1148-1167.
13. Kvålseth, T.O. and J.M. Deems, *Statistical models of the demand for emergency medical services in an urban area.* American Journal of Public Health, 1979. **69**(3): p. 250-255.
14. Baker, J.R. and K.E. Fitzpatrick, *Determination of an Optimal Forecast Model for Ambulance Demand Using Goal Programming.* Journal of the Operational Research Society, 1986. **37**(11): p. 1047-1059.
15. Tandberg, D., J. Tibbetts, and D.P. Sklar, *Time series forecasts of ambulance run volume.* The American journal of emergency medicine, 1998. **16**(3): p. 232-237.
16. Chen, A.Y., et al., *Demand Forecast using Data Analytics for the Pre-allocation of Ambulances.* IEEE Journal of Biomedical and Health Informatics, 2016. **20**(4): p. 1178-1187.
17. Yannis, G. and M.G. Karlaftis. *Weather effects on daily traffic accidents and fatalities: a time series count data approach*. in *Proceedings of the 89th Annual Meeting of the Transportation Research Board*. 2010.

18. Kjellstrom, T., et al., *Public health impact of global heating due to climate change: potential effects on chronic non-communicable diseases.* International journal of public health, 2010. **55**(2): p. 97-103.
19. Leung, W., Y. Leung, and H. Mok. *Impact of weather on human health*. in *Guangdong-Hong Kong-Macau Seminar on Meteorological Science and Technology, held in Zhongshan, China on*. 2008.
20. Rooney, C., et al., *Excess mortality in England and Wales, and in Greater London, during the 1995 heatwave.* Journal of Epidemiology & Community Health, 1998. **52**(8): p. 482-486.
21. Wong, H. and P. Lai, *Weather inference and daily demand for emergency ambulance services.* Emergency Medicine Journal, 2010: p. emj. 2010.096701.
22. McLay, L.A., E.L. Boone, and J.P. Brooks, *Analyzing the volume and nature of emergency medical calls during severe weather events using regression methodologies.* Socio-Economic Planning Sciences, 2012. **46**(1): p. 55-66.
23. Wong, H.-T. and P.-C. Lai, *Weather factors in the short-term forecasting of daily ambulance calls.* International Journal of Biometeorology, 2014. **58**(5): p. 669-678.
24. Zhou, Z., et al., *A Spatio-Temporal Point Process Model for Ambulance Demand.* Journal of the American Statistical Association, 2015. **110**(509): p. 6-15.
25. Tang, Z. and P.A. Fishwick, *Feedforward neural nets as models for time series forecasting.* ORSA journal on computing, 1993. **5**(4): p. 374-385.
26. Hill, T., M. O'Connor, and W. Remus, *Neural network models for time series forecasts.* Management science, 1996. **42**(7): p. 1082-1092.
27. Mitchell, T.M., *Machine Learning*. 1997: McGraw-Hill.
28. Agency, M.M.E., *2017 Annual Report.* 2018: p. 11.
29. Wong, H. and P. Lai, *Weather inference and daily demand for emergency ambulance services.* Emergency Medicine Journal, 2012. **29**(1): p. 60-4.
30. Hamilton, J.D., *Time series analysis*. Vol. 2. 1994: Princeton university press Princeton, NJ.
31. Hyndman, R.J. and Y. Khandakar, *Automatic time series for forecasting: the forecast package for R*. 2007: Monash University, Department of Econometrics and Business Statistics ….
32. Penner, J., J. Studnek, and A. Infinger, *Interview: EMS Call Data for Research*, C. Saydam, J. Martin, and V. Vasudev, Editors. 2016.
33. Zhang, G., B. Eddy Patuwo, and M. Y. Hu, *Forecasting with artificial neural networks:: The state of the art.* International Journal of Forecasting, 1998. **14**(1): p. 35-62.
34. Kohavi, R. and G.H. John, *Wrappers for feature subset selection.* Artificial intelligence, 1997. **97**(1-2): p. 273-324.
35. Kursa, M.B. and W.R. Rudnicki, *Feature Selection with the Boruta Package.* Journal of Statistical Software, 2010. **36**(11): p. 1-13.
36. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108.

37. Moore, T.W. and R.W. Dixon, *A Spatiotemporal Analysis and Description of Hurricane Ivan's (2004) Tornado Clusters.* Papers in Applied Geography, 2015. **1**(2): p. 192-196.
38. Naaman, M., et al. *Automatic organization for digital photographs with geographic coordinates*. in *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.* 2004. IEEE.
39. *Scikit Learn, K-Means Clustering*. [cited 2019; Available from: https://scikit-learn.org/stable/modules/clustering.html#k-means.

APPENDIX A: SUB-DAILY TIME FRAME CALL VOLUME FORECASTS

A.1 TEMPORAL CLUSTERING

In section 6.1, an implementation of the K-Means clustering algorithm was used to achieve dynamic spatial segmentation based on recorded call locations and associated call volume densities. In a similar fashion K-Means can be applied as a data-mining technique to establish dynamic time frame clusters. The observed behavior of average hourly call volume suggests that hours with comparable volume levels on individual days of the week may be clustered together to potentially create a more representative feature set. Visually examining the data plotted in Figure 5 reveals, based on slope direction and inflection points, approximately six different daily call volume states. On each day the pattern of call volumes begins with a (1) steady decline towards a daily minimum, (2) reaches that minimum point and changes directions, (3) increases towards a daily maximum, (4) reaches a daily volume peak or near peak, (5) maintains a volume level at peak or near peak, and (6) declines into the next day. These six states are depicted in Figure 16, which visualizes the output from a preliminary K-Means cluster model using the average hourly call volume data for all days of the week aggregated over 24 hours.

To further evaluate this concept of sub-daily time frame forecasts in subsequent MLP models, the hourly call volume data was divided into seven separate datasets based on the day of the week (to account seasonality). The hourly call volume values for each day of the week were then averaged, prior to clustering, using two experimental concepts. For the first concept, *(Time-Cluster Concept A) the hourly call volumes for each day of the week are averaged based on the year and hour in which they occurred*. In the second concept,

(*Time-Cluster Concept B) the hourly call volumes for each day of the week are averaged based on the year, month, and hour in which they occurred.* This resulted in two sets of call volume averages, at different levels of aggregation, for clustering hours with similar volume levels. From there, the datasets were iteratively filtered based on the day of the week and processed individually through a K-Means clustering model.

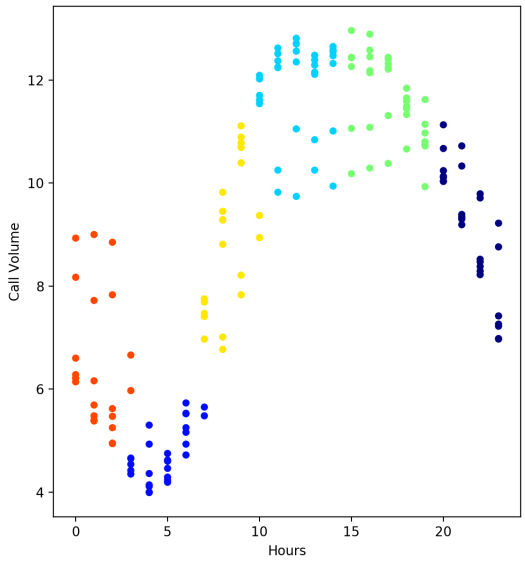

Figure 16: Avg. P1+P2 Avg. Hourly Call Volume Time Clusters (2010-2017)

Following convergence, a numeric cluster label was assigned to each hourly call volume record identifying the respective time frame cluster for each hour on a given day of the week. This process was repeated using both time-cluster concepts (TCC-A and TCC-B) datasets for multiple values of K (1-9). The average WCSS error was calculated for each cluster set, resulting in the elbow plot shown in Figure 17. Closely examining the elbow plot, 4-6 time-clusters appear to be the optimal range given the declining rate of change in the WCSS error. This supports the visual estimation that approximately six distinctly unique states of average call volume exist. A sample of the resulting K-Means clusters using TCC-A and K=6 time frame clusters are provided in Figure 18.

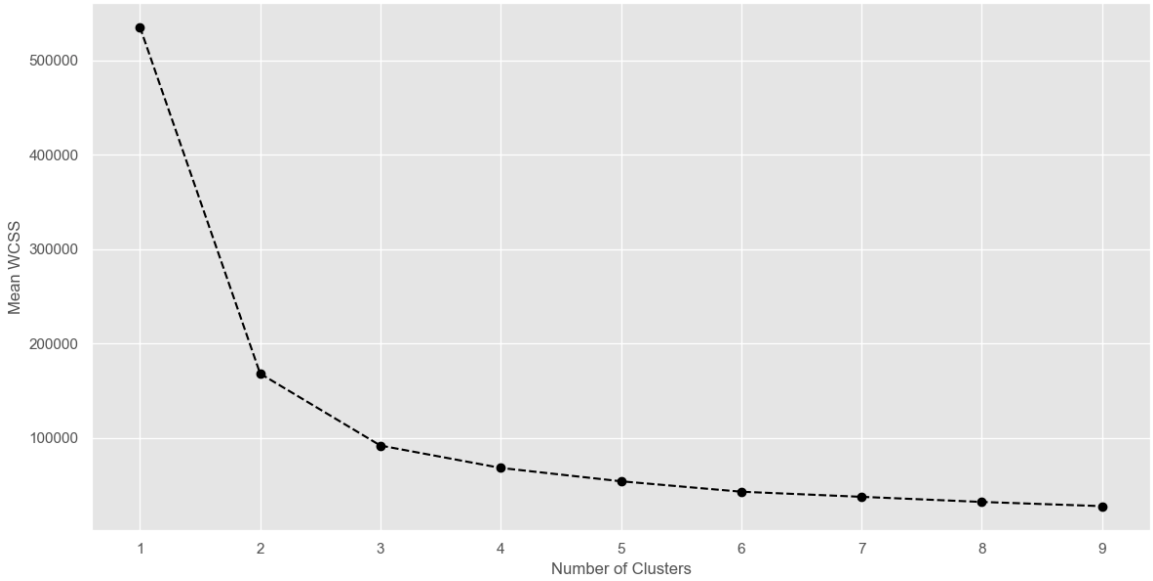

Figure 17: P1+P2 Avg. Hourly Call Volume Time Clusters Elbow Plot

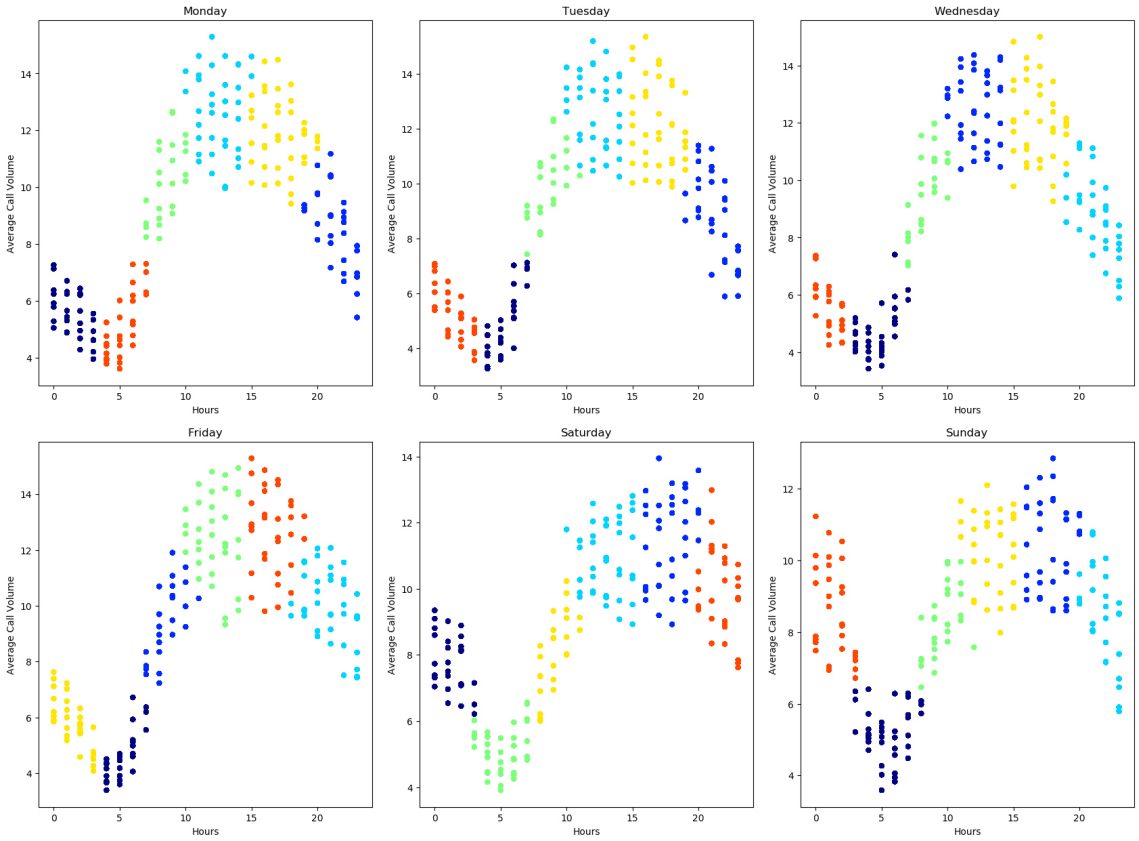

Figure 18: P1+P2 TCC-A Sample Time Clusters by Day of Week (2010-2017)

A.2 SPATIALLY DISTRIBUTED TIME FRAME CALL VOLUME FORECASTS

To evaluate the effectiveness of generating spatially distributed sub-daily time frame call volume forecasts using the two proposed time-cluster concepts, the prediction performance of each approach using K=6 time-frame clusters (TCC-A and TCC-B) is compared against a 3-hour fixed time frame approach [5, 16]. Based on the performance of spatially distributed hourly MLP models at varying levels of spatial granularity, 9-spatial clusters are used for comparative analysis of sub-daily time frame call volume forecasts. Non-spatially distributed forecasts are also generated for each time frame forecasting model. In practice, sub-daily time frame call volume predictions support an EMS agency's ability to identify peak volume time intervals based on the day of the week and support median-range workforce scheduling and fleet planning decisions.

A.3 SPATIALLY DISTRIBUTED TIME FRAME FEATURE SELECTION

Feature selection for non-spatial and spatially distributed sub-daily time frame (cluster and 3-hour fixed) call volume MLP models was carried out following the same process as before. With the addition of the time frame labels for each approach, the complete candidate feature set included: year, season, month, week, day, day of week, average daily temperature, minimum daily temperature, maximum daily temperature, precipitation in inches, binary variables indicating occurrences of daily rain and snow fall, the spatial cluster and time frame/cluster labels, and lag values of time frame call volume for periods 1-3, 6, 12, 18, 24, 30, 36, and 42. Processing each of the datasets separately, Boruta consistently rejected the snow feature and confirmed all other features as relevant for

estimating spatially distributed time frame EMS call volumes. For non-spatial time frame datasets, Boruta rejected the snow, rain, and precipitation features and confirmed all others.

A.4 ANN DEVELOPMENT FOR SPATIALLY DIST. TIME FRAME FORECASTS

MLP model development, hyperparameter tuning, and implementation for generating spatially distributed time frame call volume estimations was carried out following the same process used for previous models. Each model instance was trained using the input features confirmed during feature selection and the time frame call volume target values for each configuration. Following multiple rounds of hyperparameter tuning, the top performing MLP model parameter settings for each spatial and non-spatial configuration (TCC-A, TCC-B, and 3-hour fixed) were identified and recorded.

Table 13: Per Cluster MAPE/MAD Results by Time Frame Method (9 Clusters)

Cluster	1-HR MAPE	1-HR MAD	3-HR MAPE	3-HR MAD	TCC-A MAPE	TCC-A MAD	TCC-B MAPE	TCC-B MAD
0	38.04	0.78	25.01	1.40	40.63	1.67	28.05	2.04
1	39.56	0.92	24.29	1.62	39.07	2.00	27.32	2.23
2	34.89	0.65	24.29	1.20	44.19	1.41	26.84	1.71
3	45.72	1.30	25.10	2.33	32.24	2.85	26.91	3.31
4	31.39	0.61	23.30	1.09	44.75	1.30	26.03	1.58
5	32.07	0.67	23.26	1.20	43.51	1.44	26.58	1.71
6	43.24	0.94	26.51	1.68	39.20	1.96	28.29	2.29
7	44.36	1.07	24.32	1.85	35.03	2.28	27.56	2.64
8	21.57	0.46	22.29	0.94	48.72	1.09	25.18	1.33
Average =	36.76	0.82	24.26	1.48	40.82	1.78	26.97	2.09
Non-Spatial (County)	26.97	2.64	15.46	4.89	16.49	5.78	20.15	8.12

A.5 SPATIALLY DISTRIBUTED TIME FRAME FORECAST RESULTS

The resulting MAPE and MAD values for each time frame forecasting MLP model (spatial and non-spatial) are provided in Table 13. As a benchmark comparison, the hourly (1-hour) MLP model results from previous sections are included. Given the decreased level of temporal granularity, and related decreased estimation complexity, all three variations of

the time frame forecasting models evaluated, with the exception of TCC-A spatial model, outperformed the hourly (spatial and non-spatial) models. However, the time-cluster concept (TCC-A and TCC-B) models failed to outperform the comparatively naïve 3-hour fixed time frame approach. This indicates that the added complexity of clustering (grouping together) hours with similar average call volumes levels based on the day of the week, month, and year, does not increase the predictive performance for subsequent forecasting models. Lastly, the considerable reduction in forecast error (MAPE) between the 3-hour and 1-hour models suggest that utilizing a 3-hour planning horizon, versus the 1-hour industry standard, has the potential to significantly improve the effectiveness of short-term EMS deployment plans. Ideally, EMS agencies could institute rolling deployment and redeployment plans using a combination of daily, sub-daily (spatial and non-spatial), and hourly (spatial and non-spatial) forecasting models to maximize coverage and reduce call response times.

APPENDIX B: GEOSPATIAL ANALYSIS

To analyze the EMS call volume records spatially, a series of maps were constructed using the popular geographic information system (GIS) platform ArcMap. Given that each call record contains the precise latitude and longitude coordinates representing the call location, each record can easily be plotted on a map using GIS software. The primary purpose of this analysis is to visualize the spatial distribution of demand, identifying areas within Mecklenburg County, North Carolina that consistently experience high call volumes and, potentially identify seasonal patterns and trends related to time. Additionally, since early studies leveraged the availability of census information to identify causal links between EMS demand and socio-economic/demographic characteristics, a visual geo-spatial analysis is conducted via choropleth maps using a sample of census features explored in previous studies. The last decennial census conducted in The United States was completed in 2010, therefore, only the P1/P2 calls that occurred during 2010 will be included in this analysis. All P1/P2 calls that occurred throughout Mecklenburg County in 2010, aggregated at the census tract level, are displayed below in a choropleth map (refer to Figure A-1). The number of priority calls per census tract has been divided into six classification ranges using the quantile method in Arc Map. This ensures an equal distribution of the calls occurrences and simplifies visual analysis, providing clear visual distinction between the different quantile range classifications. Varying shades of green indicate the total volume of calls, lowest to highest, from lightest to darkest respectively. Examining the map, the majority of census tracts that consistently experience the highest call volumes (outlined in light blue) are predominantly positioned around the center of the county. This area represents the city of Charlotte and its various urban centers. More

specifically, the high-volume census tracts primarily fall to the west, east and north of the city's center. The main hospitals throughout the county have also been included as a layer on the map to provide reference points. The location of hospitals, relative to demand, is important to note as high priority patients are generally transported to the nearest emergency medical facility.

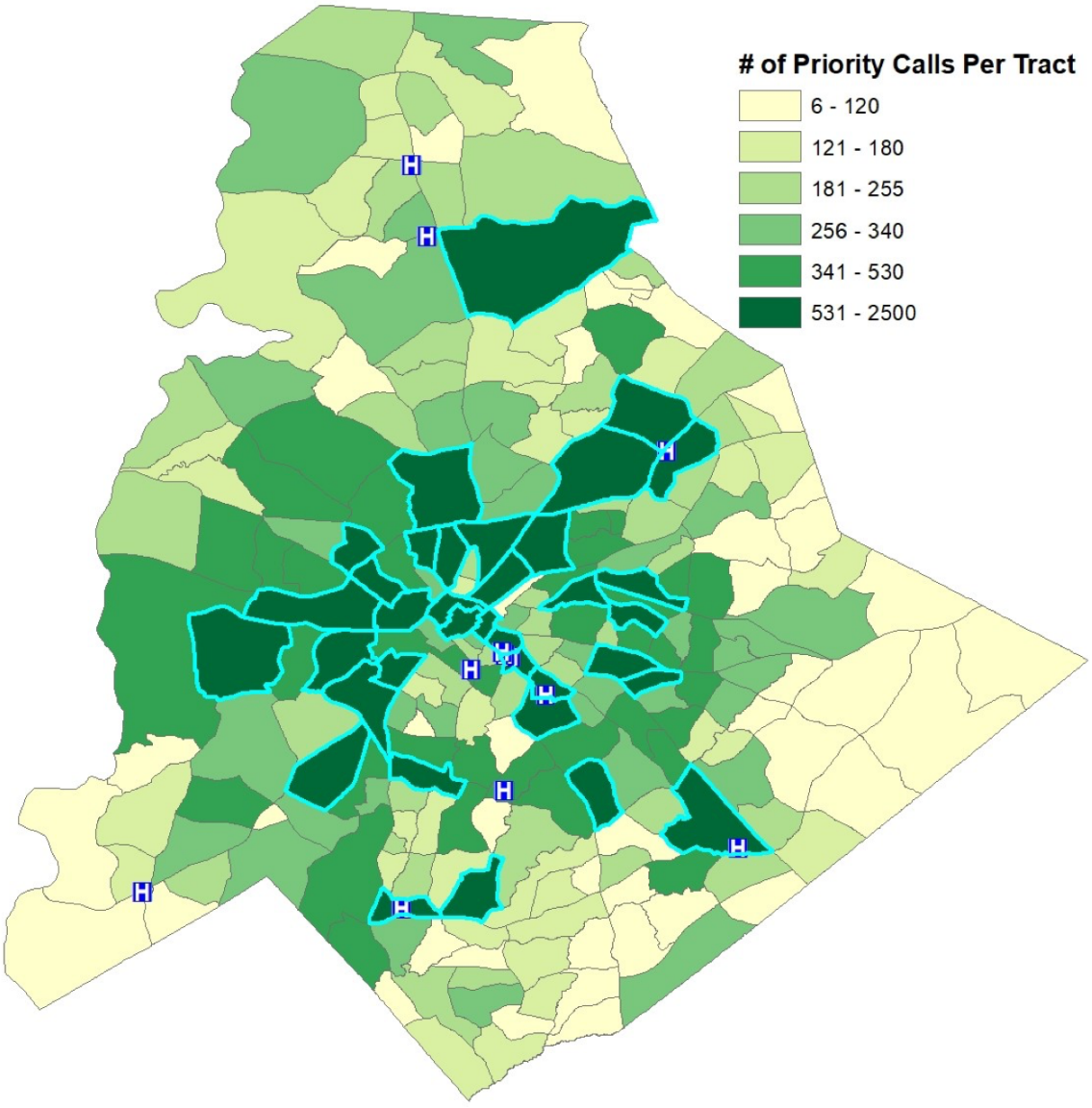

Figure A-1. P1+P2 EMS Call Occurrences by Census Tract (2010)

It has been established that EMS demand in Mecklenburg County follows a long-term positive linear trend with a distinct monthly and daily seasonal component. Figure A-2 displays the same dataset aggregated by census tract separated into annual quarters. Visually, there appears to be little variation from quarter to quarter, i.e. very few shifts in the volume level classifications of census tracts.

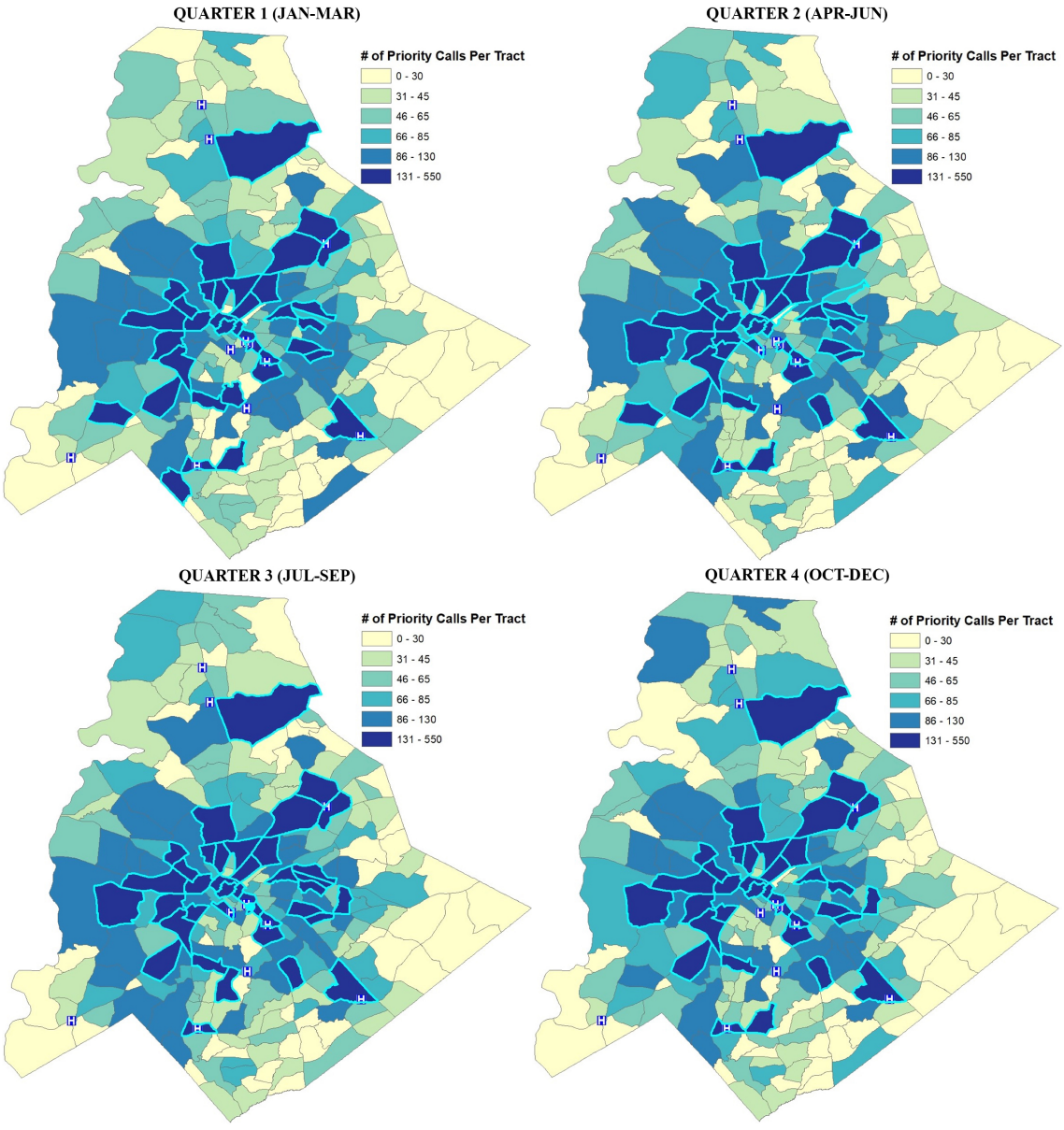

Figure A-2. Quarterly P1+P2 EMS Call Occurrences by Census Tract (2010)

The majority of tracts classified into the highest volume quantile, remain at that level each quarter of the year. This indicates that the seasonality observed from quarter to quarter over a given year has a minimal impact on the overall spatial distribution of demand. Figures A-3 and A-4 contain a series of choropleth maps depicting the number of total daily P1/P2 call occurrences separated by the day of the week. Each of the maps has been filtered to only display color in the census tracts that are classified into the highest two quantiles.

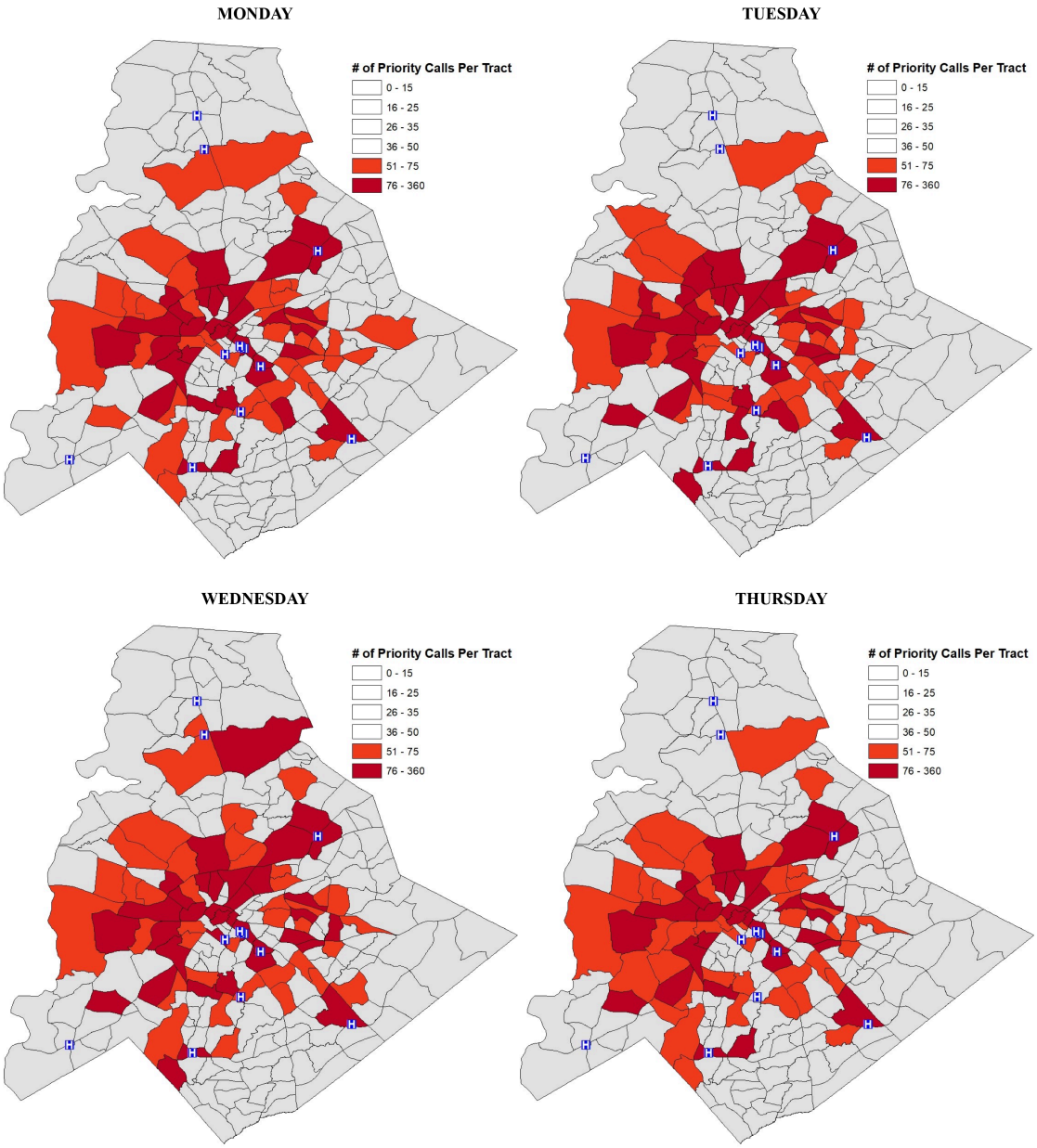

Figure A-3. Day of Week P1+P2 EMS Call Occurrences by Census Tract (Mon-Thu, 2010)

From this perspective, minor shifts in the high-volume census tracts are visible. This suggests that the day of the week has an influence on the spatial distribution of call volume.

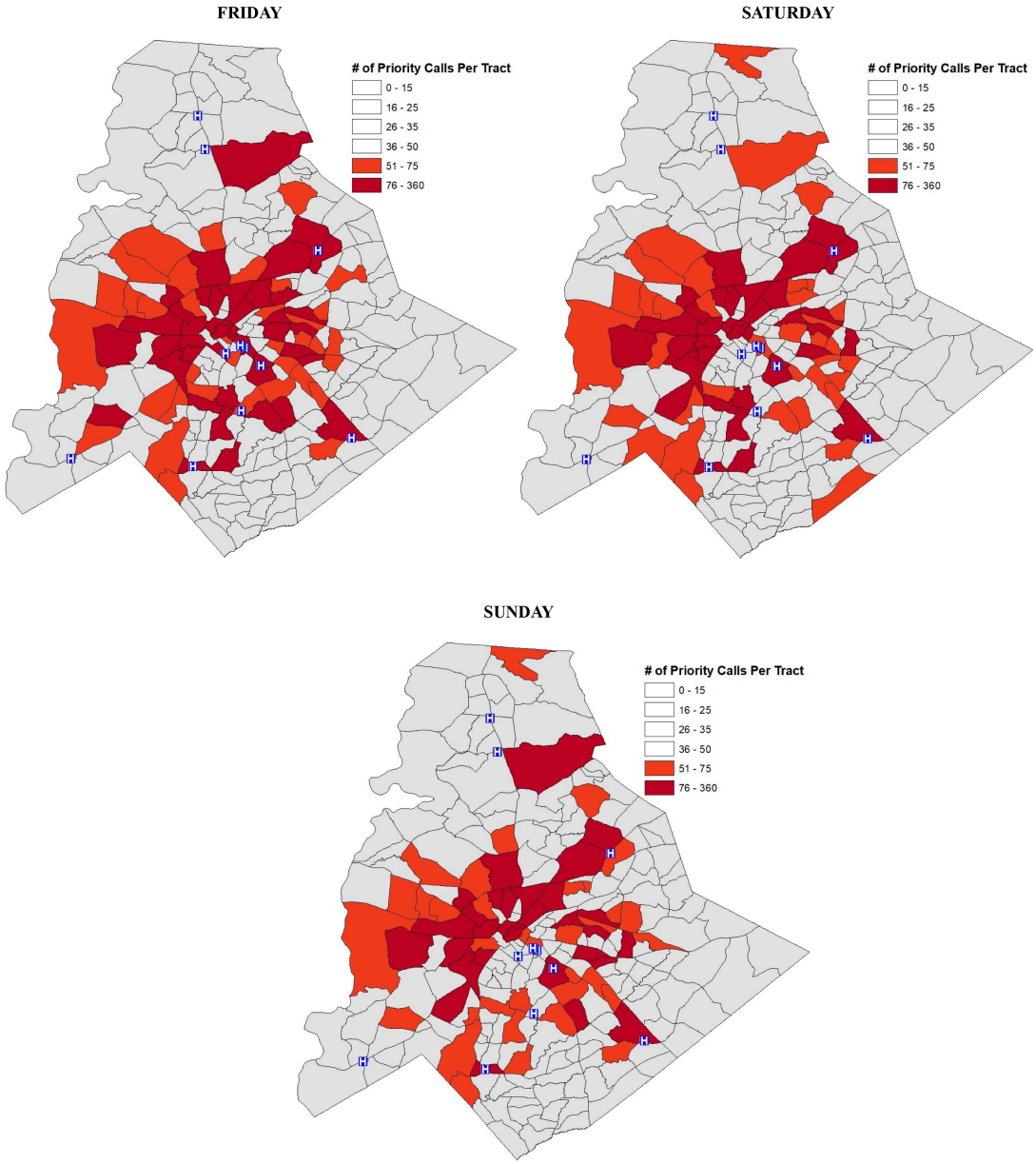

Figure A-4. Day of Week P1+P2 EMS Call Occurrences by Census Tract (Fri-Sun, 2010)

Additional EMS call volume choropleth maps of interest are provided below in Figures A-5, A-6, and A-7. These maps illustrate the distribution of calls related to traffic accidents, possible crime related calls (based on patient problem description, i.e. assault, stab wounds, and/or gunshot wounds), and patients who refused treatment or transport respectively.

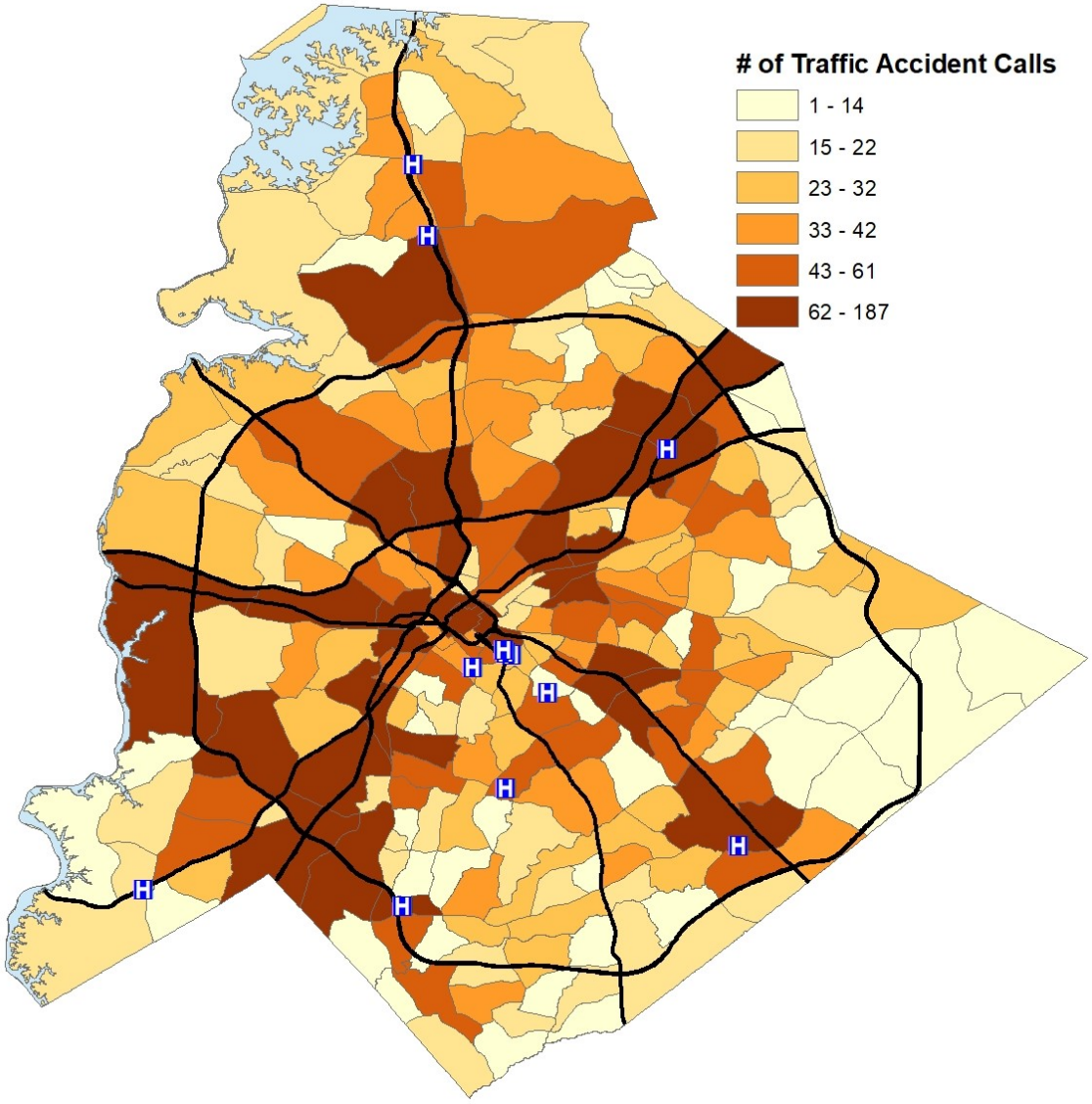

Figure A-5. Traffic Accident Related EMS Calls by Census Tract (2010)

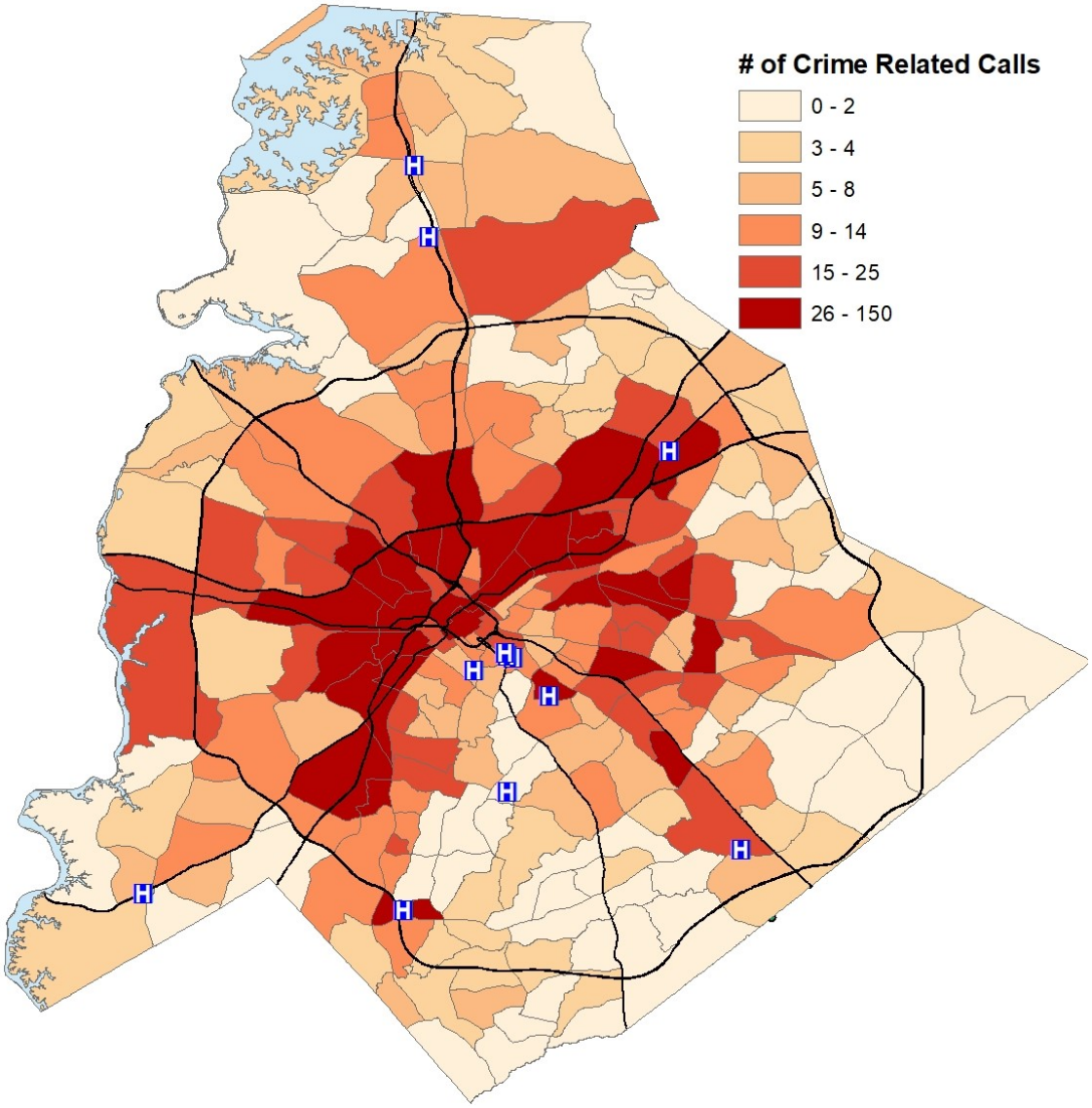

Figure A-6. Possible Crime Related EMS Calls by Census Tract (2010)

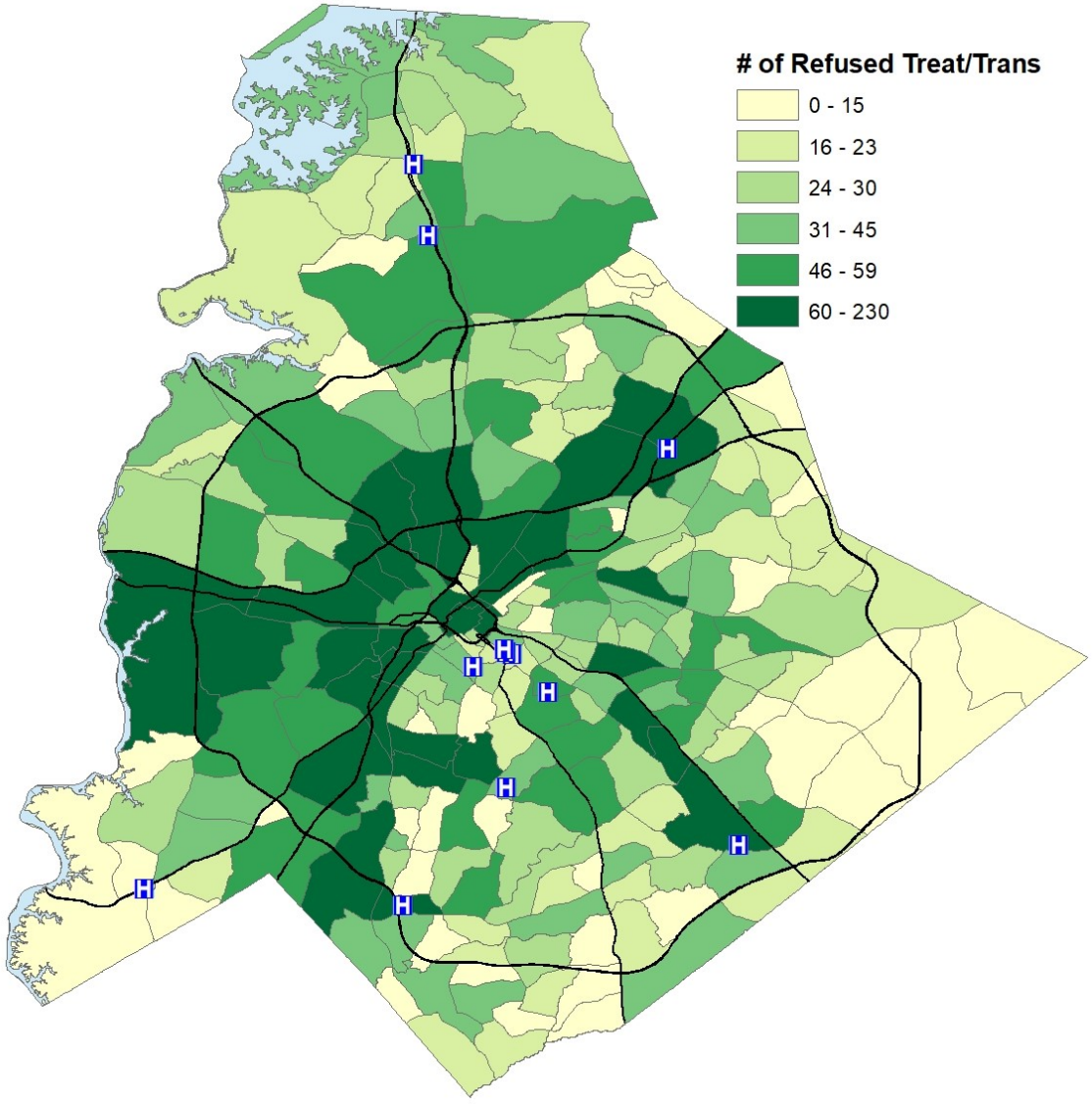

Figure A-7. Refused Treatment or Transport EMS Calls by Census Tract (2010)