GEOGRAPHIC APPLICATIONS OF KNOWLEDGE-RICH MACHINE LEARNING APPROACHES IN SPATIOTEMPORAL DATA ANALYSIS

By

Pouyan Hatami Bahman Beiglou

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Geography–Doctor of Philosophy

2021

ABSTRACT

GEOGRAPHIC APPLICATIONS OF KNOWLEDGE-RICH MACHINE LEARNING APPROACHES IN SPATIOTEMPORAL DATA ANALYSIS

By

Pouyan Hatami Bahman Beiglou

In the modern realm of pervasive, frequent, sizable and instant data capturing with advancements in instrumentation, data generation and data gathering techniques, we can benefit new prospects to comprehend and analyze the role of geography in everyday life. However, traditional geographic data analytics are now strictly challenged by the volume, velocity, variety and veracity of the data requiring analysis to extract value. As a result, geographic data science has garnered great interest in the past two decades. Considering that much of data science's success is formed outside of geography, there is an increased risk within such perspectives that location will remain simply as an additional column within a database, no more or less important than any other feature. Geographic data science combines this data with spatial and temporal components. The spatial and temporal dependence allow us to interpolate and extrapolate to fill gaps in the presence of inadequate data and infer reasonable approximations elsewhere by incorporating information from diverse data types and sources. However, within scientific communities there exist arguments regarding whether geographic data science is a scientific discipline of its own. Because data science is still in its early adoption phases in geography, geographic data science is required to develop its unique concepts, differentiating itself from other disciplines such as statistics or computer science. This becomes possible when geographers, within a community of practice, are enabled to learn and connect the current tools, methods, and domain knowledge to address the existing challenges of geographic data analysis. To take a step toward that purpose, in this dissertation, three knowledge-rich applications of data science in the analysis of geographic

spatiotemporal big datasets are studied, and the opportunities and challenges facing this research along the way are explored. The first chapter of this dissertation is allocated to review the challenges and opportunities in the era of spatiotemporal big data, followed by tackling three different problems within geography, one within the subfield of human geography, and two within physical geography. Finally, in the last chapter, some final thoughts on the current state of geographic data science are discussed and the potential for future studies are considered.

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my advisor, Dr. Lifeng Luo for his invaluable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. My deep gratitude extends to Dr. Pang-Ning Tan, my committee member from the department of computer science and engineering, for his guidance and support at every stage of the research projects. I was also honored to have Dr. Ashton Shortridge and Dr. Igor Vojnovic in my guidance committee, and I am genuinely thankful for their constructive comments and disciplinary insights for improving my research. I would like to thank my research group Tyler Wilson, Xi Liu and Jianpeng Xu, and Lisi Pei for helping me at different phases of my research. Finally, I would like to acknowledge the National Science Foundation (NSF) award numbers, 1615612 and 2006633, and the National Oceanic and Atmospheric Administration (NOAA) award NA17OAR4310132 for funding this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 SPATIOTEMPORAL BIG DATA: A SURVEY OF CHALLENGES AND OPPORTUNITIES

1.1 Introduction

Ongoing data growth has launched us into the 'Big Data' era, in which different types and formats of data resources are produced in many fields of study, due in part to advancements in instrumentation, data generation and data gathering techniques. Significant changes have been made to data gathering in terms of capacity, as well as the performance of instruments and equipment; calculations and archives from the 1970s to 2000 have been improved from 1-Dimensional to multi-Dimensional, and from megabyte to petabyte, respectively (Li et al., 2006). The surge of big data has impacted many commercial and scientific areas, and the field of geography has been no exception, moving from a data-poor to a data-rich era (Miller & Han, 2009). Geographical data have always been large resources, where most big data from phenomena of interest are recorded with stamps in three dimensions of space and one dimension in time, generally called big spatiotemporal data (Yang et al., 2020). The McKinsey Global Institute reported that location data was 1 petabyte in 2009 with a growth rate of 20% per year (Dasgupta, 2013); the United Nations Initiative on Global Geospatial Information Management (UN-GGIM) estimated 2.5 quintillion bytes of data are generated every day, and a large portion of the data is location-aware (Lee & Kang, 2015). The availability of big data has become more ubiquitous with improvements to measurement sensors (e.g., remote sensing satellites, mobile sensors), increases in computation power to run more and larger earth system simulation models, and crowdsourcing data that are generally publicly available (Dennis et al., 2012; Giachetta, 2015; Rice et al., 2012).

Geography is concerned with the study of Earth's physical structures and inhabitants spatially and temporally. Crucial challenges to Earth's inhabitants are naturally tied to study and modeling of physical features. Those challenges can include predicting climate change consequences, water

resources management, food security measurement, spread of disease during pandemics and recognition of contributing aspects in events like flood, drought, hurricanes, and earthquakes, among many others. The accessibility of spatiotemporal data delivers prospects for obtaining a new insight of complex geographic phenomena at macroscale and microscale. Furthermore, big data can facilitate innovations and productivity in various aspects of applications from hardware to software (Manyika et al., 2011). Big data was initially defined by the "3Vs": *volume, velocity* and *variety* by Laney (2001), which was then redefined to "5Vs" by adding *veracity* and *value* to it (Tiguint & Hossari, 2003; Zikopoulos et al., 2013). Li (2020) argues that although big data can be inherently beneficial for advancing science, obtaining an effective, time-sensitive and meaningful extraction of information presents some challenges. They note that volume (the size of data), velocity (high pace of data generation), variety (high data heterogeneity), and veracity (uncertainty and inadequate quality of data) are the challenges we face when extracting value in a spatiotemporal context.

As the importance of big spatiotemporal data have become clearer in recent years, more studies have been published on this topic. Yang et al. (2020) examined the number of articles in Web of Science published on topics containing related keywords, and found a rapid rise in the number of publications since 2009 (Figure 1-1). However, traditional spatial analysis techniques were established at a time when data were somewhat limited and computational capacity was not as powerful as it is today (Miller & Han, 2009; Yang et al., 2020). As a result, the capabilities of traditional data analysis methods show limitations, and spatiotemporal data analytics have become more challenging (Cheng et al., 2014; Yang et al., 2020).

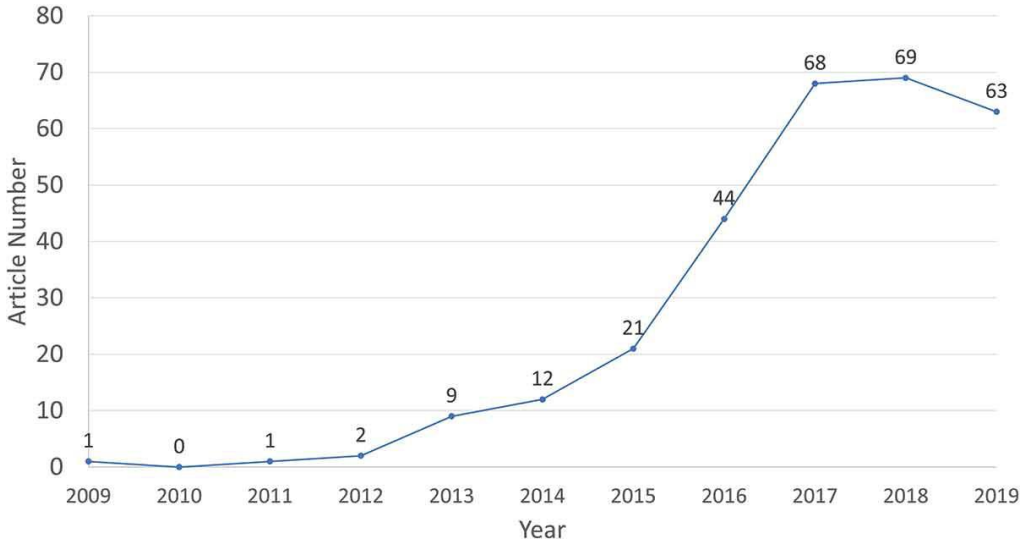


Figure 1-1 Ten-year history of the number of publications in the big spatiotemporal data area (Yang et al., 2020)

1.2 Big Spatiotemporal Data Sources, Properties and Challenges

1.2.1 Data Sources

Li (2020) organizes big geographic data into four typical sources: Earth observations, geoscience models simulations, Internet of Things (IoT) and volunteered geographic information.

1.2.1.1 Remote sensing

Earth observation refers to huge quantities of data obtained by remote sensing or sensing devices. Remote sensing data, which are measured over distances (e.g. radar, satellites, lidar) provided by space research organizations such as National Aeronautics and Space Administration (NASA), Japan Aerospace Exploration Agency (JAXA) and European Space Agency (ESA), delivers a worldwide history of geoscience variables such as land surface temperature, soil moisture and temperature at different spatial resolution and at consistent time intervals (Zhang, 2010). For specific studies on particular geographic areas of interest, devices such drones or airplanes can also be used as remote sensing methods (Frankenberg et al., 2016). Remotely sensed data are frequently captured throughout regularly spaced grid cells spatially and temporally, and the data is usually directly available, however, the time series often suffer from having a relatively short history of records.

1.2.1.2 In-situ Sensors

Another major supply of Earth observations is the in-situ sensors measuring at or near the Earth's surface, such as weather stations, or movement in the atmosphere or the ocean such as balloons, ships and ocean buoys (Bonnefond et al., 2011; Karpatne et al., 2017c). Sensor data, which are referred to as point reference data, are considered to be one of the most dependable sources of information about the geoscience variables. In-situ sensors are not evenly distributed in space and

time, however; there are inhomogeneities in the time series due to changes in the measurement sites (Horton et al., 2010).

1.2.1.3 Physical Models

One of the massive geographic data resources is produced by physics-based models known as geoscience simulation models. In these models, different elements of Earth system are simulated using laws of physics (e.g., first law of thermodynamics, Stefan-Boltzmann Law). The big data supplied by the simulation models have constantly been increasing in terms of volume, spatiotemporal resolution and coverage due to speedy progression of computing capacity. For example, Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) solely generated ten petabytes of simulated climate data (Li et al., 2017).

1.2.1.4 Internet of Things (IoT)

Internet of Things, also known as Internet of Objects, illustrates the network of everyday physical objects that are connected wirelessly via smart sensors and can work together without human interference (Ashton, 2009; Li et al., 2015). The devices that are uniquely identifiable, from sensors, cellphones, and smart factory equipment to connected appliances and wearable health monitors, can form an interconnected worldwide network for a new era of information. Objects combined with location-aware sensors, are capable of producing enormous sizes of spatiotemporal data. However, the amorphous flows of information across the globe can produce more heterogeneous and noisy data compared to more structured Earth observation data, which potentially confronts us with difficulties to take advantage of.

1.2.1.5 Volunteered Geographic Information (VGI)

Volunteered geographic information (VGI) is a term first introduced by Goodchild (2007), and in this context citizens are considered as sensors to participate in generating georeferenced data along

with other properties at their locations. Social media and crowdsourcing websites such as Twitter have proven to be beneficial during natural crises. Citizen Science, which is another form of VGI, involves a broad range of projects in which the public cooperates with scientists to acquire data for a national database. VGI has the advantage of being low cost compared to official data collection methods and significantly enhances geospatial databases. However, quality assurance is not guaranteed.

1.2.2 Spatiotemporal Data Properties and Challenges

Several common properties of spatiotemporal data that are widespread throughout many applications either arise out of the nature of spatiotemporal processes or are anticipated from the data collection techniques (Karpatne et al., 2018). Two natural and universal characteristics of spatiotemporal data that bring both challenges and opportunities for traditional data analysis algorithms are autocorrelation and heterogeneity (Shekhar et al., 2015). Conventional prediction methods assume samples hold an identical and independent distribution (i.i.d.) (Xie et al., 2017). In domains containing geographic spatiotemporal data the observations commonly reveal spatiotemporal autocorrelation because every phenomenon happens in space and time and most phenomena show short-scale dependence. This is known as Tobler's First Law of Geography (Miller, 2004). The presence of autocorrelation shows that the observations at nearby locations and time marks are correlated and cannot be considered independent; this can be problematic for classical data analysis methods as their assumption about independence between observations is not valid and can often consequently result in weak performance with salt-and-pepper errors (Jiang et al., 2014). The homogeneity, or stationarity, of observations indicates that every occurrence is from the same population and as a result has an identical distribution. However, spatiotemporal data are heterogeneous spatially and temporally at different levels. Additionally, non-stationarity

of the Earth system in time due to seasonal, decadal or long-term geological cycles can influence processes (Karpatne et al., 2018).

The rest of the spatiotemporal data properties, which are due to data collection procedures, can be identified as high dimensionality, absence of structured object definitions and boundaries, uncommon classes, multi-source and resolution data, poor data quality, insufficient sample size and in-situ observation (Karpatne et al., 2018). These properties may cause a variety of shortcomings in data-driven analysis and modeling. High dimensionality refers to the requirement to include numerous variables in the analysis due to complexity of the system. Object boundaries and their definitions are not as crisp spatially and temporally as are for common discrete spaces that data-driven methods usually deal with. Hurricanes are a good example for unclear object boundaries, as they continuously reshape in complex aspects during time. Extreme events such as heatwaves, which occur infrequently but cause significant impacts on society, are considered as uncommon categories of events. Multi-scale and resolution involve the integration of data with different sampling frequency, accuracy, as well as uncertainty within the system, as spatiotemporal information is often collected from various sources at different spatial and temporal resolutions (e.g., blending satellite images at different time intervals). Although this may appear to be a challenge, the analysis of multi-resolution geographic spatiotemporal datasets can assist in portraying processes that emerge on varying scales of space and time. Poor data quality is another characteristic of geographic spatiotemporal datasets, as many of them are subject to noise and missing values due to sensors failure, malfunctions or upgrades. Because of these differing levels of accuracy throughout time, establishing a consistent methodology of analysis is challenging. This is also the case for datasets generated by physics-based models as a result of the simplified representation of the system in the models as well as our imperfect knowledge of the initial and

boundary conditions of the system. Insufficient sample size presents yet another challenge in geographic analytical studies. Although there are many geographic datasets captured at high spatiotemporal resolutions, the historical datasets do not extend over a long time span or large spaces, which introduces additional challenges when adequate knowledge about the past and at some locations is unavailable. For instance, the majority of the satellite records available are relatively recent; satellite data has been captured since 1970s, and early records of precipitation are limited to land areas, and records for seas and oceans are lacking. Insufficient in-situ measurements may be considered to be another difficulty in geographic spatiotemporal datasets particularly when performing supervised learning problems. This is due to expensive and time-consuming procedures of high-quality data measurement which significantly limit the compilation of ground truth experiments. Some application processes, such as subsurface flow, do not have ground truth due to the system complexity, making it challenging to fully understand the state of the system. The lack of ground truth makes supervised models training, evaluation and testing difficult. This differs from commercial uses of data science, where significant quantities of labeled data have been essential for the success of machine learning methodologies.

Kanevski et al. (2008) recorded the difficulties that the typical characteristics of geospatial phenomena can place in front of the traditional data analysis algorithms. They recognized nonlinearity as geographic phenomena that may cause inadequate applicability in linear models; in many circumstances, spatial and temporal non-stationarity models can be in conflict with the hypotheses of spatiotemporal stationarity (second-order stationarity, inherent hypotheses), and the rest, including multi-scale variability, presence of noise and extremes/outliers, the multivariate nature fail practicality of traditional methods (including many geostatistical models) and extremely complex analysis, modelling and visualization of geographical data. Mennis and Guo (2009)

asserted that traditional methods for analysis regularly have several of the following limitations. First, generally current procedures emphasis on a limited perspective or a particular sort of relation model. Additionally, large volumes of data cannot be easily processed by traditional models. Finally, newly emerging data types such as trajectories of moving objects necessitate new tactics to analyze such data and uncover relationships and information.

Because of the fundamental shortcomings of the current methods due to complexity of geographic spatiotemporal datasets, there is a critical demand for more successful and efficient methods to uncover undetermined and unforeseen information. For this purpose, a new structure of information-rich systems with the use of new machine learning approaches offers the chance to meaningfully adjust geographic research procedures and acquire improved understanding from data (Gil et al., 2018).

1.3 Machine Learning

In recent years there has been unpredictable increase in the advancement of adaptive and data-driven methodologies in scientific communities; the geospatial perspective has been no exception. Today, machine learning (ML) provides important tools for intelligent geographical data analysis, processing, and visualization and is an essential aspect that complements traditional techniques like geostatistics (Kanevski et al., 2008). Research began when an overview of the topic written by Roddick and Lees (2001) brought the necessity of the study of ML into geographic information science. The use of ML in geography is often arranged under various names, such as spatial statistics, geo-computation, geo-visualization, and geo-spatial data mining, based on the procedures the research is centered on.

1.3.1 Framework

ML in a framework, that can be seen as a subfield of artificial intelligence, involved with the development, and application of algorithms and methods to let computers to discover the patterns from the data supplied. The ML process is inherently iterative (Andrienko & Andrienko, 1999), and is closely connected to nonparametric statistics. ML has grown from the simulated of a simple neuron and artificial neural networks to a solid, interdisciplinary field of fundamental and applied research with impact in many subjects (Kanevski et al., 2008). ML is an effective empirical tactic for both supervised and unsupervised learning of nonlinear systems that can be enormously multivariate containing from a few to thousands of variables (Lary et al., 2016). The use of ML is suitable for dealing with the obstacles where our theoretical understanding is still inadequate but a decent amount of data samples are available. There would not be a necessity for ML in a utopian world, if we had full theoretical understanding of phenomena.

1.3.2 Features

The most beneficial feature of the machine learning models/algorithms is their ability to learn the essential behavior of a system from training datasets. ML can be used in cases where the modeled phenomena and the nature of the relationships between variables is not well described, and we do not have or need prior knowledge about it, which is the case in many applications of geospatial data (Lary, 2010). Data-driven models built by ML are adaptive tools, which are broadly used to answer prediction, classification, optimization, and many other challenges. Lary et al. (2016) recorded three situations where the applications of ML in geoscience shines: (1) the use of a physics-based model is computationally costly, (2) no physics-based model exists but an empirical ML model may be developed using the available data, and (3) classification problems. Fayyad et al. (1996) listed two types of ML tasks: descriptive tasks, which describe the intrinsic

characteristics of the existing data, and predictive tasks that make an effort to achieve predictions based on inference from available data. Building a data model for the given dataset is the ultimate goal. The major tasks of ML in the analysis of spatiotemporal datasets incorporate regression, clustering, classification, and visualization, and the approaches designed to implement these tasks should consider spatiotemporal autocorrelation and heterogeneity, which differentiate them from older data mining procedures (Cheng et al., 2014).

Various tasks in ML such as regression, classification, association, clustering, ensemble learning, feature extraction, dimensionality reduction, principal component analysis (PCA), maximum likelihood estimation (MLE) fall within four particular learning approaches: supervised, unsupervised, semi-supervised and reinforcement learning. Common algorithms in ML include linear regression, logistic regression, Naïve Bayes, K-Nearest Neighbors, K-mean clustering, dimensionality reduction algorithms such as PCA and Factor Analysis, artificial neural networks (ANN), support vector machines (SVM), decision trees (DT), ensemble learning techniques such as random forests (RF), and etc.

In the next section, the applications that ML can be deployed will be discussed along with the challenges arising from the spatiotemporal data properties and the possible opportunities for the ML field for further advancement.

1.4 ML Applications, Challenges and Opportunities in Spatiotemporal Data Analysis

ML algorithms have been applied to analyze numerous application domains containing big spatiotemporal data. The domains can include urban studies such as crime prediction (Kim et al., 2018; Lin et al., 2018; Yu et al., 2020), infectious disease spread and control (Barratt & Sapp, 2020; Torrats-Espinosa, 2021; Valdes-Donoso et al., 2017), poverty distribution (Li et al., 2019; McBride et al., 2021; Vaz et al., 2021), transportation dynamics such as travel pattern (Hagenauer

& Helbich, 2017; Zhou et al., 2019), traffic dynamics (McCarthy, 2020; Rahman, 2020), environmental science such as air and water quality management (Chen et al., 2018; Lee et al., 2020; Ma et al., 2020; Muharemi et al., 2019; Wu et al., 2021), natural hazards (Resch et al., 2018) such as flood (Costache et al., 2020a; Costache et al., 2020b; Zhao et al., 2019), heatwaves (Park et al., 2020; Shi et al., 2021) and earthquakes (Akyol et al., 2020; Ghorbanzadeh et al., 2019), ecology such as land-use land cover classification (Talukdar et al., 2020) , and Earth system science such as climate science (Liu et al., 2018; Rolnick et al., 2019a; Xu et al., 2018), meteorology (Camporeale, 2019; Scher & Messori, 2018), ecosystem (Valerio et al., 2021; Willcock et al., 2018) and oceanographic (Hicks & Abuomar, 2019; Sonnewald et al., 2019) where a vast amount of spatiotemporal data are generated. In addition to the large amount of ML in geographic applications, there have been a number of research surveys in which the challenges and opportunities are discussed as a general standing of ML in geography or a specific domain. Table 1-1 briefly describes the different focused domains of surveys, as well as the opportunities and suggestion for covering the current gaps.

Table 1-1 A summary of surveys conducted on the use of spatiotemporal data

Study	Title keywords	Key points	Challenges and Opportunities
Kiwelekar et al. (2020)	Deep learning for geospatial data analysis	• Overview of DL algorithms • Geospatial analysis with data science • DL for analyzing remote sensing, GPS data and RFID data • CNN and Autoencoders vastly used in remote sensing and UAV in applications such as land use land cover • RNN along with CNN vastly used for GPS in applications such as traffic and mobility • CNN is used for RFID devices over smaller study areas	• Small sample size • Large number of objects in images to be detected
Joshi and Miller (2021)	Machine learning for mosquito control	• Reviewed 120 papers in ML techniques for mosquito control in urban areas • Geospatial, visual and audio models for mosquito control • Geospatial approaches use environmental factors on macro-scale for population modeling and prediction • Disease forecasting for dengue, malaria	• Use of citizen science and crowd-sourced data is essential in global awareness and prevention efforts • Open-source ML pipeline to use more private datasets and ability for model replication • Use of new techniques; transfer learning for local contexts, reinforcement learning for optimized resource distribution for mosquito control • Explore the use of online, active and multi-task learning • Integrate workspace between experts from multiple disciplines
Yekeen and Balogun (2020)	Advances in remote sensing, ML and DL in marine oil spill detection, prediction, and vulnerability assessment	• Reviews different oil spill detection by remote sensing methods • There is no single best remote sensing technique • Automatic detection techniques • ML classifiers for feature classification • SVM and ANN are the most used algorithms in oil spill detection	• Challenge of false positive appearance of similar oil spills in the imageries • Improve oil spill classification using ML and DL • Explore the use of image fusion methods • DL can help to develop a universal model for oil spill detection • No uncertainty measurement

Table 1-1 (cont'd)

Kovacs-Györi et al. (2020)	Geospatial analysis with big data and ML for promoting urban livability	• Reliability of crowdsourced and VGI data • ML in urban livability assessment and planning • Identification of relevant information in urban big data for livability progress	• Issue of users data privacy and ethics • Collaborative work between academia, stakeholders and policymakers is necessary • Data-driven approaches need to be combined with qualitative considerations
Singleton and Arribas-Bel (2019)	Geographic data science	• General review of the use of data science in geography	• There is a need to move from • Development of spatial databases and file formats for geographic Big Data • Data-driven geographic epistemology modeling (extension of scientific theories instead of testing the existing theories)
Atluri et al. (2017)	Spatiotemporal data mining	• General review of problems and views in spatiotemporal data mining	• New research methods in spatiotemporal data mining are needed • Novel representation of dynamic edges of Spatiotemporal raster data (as compared to existing methods focused on static edges) • Develop more multi-modal spatiotemporal datasets • Theory-guided data science is needed
Li et al. (2016)	Geospatial big data theories and methods	• General review and examination of the existing geospatial data handling methods and theories	• Data granularity is a challenge • Develop real-time modeling • Develop methods for explanatory relationships • Develop 3D spatial and 1D temporal displaying methods
Xie et al. (2017)	Transdisciplinary foundations of geospatial data science	• General review of data mining methods from mathematics, statistics and computer science perspectives	• Statistical strength of existing techniques needs improvement (p-value is not enough) • Transdisciplinary foundations instead of siloed (many techniques are strong from computational and mathematical perspective, but less statistical robustness) • Develop new techniques particularly for spatiotemporal data analytics

Table 1-1 (cont'd)

Yuan et al. (2020)	Deep learning in environmental remote sensing	• DL potential for tasks such as land cover mapping, environmental parameter retrieval, data fusion and downscaling, handling missing values • Popular DL algorithms in remote sensing applications	• DL cannot fully replace physical models • DL can be used for forward simulation of physical models to save computation cost • Physical model calibration with DL • Physics-guided DL architecture design • Combining geographical laws into DL such as introducing autocorrelation as input variable • Limited sample size can be addressed by transfer learning
Zhao and Tang (2018)	Crime in urban areas from a data mining perspective	• A review of theories in criminology and crime analysis algorithms	• Use of deep learning to better capture complex spatiotemporal patterns • Reinforcement learning to capture the dynamic nature of urban crime • Urban environment simulation to gain insights for policing strategies
Jain et al. (2020)	ML in wildfire science and management	• Fuel characterization, Fire detection and mapping • Fire weather and climate change • Fire occurrence prediction, susceptibility mapping and landscape control • Fire behavior prediction • Fire effects such as soil erosion or smoke level • Fire management	• ML has not been applied enough in predictive or optimization analytics • Deep learning can be used considering the vast amount of available data • Domain knowledge needs to be considered more • Wildfire is a diverse discipline so it needs a diverse analysis aspect
Moreno-Indias et al. (2021)	Statistical and ML in human microbiome studies	• Review of dimensionality reduction, clustering, classification, deep learning, association • Spatiotemporal modeling of microbiome as well as biogeographical variation	• Bayesian ML techniques can help to deal with uncertainties • Limited labeled data can be addressed using semi-supervised methods • Prospective analysis predicting long-term disease risks is still at early stages

Table 1-1 (cont'd)

Niu and Silva (2020)	Crowdsourced data mining for urban activities	• VGI and crowdsourced data sources • Urban activity types and analysis such as mobility pattern, functional areas and event detection • Sociodemographic and perception analysis such as city attractiveness and sentiment detection	• Challenges in inherent sampling bias and representativeness • Reliability of the data needs more attention • Multisource and multi-format data processing challenges

In order to cope with the challenges of analyzing spatiotemporal big data, ML has found its way and has proven to be helpful. Although ML models generated from data alone are not enough and new machine learning tactics that integrate domain knowledge will be essential so that achieved conclusions will be more meaningful than from data alone.

Gil et al. (2018) discussed three broad shortcomings of the current approaches due to complexity of the system, including domain theories in developing models instead of using the data alone, employment of more effective and efficient data collection by taking advantage of prior knowledge of the problem, and blending different data and models throughout different disciplines needs to be context intellectual to validate the combination. They also detail the challenges and opportunities for ML that appear before or during the analysis of spatiotemporal data, particularly in geoscience, and offer the existing or possible future research path.

One of the general prospects is the integration of domain knowledge into ML algorithms (Karpatne et al., 2017b) to reduce the phenomena complexities and nonlinearity in order to learn from smaller sample size. This approach is necessary due to the scarcity of labeled data and the presence of noise and missing values within the data. By integrating prior domain knowledge, it is possible to catch the underlying relationships among the variables with less data, and consequently, the complexity of the learning task is reduced. Active learning is an area of research in ML which can reduce the demand for labeled data by leveraging the information from areas with rich labeled data to the areas with few or no recorded data. However, this branch of ML is still in its infancy, and there is still a great deal to be studied and developed. Combining ML and physics-based simulation models known as hybrid modeling, is another tactic to avoid developing expensive physics-based models for the entire analysis to become more effective and efficient. Modeling of extreme events is already a challenge for simulation models with untrustworthy results. Likewise, this is currently

a challenge for ML algorithms due to spatiotemporal nature which needs to be studied further. Lack of ground-truth data poses additional challenges for supervised ML methods while estimating values because they are strongly dependent on benchmark values for evaluation during training the models. One possible solution could entail using simulation data during training, which provides an opportunity to train, evaluate and test ML algorithms. Causal discovery, which is the process of inferring the causal structure of a closed system using observational data, such as the cause of sea surface temperature and heatwaves, is another area that ML can be very effective by using graphical models, particularly in this era where there is a plethora data. A large array of ML methods can be efficiently applied to geoscience problems. Additionally, geoscience problems lead researchers to create completely new machine learning algorithms. Another challenging area – which is not from data, but from ML itself – is that ML algorithms are usually regarded as a black box with lack of interpretability, but so far have been acknowledged given their modeling accuracy. However, in geography it is required to be able to explain and interpret the models. An important research area is to integrate domain knowledge and causal inference to facilitate the structure of interpretive machine learning methods. Karpatne et al. (2018) reviews challenges and opportunities in four general types of problems in geoscience and includes in-detail ML solutions in addition to exploring the challenges along the way that ML can confront, as presented in Table 1-2.

Table 1-2 A Summary of the existing challenges and opportunities For ML in the analysis of geoscientific data by Karpatne et al. (2018)

Task	Example	Current Challenge	Solution by ML	Possible Challenges from Data Properties
Identifying objects and events	Cyclones, weather fronts, atmospheric rivers, ocean eddies	Conventional techniques are founded on hand-coded features	• Pattern mining techniques • Convolutional Neural Network	• Absence of structured object definitions and boundaries
Approximating variables	Methane concentration, groundwater seepage in soil	Difficult to monitor directly	• Supervised learning • Multi-task learning (to tackle joint effect of heterogeneity and paucity of ground-truth) • Online learning (in case of heterogeneity) • Downscaling variables (in case of heterogeneity) • Semi-supervised learning (in case of paucity of ground-truth)	• Heterogeneity, unidentified source of heterogeneity (changes in topography, land cover, season, etc) • Insufficient in-situ observation for developing different model for every homogeneous part • Uncommon classes in case of studying rare phenomena, poor data quality
Long-Term Forecasting	Temperature, greenhouse gas concentrations	Computationally expensive physics-based models	• Time-series regression (exponential smoothing, ARIMA, Markov models and Kalman filters)	• Insufficient sample size
Mining Relationships	Variation in sea surface temperature and ENSO and impacts on flood, droughts and wildfires	Study teleconnections	• Graph-based representations of locations • Causality-based network (Granger causality, Pearl causality)	• High dimensionality (high number of variables to include), insufficient sample size (limited number of years of data)

1.5 ML Is Not Perfect, but Necessary

Despite the capabilities of ML, there are always drawbacks and limitations along the way that must be tackled. ML, like anything else, does not always create a better world, but can become part of

the unraveling; is a capable mean that unlocks other paths and activates other tools across fields (Rolnick et al., 2019b). The problems that were discussed emphasize innovative areas of ML, such as interpretability, causality, and uncertainty quantification. However, profound action on real geographic challenges from a ML standpoint necessitates collaboration with fields inside and beyond computer science to move toward an interdisciplinary methodological innovation. As an example, Rolnick et al. (2019) studied the application of ML in climate change mitigation and adaptation from a diverse standpoint, combining the need to include versatile perspectives such as electricity systems, transportation, buildings, industry and land use in order to have a successful contribution of ML. They argued that ML can bring benefits to the scientific community in dealing with climate change by monitoring automation, expediting the progression of scientific findings, optimizing systems for better effectiveness, and expediting the computationally of expensive physics-based simulations through hybrid modeling with ML.

While utilizing machine learning is a key discipline for dealing with many challenges, there is also the potential to mutually benefit society and to improve the field of ML. In other words, the rising accessibility of big spatiotemporal data provides great possibility for ML to advance. Due to the growing number of successful results, ML has founded its valuable position in geographic conferences and journals. Yang et al. (2020) studied the emerging concepts through publications from the Web of Science, and discovered the top keywords were mostly related to *human dynamics*, *technology and methods*, such as spatiotemporal analyses, data mining, machine learning, deep learning, cloud computing, Hadoop/Spark, network, and *data and information* to assert that spatiotemporal data analytics has become a booming research route with a wide influence among diverse disciplines (Figure 1-2).

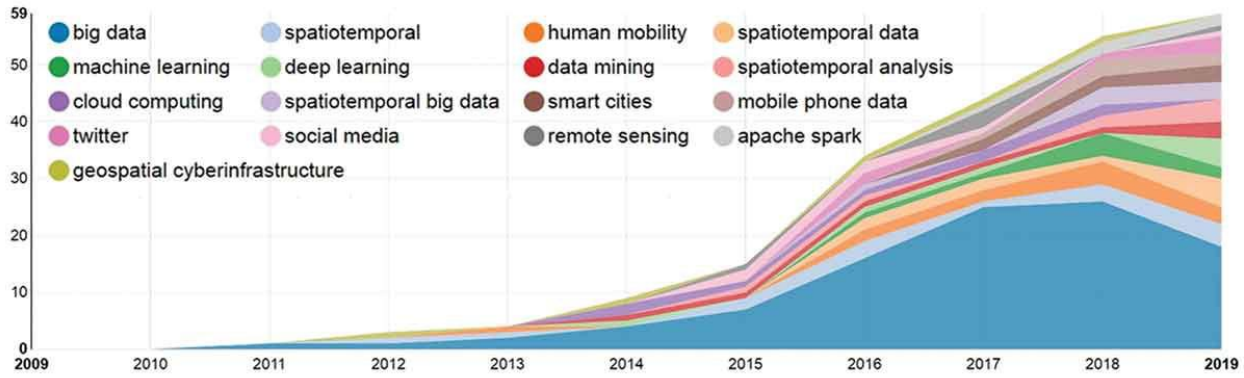


Figure 1-2 The emerging keywords obtained from the Web of Science publications (Yang et al., 2020)

1.6 Geographic Spatiotemporal Data Science Research Prospects

Gil et al. (2018) argues that a new research theme for the use of ML should consist of an "integrative workspace" where researchers from different backgrounds related to the study are able to communicate for a better understanding of the assumptions and uncertainties. These new crossing points and collaboration processes will sustain the discovery of data, as well as unearthing the knowledge to provide context to the data. This paradigm was called 'theory-guided data science (TGDS)" by (Karpatne et al., 2017a) where they argue that the popular commercial data science models have restricted applicability in scientific problems concerning physical phenomena. TGDS seeks to take advantage of the plethora of scientific knowledge available to expand the effectiveness of data science models in empowering scientific discovery. In other words, TGDS tries to lessen the challenges of physics only and data-only techniques by finding a balance between physics and data. For instance, to understand the Earth system, there is a need to combine broad information about the physical, geological, chemical, biological, ecological, and anthropomorphic elements that influence system by employing the most recent data science approaches. In a recent study, Karpatne et al. (2017d), presented a new framework blending physics-based models and deep learning methods, called physics-guided neural networks (PGNN)

to model lake temperature. In their work, the combined model was developed to use the output of physics-based models as the input of a deep learning architecture, where the results were significantly superior when compared to the absolute physics-based models or deep learning models. Additionally, the used loss functions in the proposed PGNN model evaluated the prediction to stay in accordance with the physics-based equations so it was capable of producing generalizable results as perfectly as physical models, even in the midst of scarcity in ground-truth data.

1.7 Conclusion and Dissertation Layout

Prospects in the data science analytics of geographic spatiotemporal datasets inspired by the challenges were discussed. Key skills are needed that require significant research in data collection and sampling, knowledge representation and integration, machine learning, and collaborative analytics to enable new findings. Being in the era of "big data", data science has formed as an interdisciplinary method that transforms large amounts of data into information. However, despite being common in other fields of science, data science is still in its early adoption phases in Geography. Considering that much of data science's success is formed outside of geography, there is an increased risk within such perspectives that location stays only as an additional column within a database, no more or less important than any other feature (Singleton & Arribas-Bel, 2019).

Such separation amid geography and data science, encouraged this research with an opportunity for pairing the two fields from a geographer's viewpoint and not that of a computer scientist. This research essentially takes one step forward to explore the challenges and opportunities that ML can encounter in three different problems within geography in order to continue constructing scientific ties between data science and geography. The first chapter is covered from a topic in the subfield of human geography and the next two are included from physical geography, particularly

focused on the area of hydroclimatology. The contributions of the three topics are going to better account some of the key challenges in building models with spatiotemporal big data. The proposed frameworks in all three studies are knowledge-rich, which means that they not only apply data science methods for improving the predictive power within the field of geography, but also extend our ability to integrate domain knowledge and causal inference to facilitate the shape of interpretive machine learning. As Graham and Shelton (2013) stated: "the futures of geography and big data are still to be made."

BIBLIOGRAPHY

BIBLIOGRAPHY

Akyol, A., Arikan, O., & Arikan, F. (2020). A machine learning-based detection of earthquake precursors using ionospheric data. *Radio Science*, *55*(11), 1-21.

Andrienko, G. L., & Andrienko, N. V. (1999). Data mining with C4. 5 and interactive cartographic visualization. User Interfaces to Data Intensive Systems, 1999. Proceedings,

Ashton, K. (2009). That 'internet of things' thing. *RFID journal*, *22*(7), 97-114.

Atluri, G., Karpatne, A., & Kumar, V. (2017). Spatio-Temporal Data Mining: A Survey of Problems and Methods. *arXiv preprint arXiv:1711.04710*.

Barratt, J. L., & Sapp, S. G. (2020). Machine learning-based analyses support the existence of species complexes for Strongyloides fuelleborni and Strongyloides stercoralis. *Parasitology*, *147*(11), 1184-1195.

Bonnefond, P., Haines, B., & Watson, C. (2011). In situ absolute calibration and validation: a link from coastal to open-ocean altimetry. In *Coastal altimetry* (pp. 259-296). Springer.

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166-1207.

Chen, S., Kan, G., Li, J., Liang, K., & Hong, Y. (2018). Investigating China's Urban Air Quality Using Big Data, Information Theory, and Machine Learning. *Polish Journal of Environmental Studies*, *27*(2).

Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., & Wang, J. (2014). Spatiotemporal data mining. In *Handbook of Regional Science* (pp. 1173-1193). Springer.

Costache, R., Bao Pham, Q., Corodescu-Roșca, E., Cîmpianu, C., Hong, H., Thi Thuy Linh, N., Ming Fai, C., Najah Ahmed, A., Vojtek, M., & Muhammed Pandhiani, S. (2020a). Using GIS, remote sensing, and machine learning to highlight the correlation between the land-use/land-cover changes and flash-flood potential. *Remote Sensing*, *12*(9), 1422.

Costache, R., Pham, Q. B., Sharifi, E., Linh, N. T. T., Abba, S. I., Vojtek, M., Vojteková, J., Nhi, P. T. T., & Khoi, D. N. (2020b). Flash-flood susceptibility assessment using multi-criteria decision making and machine learning supported by remote sensing and gis techniques. *Remote Sensing*, *12*(1), 106.

Dasgupta, A. (2013). *Big data: The future is in analytics*. Retrieved April 17 from https://www.geospatialworld.net/article/big-data-the-future-is-in-analytics/

Dennis, J. M., Vertenstein, M., Worley, P. H., Mirin, A. A., Craig, A. P., Jacob, R., & Mickelson, S. (2012). Computational performance of ultra-high-resolution capability in the Community Earth System Model. *The International Journal of High Performance Computing Applications*, *26*(1), 5-16.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Frankenberg, C., Thorpe, A. K., Thompson, D. R., Hulley, G., Kort, E. A., Vance, N., Borchardt, J., Krings, T., Gerilowski, K., Sweeney, C., Conley, S., Bue, B. D., Aubrey, A. D., Hook, S., & Green, R. O. (2016). Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region. *Proceedings of the National Academy of Sciences*, *113*(35), 9734--9739. https://doi.org/10.1073/pnas.1605617113

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, *11*(2), 196.

Giachetta, R. (2015). A framework for processing large scale geospatial and remote sensing data in MapReduce environment. *Computers & Graphics*, *49*, 37-46.

Gil, Y., Pierce, S. A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., Horel, J., Hsu, L., Kinter, J., Knoblock, C., Krum, D., Kumar, V., Lermusiaux, P., Liu, Y., North, C., Pankratius, V., Peters, S., Plale, B., Pope, A., Ravela, S., Restrepo, J., Ridley, A., Samet, H., Shekhar, S., Skinner, K., Smyth, P., Tikoff, B., Yarmey, L., & Zhang, J. (2018). Intelligent systems for geosciences: an essential research agenda. *Communications of the ACM*, *62*(1), 76--84. https://doi.org/10.1145/3192335

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255--261. https://doi.org/10.1177/2043820613513121

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert systems with applications*, *78*, 273-282.

Hicks, R., & Abuomar, S. (2019). Data Analytics for Clustering and Statistical Modeling of Oceanography Data. AGU Fall Meeting Abstracts,

Horton, R., Gornitz, V., Bowman, M., & Blake, R. (2010). Climate observations and projections. *Annals of the New York Academy of Sciences*, *1196*(1), 41-62.

Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, *28*(4), 478-505.

Jiang, Z., Shekhar, S., Zhou, X., Knight, J., & Corcoran, J. (2014). Focal-test-based spatial decision tree learning. *IEEE Transactions on Knowledge and Data Engineering*, *27*(6), 1547-1559.

Joshi, A., & Miller, C. (2021). Review of machine learning techniques for mosquito control in urban environments. *Ecological Informatics*, 101241. https://doi.org/10.1016/j.ecoinf.2021.101241

Kanevski, M., Pozdnukhov, A., & Timonin, V. (2008). Machine learning algorithms for geospatial data. Applications and software tools.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017a). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318-2331.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017b). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318--2331. https://doi.org/10.1109/tkde.2017.2720168

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2017c). Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544--1554. https://doi.org/10.1109/tkde.2018.2861006

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544-1554.

Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2017d). Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*.

Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018). Crime analysis through machine learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),

Kiwelekar, A. W., Mahamunkar, G. S., Netak, L. D., & Nikam, V. B. (2020). Deep Learning Techniques for Geospatial Data Analysis. In *Machine Learning Paradigms* (pp. 63-81). Springer.

Kovacs-Györi, A., Ristea, A., Havas, C., Mehaffy, M., Hochmair, H. H., Resch, B., Juhasz, L., Lehner, A., Ramasubramanian, L., & Blaschke, T. (2020). Opportunities and Challenges of Geospatial Analysis for Promoting Urban Livability in the Era of Big Data and Machine Learning. *ISPRS International Journal of Geo-Information*, *9*(12), 752. https://doi.org/10.3390/ijgi9120752

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Lary, D. J. (2010). Geoscience and Remote Sensing. *Geoscience and Remote Sensing: New Achievements*, 105.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3-10.

Lee, J.-G., & Kang, M. (2015). Geospatial big data: challenges and opportunities. *Big Data Research*, *2*(2), 74-81.

Lee, M., Lin, L., Chen, C.-Y., Tsao, Y., Yao, T.-H., Fei, M.-H., & Fang, S.-H. (2020). Forecasting air quality in Taiwan by using machine learning. *Scientific reports*, *10*(1), 1-13.

Li, D., Wang, S., & Li, D. (2006). Theory and application of spatial data mining. *Press of Science*.

Li, G., Cai, Z., Liu, X., Liu, J., & Su, S. (2019). A comparison of machine learning approaches for identifying high-poverty counties: Robust features of DMSP/OLS night-time light imagery. *International journal of remote sensing*, *40*(15), 5716-5736.

Li, S., Da Xu, L., & Zhao, S. (2015). The internet of things: a survey. *Information Systems Frontiers*, *17*(2), 243-259.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., & Stein, A. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119-133.

Li, Z. (2020). Geospatial big data handling with high performance computing: Current approaches and future directions. In *High Performance Computing for Geospatial Applications* (pp. 53-76). Springer.

Li, Z., Hu, F., Schnase, J. L., Duffy, D. Q., Lee, T., Bowen, M. K., & Yang, C. (2017). A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. *International Journal of Geographical Information Science*, *31*(1), 17-35.

Lin, Y.-L., Yen, M.-F., & Yu, L.-C. (2018). Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information*, *7*(8), 298.

Liu, X., Tan, P.-N., Abraham, Z., Luo, L., & Hatami, P. (2018). Distribution preserving multi-task regression for spatio-temporal data. 2018 IEEE International Conference on Data Mining (ICDM),

Ma, J., Ding, Y., Cheng, J. C., Jiang, F., Tan, Y., Gan, V. J., & Wan, Z. (2020). Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *Journal of Cleaner Production*, *244*, 118955.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition*.

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., Sun, Y., & Wen, J. (2021). *Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning*.

McCarthy, L. (2020). Spatio-temporal analysis and machine learning for traffic speed forecasting. *Signature*.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, *33*(6), 403-408.

Miller, H. J. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, *94*(2), 284--289. https://doi.org/10.1111/j.1467-8306.2004.09402005.x

Miller, H. J., & Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC Press.

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., Aydemir, O., Bakir-Gungor, B., Santa Pau, E. C.-d., & D'Elia, D. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Frontiers in microbiology*, *12*, 277.

Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294-307.

Niu, H., & Silva, E. A. (2020). Crowdsourced data mining for urban activity: Review of data sources, applications, and methods. *Journal of Urban Planning and Development*, *146*(2), 04020007.

Park, M., Jung, D., Lee, S., & Park, S. (2020). Heatwave Damage Prediction Using Random Forest Model in Korea. *Applied Sciences*, *10*(22), 8237.

Rahman, F. I. (2020). SHORT TERM TRAFFIC FLOW PREDICTION USING MACHINE LEARNING-KNN, SVM AND ANN WITH WEATHER INFORMATION. *International Journal for Traffic & Transport Engineering*, *10*(3).

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362-376.

Rice, M. T., Paez, F. I., Mulhollen, A. P., Shore, B. M., & Caldwell, D. R. (2012). *Crowdsourced Geospatial Data: A report on the emerging phenomena of crowdsourced and user-generated geospatial data*.

Roddick, J. F., & Lees, B. G. (2001). Paradigms for spatial and spatio-temporal data mining. *Geographic data mining and knowledge discovery*, 33-50.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., & Waldman-Brown, A. (2019a). Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., & Bengio, Y. (2019b). Tackling Climate Change with Machine Learning. *arXiv*.

Scher, S., & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830-2841.

Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V., & Zhou, X. (2015). Spatiotemporal data mining: A computational perspective. *ISPRS International Journal of Geo-Information*, *4*(4), 2306-2338.

Shi, Y., Ren, C., Luo, M., Ching, J., Li, X., Bilal, M., Fang, X., & Ren, Z. (2021). Utilizing world urban database and access portal tools (WUDAPT) and machine learning to facilitate spatial estimation of heatwave patterns. *Urban Climate*, *36*, 100797.

Singleton, A., & Arribas-Bel, D. (2019). Geographic Data Science. *Geographical Analysis*. https://doi.org/10.1111/gean.12194

Sonnewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, *6*(5), 784-794.

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.-A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote Sensing*, *12*(7), 1135.

Tiguint, B., & Hossari, H. (2003). Big data analytics and artificial intelligence: a meta-dynamic capability perspective. *changes (Eisenhardt & Martin, 2000)*.

Torrats-Espinosa, G. (2021). Using machine learning to estimate the effect of racial segregation on COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, *118*(7).

Valdes-Donoso, P., VanderWaal, K., Jarvis, L. S., Wayne, S. R., & Perez, A. M. (2017). Using machine learning to predict swine movements within a regional program to improve control of infectious diseases in the US. *Frontiers in veterinary science*, *4*, 2.

Valerio, C., De Stefano, L., Martínez-Muñoz, G., & Garrido, A. (2021). A machine learning model to assess the ecosystem response to water policy measures in the Tagus River Basin (Spain). *Science of The Total Environment*, *750*, 141252.

Vaz, E., Bação, F., Damásio, B., Haynes, M., & Penfound, E. (2021). Machine learning for analysis of wealth in cities: A spatial-empirical examination of wealth in Toronto. *Habitat International*, *108*, 102319.

Willcock, S., Martínez-López, J., Hooftman, D. A., Bagstad, K. J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., & Voigt, B. (2018). Machine learning for ecosystem services. *Ecosystem services*, *33*, 165-174.

Wu, J., Song, C., Dubinsky, E. A., & Stewart, J. R. (2021). Tracking Major Sources of Water Contamination Using Machine Learning. *Frontiers in microbiology*, *11*, 3623.

Xie, Y., Eftelioglu, E., Ali, R. Y., Tang, X., Li, Y., Doshi, R., & Shekhar, S. (2017). Transdisciplinary Foundations of Geospatial Data Science. *ISPRS International Journal of Geo-Information*, *6*(12), 395. https://doi.org/10.3390/ijgi6120395

Xu, J., Liu, X., Wilson, T., Tan, P.-N., Hatami, P., & Luo, L. (2018). MUSCAT: Multi-Scale Spatio-Temporal Learning with Application to Climate Modeling. IJCAI,

Yekeen, S. T., & Balogun, A.-L. (2020). Advances in Remote Sensing Technology, Machine Learning and Deep Learning for Marine Oil Spill Detection, Prediction and Vulnerability Assessment. *Remote Sensing*, *12*(20), 3416. https://doi.org/10.3390/rs12203416

Yu, H., Liu, L., Yang, B., & Lan, M. (2020). Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method. *ISPRS International Journal of Geo-Information*, *9*(12), 732.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., & Wang, J. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716.

Zhang, J. (2010). Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, *1*(1), 5-24.

Zhao, G., Pang, B., Xu, Z., Peng, D., & Xu, L. (2019). Assessment of urban flood susceptibility using semi-supervised machine learning model. *Science of The Total Environment*, *659*, 940-949.

Zhao, X., & Tang, J. (2018). Crime in urban areas: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *20*(1), 1-12.

Zhou, X., Wang, M., & Li, D. (2019). Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning. *Journal of transport geography*, *79*, 102479.

Zikopoulos, P. C., Deroos, D., & Parasuraman, K. (2013). *Harness the power of big data: The IBM big data platform*. McGraw-Hill.

Chapter 2 FINE-RESOLUTION PREDICTION OF THE NUMBER OF CRIMES USING MULTI-TASK LEARNING

2.1 Introduction

Crime is a ubiquitous social problem that could potentially become more serious as urbanization increases. Urbanization helps growth of industries and economic development; however, one of the drawbacks of urbanization may be the boost of crime occurrence as well because crimes happen more frequently in large cities (Malik 2016). Factors such as size, density, heterogeneity and impersonality of urban areas and the tendency toward crime have been studied as undeniable evidence for the connection between urbanization and more crime (Wirth 1938, Wirth 1964, Clinard 1942). Furthermore, crime can affect the life quality of a society; it may influence opportunities for new investments, tourism, or other aspects of the economy (Arulanandam, Savarimuthu and Purvis 2014).

The importance of safety has led law enforcement agencies to demand scholars and practitioners to focus on crime prevention by improving crime analytics and predictions. However, it is a complex phenomenon to give a comprehensive cause for the crime (Weatherburn 2001) and in contrast to many foreseeable events, crime is sparse (Wang et al. 2017). The associated distinctiveness and arbitrariness with crime makes the prediction a difficult task, though, there are patterns (Gorr and Harries 2003)**.**

Researchers have developed various crime prediction frameworks using different statistical to machine learning techniques along with combination of multiple data sources in addition to historical crime data. Crime occurrence is a multi-dimensional phenomenon associated with temporal, spatial, societal, and ecological factors (Yu et al. 2014) so the research community have been attempting to create more accurate predictive models with the assistance of different types of

data sources besides historical crime. Initially, crime prediction models mostly relied on demographics as the only additional data tributary, however, due to its failure to obtain the dynamic aspects of human activity (Zhao and Tang 2017), they gradually moved toward contributing other data sources into the analysis. New data streams such as weather (Chen, Cho and Jang 2015), social media (Chen et al. 2015, Wang and Gerber 2015, Gerber 2014), Point-of-Interests (POI) (Wang et al. 2016), transit flow (Smith, Quercia and Capra 2013, Kadar and Pletikosa 2018), mobile data (Bogomolov et al. 2014) have been used either solely or combined in the crime prediction models.

Since crime does not occur randomly, and the frequency of crime occurrence tends to be correlated with the location of victims, offenders, and the opportunity of committing crime (Chainey, Tompson and Uhlig 2008), hotspot mapping became popular among researchers (Gerber 2014, Gruenewald et al. 2006, Yang et al. 2017, Das and Choudhury 2016). Hotspots are the areas with higher concentration of crime events compared to the rest of the study region and historical data show that crimes do occur in concentrated patterns (Chainey and Ratcliffe 2013). Crime hotspot prediction uses historical data to detect geographical areas vulnerable to crime events in the future. The drawback of most of the frameworks in crime hotspot prediction is that it is limited to the employment of historical crime records (Yang et al. 2017, Wang and Gerber 2015), while disregarding the use of other types of data such as environmental factors and urban data. A large number of hotspot mappings (Liu and Brown 2003, Xue and Brown 2006, Brown, Dalton and Hoyle 2004) solely focused on the spatial distribution of crimes, whereas knowing the temporal likelihood of crime occurrence is needed for tactical purposes including for urban planning and police protection. In other words, from the predictive perspective, temporally aggregated hotspots which are relying on shorter prior time periods is less operative (Groff and La Vigne 2002).

Additionally, hotspot models are not generalizable to ranges without historical data (Mookiah, Eberle and Siraj 2015).

Similar to spatial crime studies, there are a great deal of spatiotemporal crime statistics studies as well in which they quantify the count or rate of crime variation over space and in each of the time periods under study (Chainey and Ratcliffe 2013). A majority of the previous studies suffer, as they do not take spatial dependencies and heterogeneity into consideration (Liu 2017). In reality, crime rate and type fluctuates from region to region while existing crime prediction models have not accounted for this variation, resulting in a preference for global models that compensate for low resolution (Yu et al. 2014).

To fill this gap, for the first time, we attempt to capture the related spatial and temporal information in crime prediction using a multi-task learning model with a Graph Laplacian regularization. Essentially, we aim to examine the spatiotemporal crime-prediction performance of a multi-task learning method against linear local models and global models in the role of the commonly used crime prediction methods. We also contribute a new combination of variables in the modeling which could represent the social, environmental and ecological factors in crime occurrence. The rest of the paper is structured as follows. We discuss the data and the methods in Section 2.2. In this section, we explain and justify the use of local, global and multi-task learning methods in our analysis. In Section 3, we present and discuss the results of each step and finally, in Section 4, we conclude this paper with a brief discussion of the significance of the results.

2.2 Data and Methods

This section presents the data, the method used, and the description of algorithms and metrics.

2.2.1 Data

The selected study area in this research was New York City (NYC), which is the most populous city in the United States. The population of the city in 2016 was estimated about 8.5 million over a land area of about 303 square miles. NYC is an important city and its known as a global city due to its significant political and socio-economic impact around the world (Sassen 2016). Hence, given its socio-demographic profile, NYC could provide a data rich location with abundant available information which would help us to better understand the potential underlying factors in crime occurrence.

To examine this research question, we obtained twelve years of historical crime data of NYC from www.data.cityofnewyork.us, which is recorded by the NYC Police Department. The record spans from January 2005 through December 2016 with 5,002,053 incidents in which for every occurrence includes date and time, offense description, law enforcement offense category (i.e. violation, misdemeanor, felony), borough, precinct, latitude, and longitude. NYC is composed of five boroughs - Manhattan, Brooklyn, Queens, the Bronx and Staten Island - and 77 Police Precincts. To build better predictive models, we collected additional information which could have a significant or near-significant relationship with crime occurrence. The information included demographics, daily weather data and zoning districts of NYC, which have been previously noted as useful statistics for crime control and prevention (Flowers 1989, Cohn 1990, Horrocks and Menclova 2011, Poulsen and Kennedy 2004). Demographics were downloaded from www1.nyc.gov for all five boroughs from 2005 to 2016 including the population of male, female, white, black, Indian American and Alaska Native, Asian, Hawaiian, Hispanic, other races, and the

number of total households. Weather data including daily summary of precipitation, snowfall, wind, and average temperature was obtained from NCDC (National Climatic Data Center)[1] for the time span of January 1, 2005 through December 31, 2016 from three land-based stations located in NYC. Among all available weather stations located inside or close to the border of NYC, only three stations covered the desired data range. Lastly, zoning data which included residential, commercial, park, and manufacturing areas of the city was obtained from www1.nyc.gov. Figure 2-1 shows the map of NYC zoning districts and police precincts.

[1] http://www.ncdc.noaa.gov/

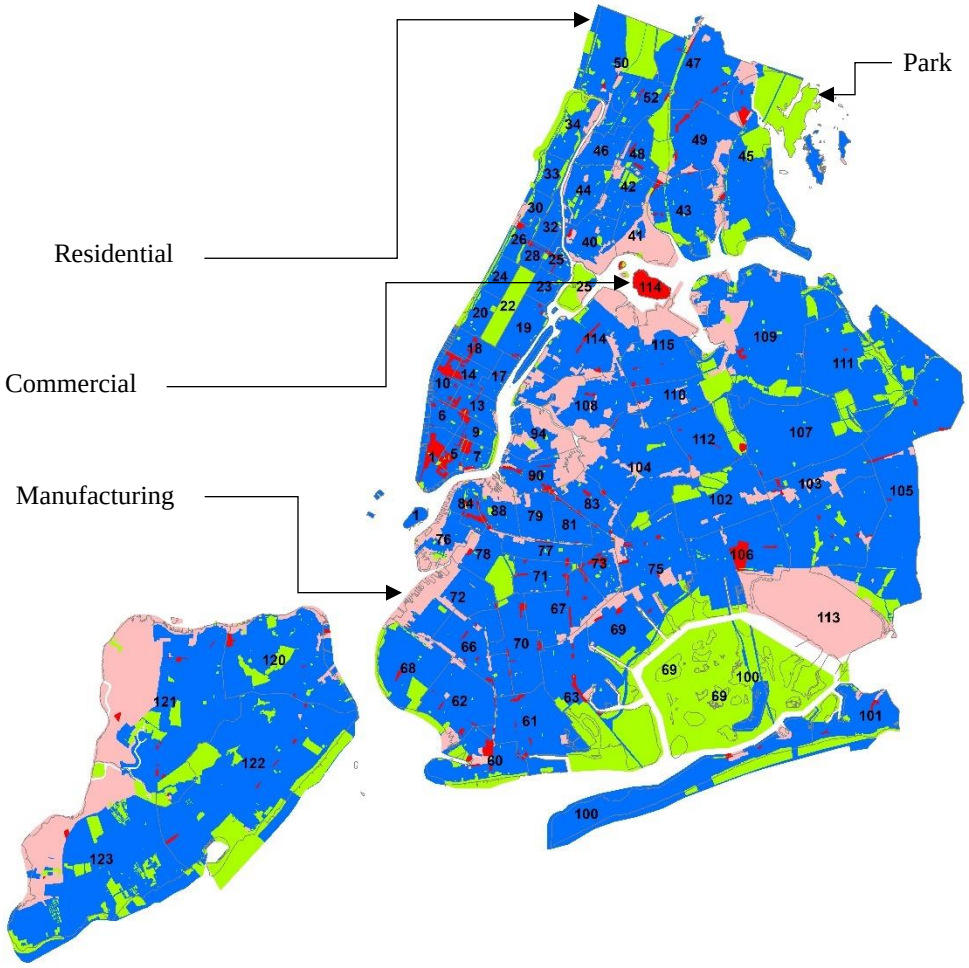


Figure 2-1 Geographic boundary of NYC with Zoning Districts and Police Precincts Numbers (Polygons in Blue, Red, Pink, Green are Residential, Commercial, Manufacturing and Park, respectively)

2.2.2 Methods

The methods considered in this study include three approaches to predict the number of crime occurrence; 1) linear local models, 2) a linear global model and 3) local models built by a multi-task learning (MTL) method. First, we need to define the temporal and spatial resolution of the study in an optimal resolution, which was a trade-off between the available data and the practicality of crime prediction. The practicality of prediction means that the temporal and spatial resolution

should be somehow informative enough if there is a need to know the probability of crime occurrence at a specific time and location. Toward this purpose, based on the available data, the temporal and spatial resolutions of the study were designed to be a 6-hour window of time and the police precincts, respectively. Therefore, we split a day into four 6-hour time-frames: 12 a.m. to 6 a.m., 6 a.m. to 12 p.m., 12 p.m. to 6 p.m., and 6 p.m. to 12 a.m. This means that using the models, we would be able to predict the number of crimes in a desired precinct for each quarter of a day.

As this study was the examination of MTL as an emerging machine learning method against traditional statistical methods in spatiotemporal crime prediction, all types of reported incidents were stacked as only one crime type. The goal was to establish MTL to aid future research into crime prediction which could be improved upon the current one. In the following sections, we begin to explain all three approaches along with the data preprocessing steps.

2.2.2.1 Local models

To examine this research question, as the initial investigation, we started with building local models for all 77 precincts of the NYC as the traditional single-task learning. A local model is an isolated model which is created specifically for a precinct with its data, regardless of the information in other existing precincts. To this end, we conducted preprocessing to create a suitable format with the data that was already collected. Since variation in precincts' demographics and zoning information were almost static during the twelve years of data, we did not reflect them as local model features; however, they were being featured in the global model that will be explained in the next section. Because a global model takes the data of precincts all together, the variation in demographics can be meaningful for the model to better learn the attributes of the precincts while they do not make a change in a local model by zero variation.

The taken preprocessing steps are as follows: 1) extract twelve years data of every 77 precincts from the entire historical crime data, 2) count the number of crimes for 6-hour time window based on the occurrence time of the crimes from January 1, 2005 to December 31, 2016 (17532 cases), 3) assign a number from 1-4 for the time, 1-7 for the day of week, 1-53 for the week of the year to each case, 4) add weather features from the closest weather station to the geographical coordinates of the crime for each case. Precipitation and snowfall were designed as a binary classification as rainy and non-rainy days and snowy and non-snowy days, 5) add the moving average of the number of crimes in the past 7 days, past 14 days and past 21 days of every case in order to investigate the crime periodicity influence in crime occurrence (Zhao and Tang 2018). To create these features, the first 21 days of the records were eliminated and the first day of our crime record with 21 days behind was day 22 of the year 2005.

As previously discussed, every precinct has its own specific model which means there are 77 local models for the 77 precincts of NYC. In the Modeling section, we will discuss the steps taken to build the local models.

2.2.2.2 Global Models

In contrast to a local model, a global model is a model which is built using information from all precincts. In this study, as another examination of single-task learning methods, we built two global models. The first one was with the same exact information as for the local models in the previous sections, however, the only difference was adding precinct number variables so that we could have the spatial predictive feature in the global model. In other words, for every precinct we had a model in local modeling whereas with a global model, there is only one single model for all the precincts for prediction, therefore we needed to define a spatial feature in addition to temporal one so that the model distinguishes the locations.

The other global model was created using the data with the extra information which were not used in the local models because of close-to-zero variation for twelve years. That information were demographics (population of Male, Female, White, Black, American Indian/Alaska Native, Asian, Hawaiian, Hispanic, Other races, Total households), boroughs (Manhattan, Brooklyn, Queens, the Bronx and Staten Island), and zoning areas (Residential, Commercial, Manufacturing and Park). As the final step, we aggregated precincts' crime data and joined the additional data during the preprocessing procedure.

The following preprocessing step for the global model 1 stops at step one; however, for global model 2, we had to take multiple extra steps forward: 1) counting the number of crimes for 6-hour time window based on the occurrence time of the crimes from January 1, 2005 to December 31, 2016 (17532 cases) for each precinct, 2) adding the percentage of zoning areas within each precinct. Utilizing this feature could help the model to learn to differentiate between precincts with diverse environmental designs (Carter, Carter and Dannenberg 2003), 3) demographics with borough-sized spatial resolution and yearly temporal resolution were added to the data.

2.2.3 Modeling

2.2.3.1 Ridge Regression

Once the data was preprocessed accordingly, the modeling for local models and the global model was fulfilled using the ridge regression method. Based on our experience with the type of data being used in this study, multiple linear regression is prone to the circumstance of multicollinearity between predicters, so ridge regression was selected as the alternative. Multicollinearity between variables causes a multiple linear regression to gain incorrect magnitude or sign of coefficients with large standard errors (Morrow-Howell 1994, Hoerl and Kennard 1970). Instead, ridge regression is a regularized form of linear regression with a regularization term $\alpha \sum_{i=1}^{n} \theta_i^2$ which is

added to the cost function. The regularization forces the learning algorithm to fit the data and simultaneously maintain the model coefficients as small as possible (Hoerl and Kennard 1970). The hyperparameter α in the model controls the magnitude of regularizing the model. To select the best hyperparameter value and subsequently the best model, 10-fold cross validation for each model with the split of 75 percent training and 25 percent testing data was utilized.

To create a meaningful predictive model which is able to hold both categorical and numerical data, we used one-hot encoding to handle categorical features[2]. One-hot encoding is a tool which transforms categorical features to a binary format for a model which work with numerical data. Using one-hot encoding the categorical encoded variable is removed and a new binary variable is added for each unique categorical value. Since the we dealt with relatively big data, we fulfilled one-hot encoding of the categorical features during the modeling process as a storage management strategy. For the rest of the features[3], since each of which has a different range, we used normalization; $x = \frac{x - x_{min}}{x_{max} - x_{min}}$ to rescale them to the same range of values which was between 0 and 1. Coefficient of determination (R^2) was used to evaluate the goodness of fit of the models.

2.2.3.2 Multi-Task Learning

Countless daily real-world prediction applications that we deal with consist of multiple correlated tasks. However, the standard strategy to resolve such problems has normally assumed independence between the tasks known as single-task learning (STL). STL does not leverage knowledge from nearby regions and may produce poor results in complex circumstances with

[2] For the local models the categorical features are Time, Day, Week, Rain and Snow and for the global model the features are the above-mentioned features plus Precinct and Borough.

[3] For the local models the features are Wind, Average Temperature, Past 7-days Average Crime, Past 14-days Crime Average and Past 21-days Crime Average. For the global model the features are the above-mentioned features plus Park_area, Commercial area, Manufacturing area, Residential area, Male, Female, White, Black, American_Indian/Alaska_Native, Asian, Hawaiian, Other races, Hispanic, Total households

insufficient data. To take the tasks' dependencies into consideration, multi-task learning (MTL) is proposed. MTL introduced by Caruana (1997), is a machine learning paradigm in which multiple learning tasks are solved simultaneously in prediction tasks. An MTL advances the generalization performance of all the tasks by leveraging valuable information included in several related tasks (Zhang and Yang 2017).

Crime prediction is not an exception, and there are many spatial and temporal relatedness among different locations in a city that cannot be disregarded. Therefore, it could provide the research communities a better opportunity to take the advantage of methods such as MTL so that it enables them to catch related information from similar locations to possibly improve the overall prediction results.

A frequently used minimization of penalized loss in predictive algorithms of machine learning is $\min_{W} \mathcal{L}(W) + \Omega(W)$, where w is approximate coefficients, L(W) is the loss on the training set, and $\Omega(W)$ is the regularization that determines the tasks similarity (Zhou, Chen and Ye 2011). Based on the assumption on tasks relatedness, a distinct regularization terms can be derived. There have been numerous studies utilizing novel regularizations on modeling the tasks relatedness (Tibshirani 1996, Evgeniou and Pontil 2004, Argyriou, Evgeniou and Pontil 2007, Jalali et al. 2010, Ji and Ye 2009, Chen, Liu and Ye 2012, Chen et al. 2009, Zhou, Chen and Ye 2011, Zhou et al. 2012, Zhou et al. 2011). In some applications, the task association can be characterized employing a graph where every task is a node, and two nodes are linked through an edge if they are related. For more information refer to Zhou et al. (2011). The graph is described as $\|WR\|_F^2 = tr\big((WR)^T(WR)\big) = tr(WRR^TW^T) = tr(W\mathcal{L}W^T)$ where $\mathcal{L} = RR^T$, known as the Laplacian matrix, which is symmetric and positive definite.

In this study, we defined R matrix as:

$$\begin{bmatrix} \sqrt{D_{12}} & \sqrt{D_{13}} & \dots & 0 & 0 & \dots & 0 \\ -\sqrt{D_{12}} & 0 & \dots & \sqrt{D_{23}} & \sqrt{D_{24}} & \dots & 0 \\ 0 & -\sqrt{D_{13}} & \dots & -\sqrt{D_{23}} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & \sqrt{D_{N-1,N}} \\ 0 & 0 & 0 & 0 & 0 & \dots & \sqrt{D_{N-1,N}} \end{bmatrix} \quad (1)$$

where Dpq is the spatial proximity between p^{th} and q^{th} nodes.

The size of the R matrix equals to $n \times \frac{(n-1)\times n}{2}$. In this study R is a 77 × 2926 matrix. To apply MTL with Graph Laplacian regularization on the case study, the precincts were regarded as nodes. To measure the spatial proximity (D) between precincts, we defined a new measurement named *Weighted Four-Dimensional Spatial Similarity Matrix*.

2.2.3.3 Weighted Four-Dimensional Spatial Similarity Matrix

The spatial proximity between two precincts is a metric to find the similarity between them which varies between 0 to 1. This means that a precinct which is the most similar one to itself, receives a proximity equal to 1 and the rest of the precincts based on the similarity definition could receive a number between 0 to 1. The simplest metric of spatial similarity can be calculated using a Euclidean Distance in which the closest precincts are more similar while the farthest are the most dissimilar. However, to create a more meaningful similarity matrix, we contributed spatial factors which were percentage of zoning areas within each precinct and the geographical distance. Primarily, the dissimilarity of two precincts was defined and then it was converted to a similarity matrix.

The dissimilarity between two precincts was defined as:

$$D(P_i, P_j) = w_{ij} \times \sqrt{(R_{P_i} - R_{P_j})^2 + (C_{P_i} - C_{P_j})^2 + (M_{P_i} - M_{P_j})^2 + (P_{P_i} - P_{P_j})^2} \quad (2)$$

$$w_{ij} = \frac{distance(P_i, P_j)}{Maximum\ distance} \quad (3)$$

where R, C, M and P are the percentage of the residential, commercial, manufacturing and park areas within a precinct, respectively. The greater the difference between the environmental design of two precincts, the greater the dissimilarities. We also added the weight w_{ij} to the formula which is a normalized geographical distance between two precincts. The distance between two precincts is divided by the maximum magnitude of the distances between precincts. It intensifies/abates the overall dissimilarity of distant/close precincts in the final matrix product. For example, say the environmental design dissimilarities of precincts 1 and 5 to precinct 6 are equal, the more distant precinct to 6 would become more dissimilar using the weighting system. The dissimilarity of each pair of precincts were calculated and used to create a matrix of 77×77 with diagonal elements all equal to zero.

As the dissimilarity to similarity conversion, we used $\frac{1}{1+dissimilarity}$ to convert all the elements of the dissimilarity matrix to the similarity. Obviously, the final similarity matrix would become a matrix with diagonal elements all equal to one.

Ultimately, the 75 percent training and 25 percent testing data which were used in the local modeling were input in the MTL with Graph Laplacian regularization using MALSAR package (Zhou et al. 2011). The MALSAR (Multi-Task Learning Via Structural Regularization) is a MTL package with different regularizations which is only available in MATLAB[4]. To tune the

[4] http://www.mathworks.com/products/matlab/

hyperparameters of the regularization and select the best model, a 10-fold cross validation was implemented.

In the following section, we present our assessment of the utility of MTL against local and global modeling in crime prediction. We begin by examining the local and global models' performance. We then examine MTL with Graph Laplacian regularization in crime prediction and lastly is compared to the local models and the global model.

2.3 Results and Discussion

2.3.1 Local Modeling

The data for 77 precincts of the NYC was preprocessed according to the steps explained previously. Of the 4,383 days across the time period of January 1, 2005 to December 31, 2016, in order to create the three features, the average of number of crimes in the past 7 days, past 14 days and past 21 days for each 6-hour time window, the first 21 days were removed. The remaining 4,362 days resulted in 17,448 cases as a day contains four time intervals. Overall, 77 data files with 10 input features of Time, Day, Week, Wind, Precipitation, Snow, Temperature, Past 7 days moving average number of Crime, Past 14 days moving average Number of Crimes and Past 21 days moving average number of crimes were produced. With one-hot encoding the categorical features[5], a matrix size of 17,448 rows and 73 columns for each precinct was formed. The R^2 of the best model using 10-fold cross validated ridge regression modeling with 75 percent training and 25 percent testing for each precinct is presented in Figure 2-2. Note that the predictand of the modeling was the number of crimes.

[5] Time (4), Day (7), Week (53), Wind (1), Precipitation (2), Snow (2), Temperature (1), Past 7-days moving average number of crime (1), Past 14-days moving average number of crimes (1) and Past 21-days moving average number of crimes (1). The numbers greater than 1 for the features are the number of features after transformation with one-hot encoding.

Precincts 22 and 121 with R^2 equal to 0.08 and 0.64 had the lowest and the highest predictive performances where the average R^2 of the local models was equal to 0.39. Evaluation of the coefficients of the regression models indicated that time interval 3 (12 pm-6pm) had the largest magnitude among the four time intervals. The total number of crimes across NYC for the entire twelve years of data are presented in Figure 2-2. Accordingly, the local models were capable to properly recognize its significance. Additionally, Friday had the largest magnitude of the coefficients compared to other six days of week. Figure 2-3 illustrates the statistics of crime in each day of the week. The models were again able to distinguish the importance of Friday as the most dangerous day of the week. However, the models did not signify any week of year as the most likely week for crime occurrence. The 12 years of data for the total number of crimes in each week is presented in Figure 2-4, where the models and the data show agreement. Finally, the models recognized that possibility of crime occurrence in snowy and rainy days by showing negative coefficients, where the coefficient of temperature was positive. In other words, an increase in temperature correlates with the increase in crime occurrence. Windy days coefficient is also positive which results in more crime occurrence.

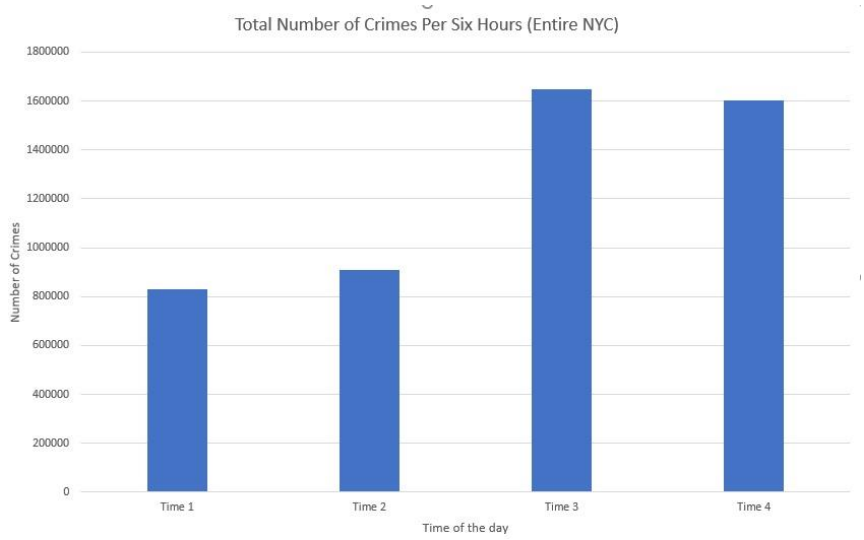


Figure 2-2 Total number of crimes in each 6 hours of day

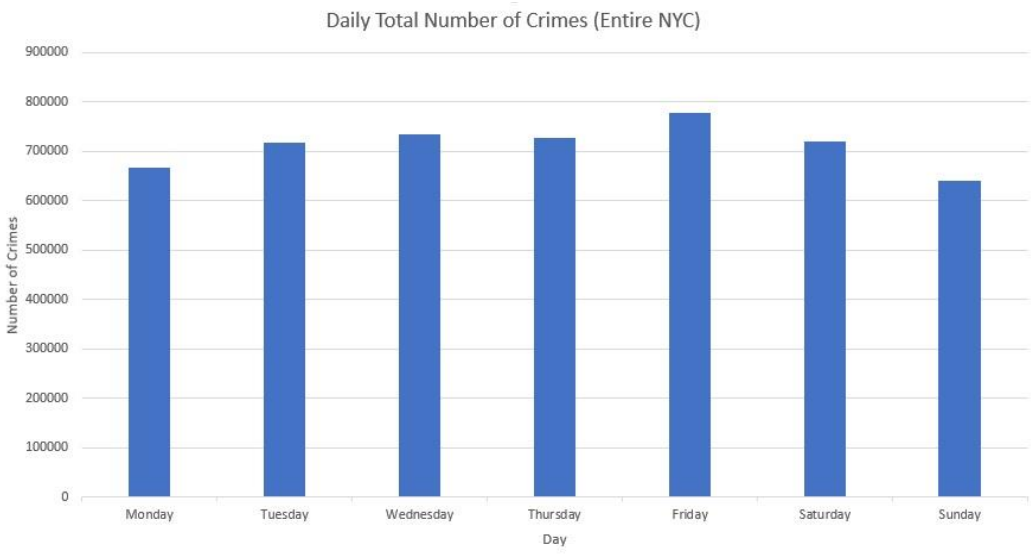


Figure 2-3 Total number of crimes during each day of week

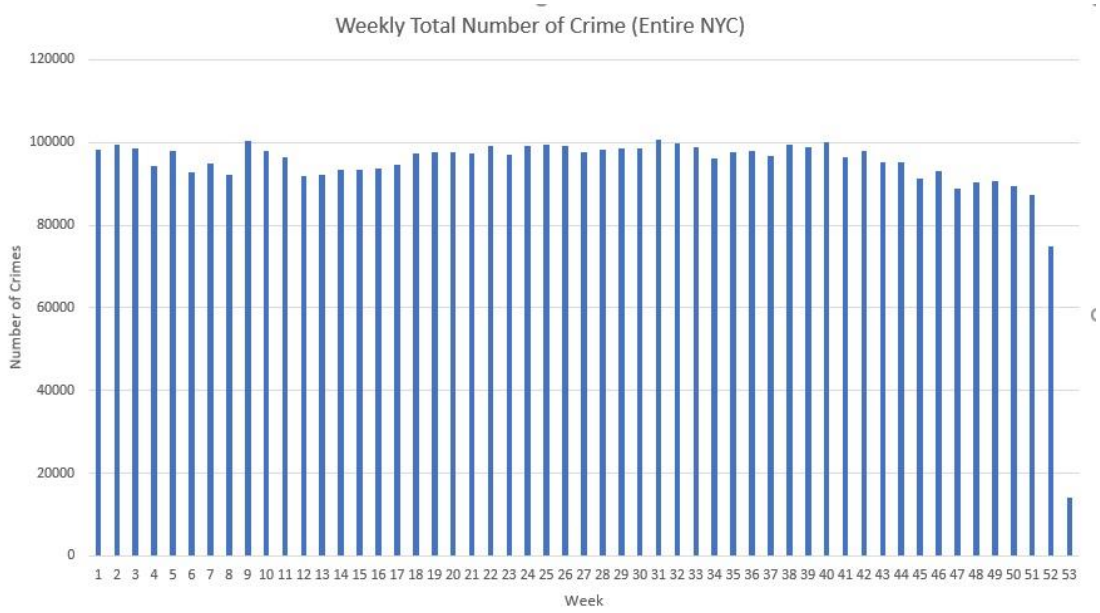


Figure 2-4 Total number of crimes during each week of year

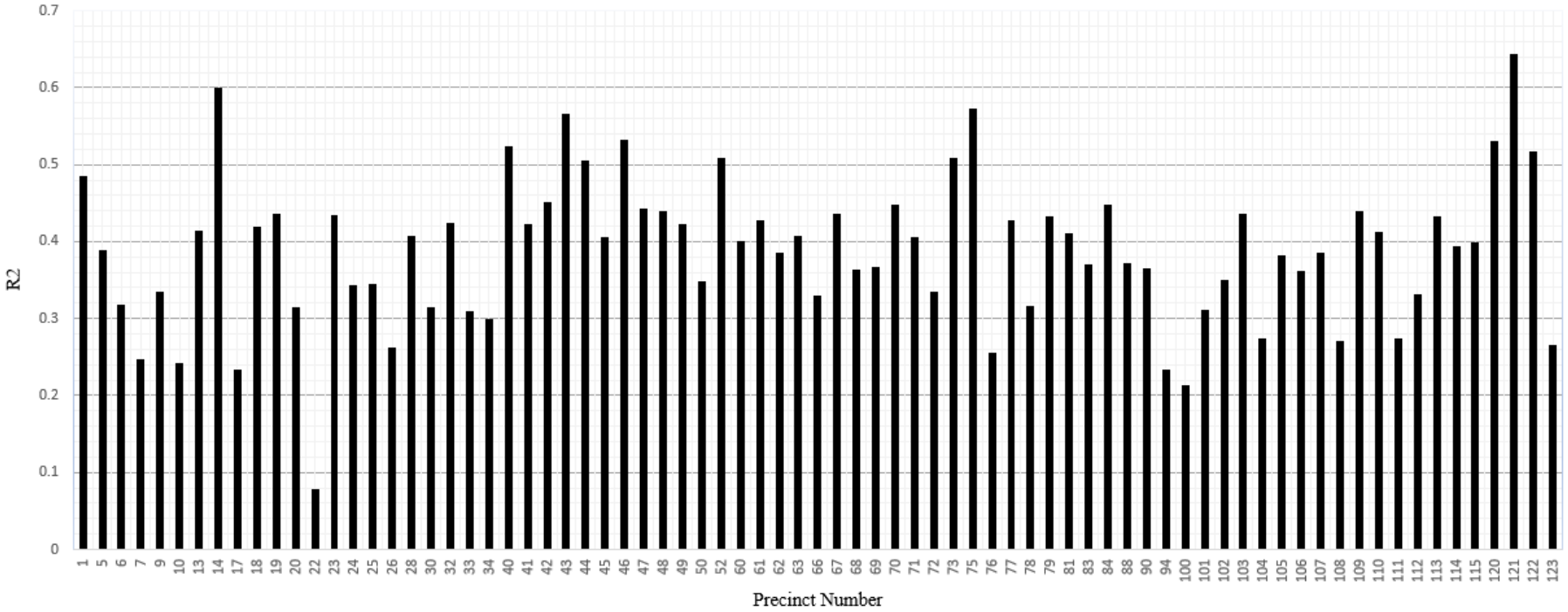


Figure 2-5 R2 of the local models

2.3.2 Global Modeling

The data for the first global model was preprocessed and a matrix size of 1343496 × 11 with the same exact features as was input in the local models plus the precinct feature was created. One-hot encoding of the categorical features[6] increased the size of the matrix to 1343496 × 150. The R^2 of the best model using 10-fold cross validated ridge regression modeling with 75 percent training and 25 percent testing was equal to 0.33. The results show that this global model on the average, is performing worse than the local models. However, the second global model including demographics and zoning area percentage, was supplied with an 1343496 × 26 matrix, which was resized to a 1343496 × 169 matrix after one-hot encoding the categorical features[7]. Since we added more features to the model, the R^2 of the best model using 10-fold cross validated ridge regression modeling with 75 percent training and 25 percent testing improved to 0.48. Although the accuracy of the models may not be satisfying, however, in a fair comparison between the local and global models, the results indicated that local models were better suited in modeling as there was less heterogeneity in the used data.

[6] Time (4), Day (7), Week (53), Wind (1), Precipitation (2), Snow (2), Temperature (1), Past 7-days Average Number of Crime (1), Past 14-days Average Number of Crimes (1), Past 21-days Average Number of Crimes (1) and Precinct (77). The numbers greater than 1 for the features are the number of features after transformation with one-hot encoding.

[7] The features being used in the first global model plus Male (1), Female (1), White (1), Black (1), American Indian/Alaska Native (1), Asian (1), Hawaiian (1), Hispanic (1), Other races (1), Total households (1), Boroughs (5), Residential (1), Commercial (1), Manufacturing (1) and Park (1).

2.3.3 MTL modeling

As it was previously discussed, an MTL model captures the spatial and temporal correlation between the tasks. Considering each precinct as a single task, the preprocessed data of each precinct in the local modeling part along with the R matrix were imported into the MTL model. The R^2 of the best models using 10-fold cross validated ridge regression modeling with 75 percent training and 25 percent testing for each precinct is presented in Figure 2-6. Similar to the local models, the highest prediction performance belonged to precinct 121 while the precinct 22 showed the weakest performance. Interestingly, the average R^2 of MTL modeling was equal to 0.39 equal to the local models average while both the MTL and the local models outperformed the first global model. However, the second global model was still superior in terms of results because of benefiting from more input information. A side-by-side comparison of the performance of the local models and MTL models for each precinct is presented in Figure 2-7.

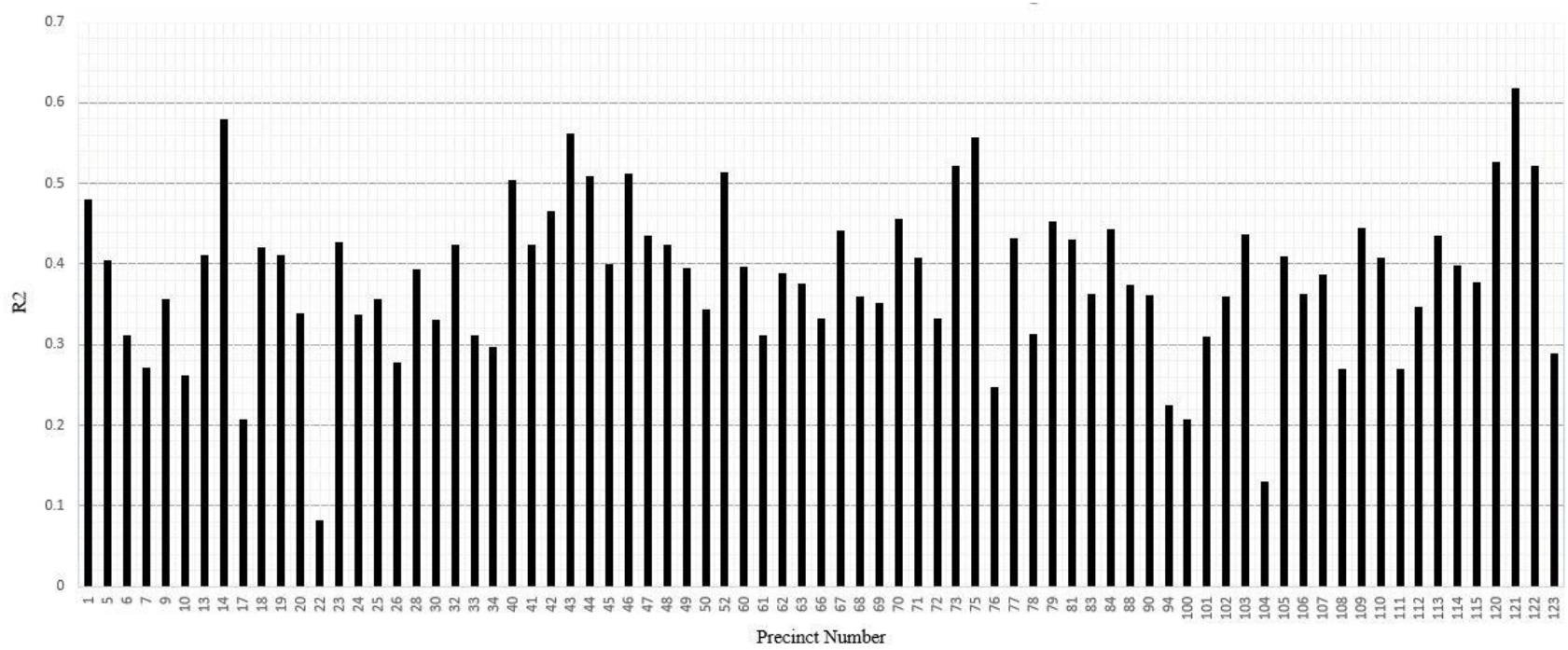


Figure 2-6 R2 of the MTL models

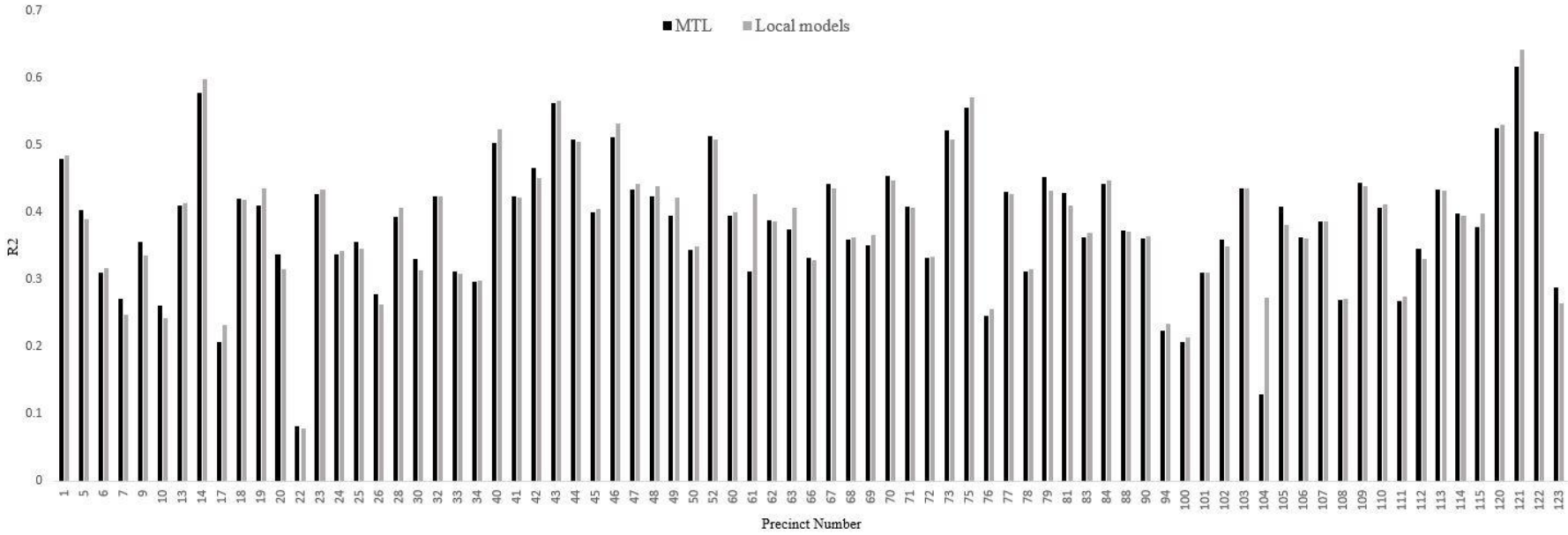


Figure 2-7 Local models versus MTL models

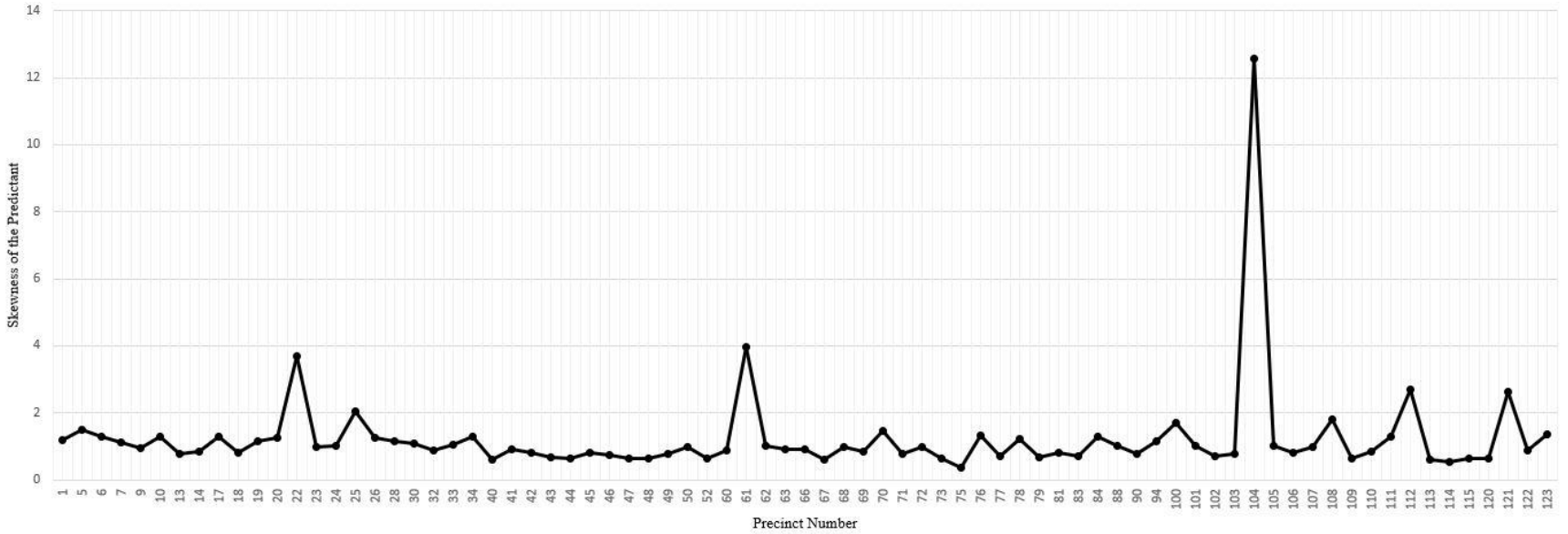


Figure 2-8 Skewness of the distribution of the dependent variable (number of crimes) in every precinct

2.3.4 Data Sparsity Impact

To better understand the problem, we took a deeper look into the data using a skewness test of the distribution of the dependent variable, the number of crimes in each precinct. Since the defined temporal resolution of the study was fine, in most of the 6-hour time intervals of every precinct, there was no crime happening so the predictand contained too many zeros. The skewness of the distribution of number of crimes for every precinct is presented in Figure 2-8. Positive skewness in all precincts implies that we face a heavily right-skewed distribution in which the peak is closer to zero (data sparsity). To evaluate the susceptibility of the local and MTL modeling, we fulfilled a Pearson's correlation with two-tailed test of significance. The test measured the linear correlation of the obtained R^2 and the skewness of the predictand distribution. Table 2-1 presents the correlation coefficients, significant at 0.01 level.

Table 2-1 Pearson's Correlation Test of the Methods' Performances and Skewness of the Number of Crimes Distribution

Variable	Skewness of number of crimes
Local Modeling R^2	-0.275
MTL Modeling R^2	-0.449

The results indicate that MTL is more susceptible to the skewness of the predictand and as the skewness increases, there is a stronger decline in the performance of the MTL compared to the local models. Essentially, the result imply that we needed a coarser temporal resolution to subsequently decrease the data sparsity to experience an improvement with MTL performance.

2.3.5 Effect of Precinct Spatial Design

Another perspective is to investigate the relationship between the modeling results and the effect of spatial properties. Figure 2-10 portrays the ratio of crime per area (CPA) in each precinct of NYC. In one test, we evaluated the results of both local models and MTL against CPA. Figure 2-9 shows the obtained R^2 in each precinct and the corresponding CPA. Although, there is no observed consistency between the methods performances and CPA, we assessed the correlation of the performance of the two methods with the city spatial design including percentage of four different zonings of the city.

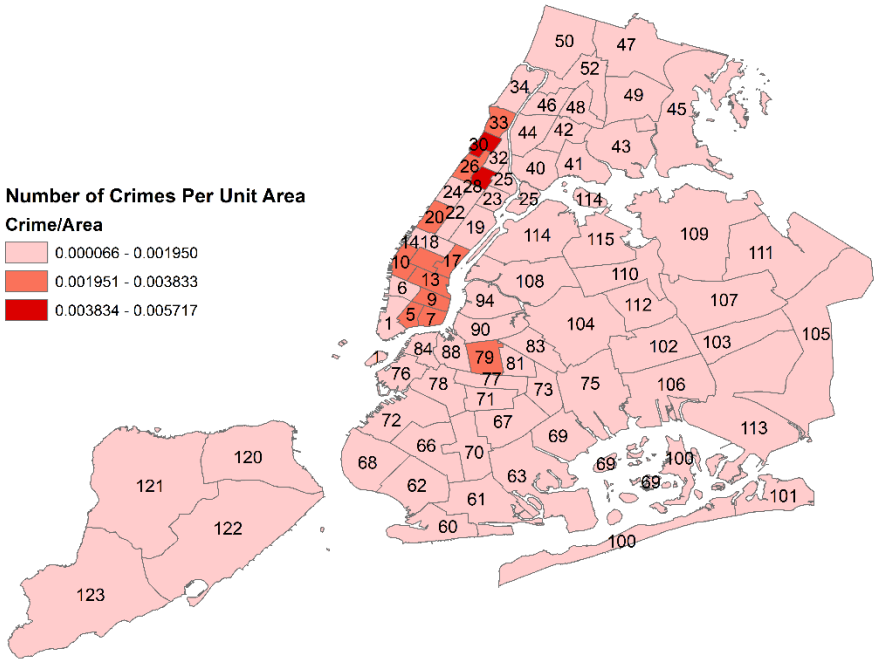


Figure 2-9 Ratio of crime per area in every precinct of NYC

The results can be found in Table 2-2. The result indicates that there is a significant negative correlation between the performances of the models as the park area increases, however the rest of the correlations are not significant. With the help of those information, a deeper look into the

percentage of zoning areas in each precinct assisted to find that precinct 22 contains 95 percent park area which is the highest percentage and the reason for poor performance of its models. The median of park area percentages in all precincts was equal to 7 percent whereas the 95 percent park area in precinct 22 would affect the models' generalizability.

Table 2-2 Correlation analysis of the impact of spatial design of the NYC with the models performances

		MTL R^2	**Local R^2**
Park	Pearson Correlation	-0.325**	-0.327**
	Sig. (2-tailed)	0.004	0.004
Commercial	Pearson Correlation	0.099	0.091
	Sig. (2-tailed)	0.389	0.430
Manufacturing	Pearson Correlation	-0.007	-0.002
	Sig. (2-tailed)	0.951	0.985
Residential	Pearson Correlation	0.152	0.156
	Sig. (2-tailed)	0.188	0.176

** Correlation is significant at the 0.01 level (2-tailed)

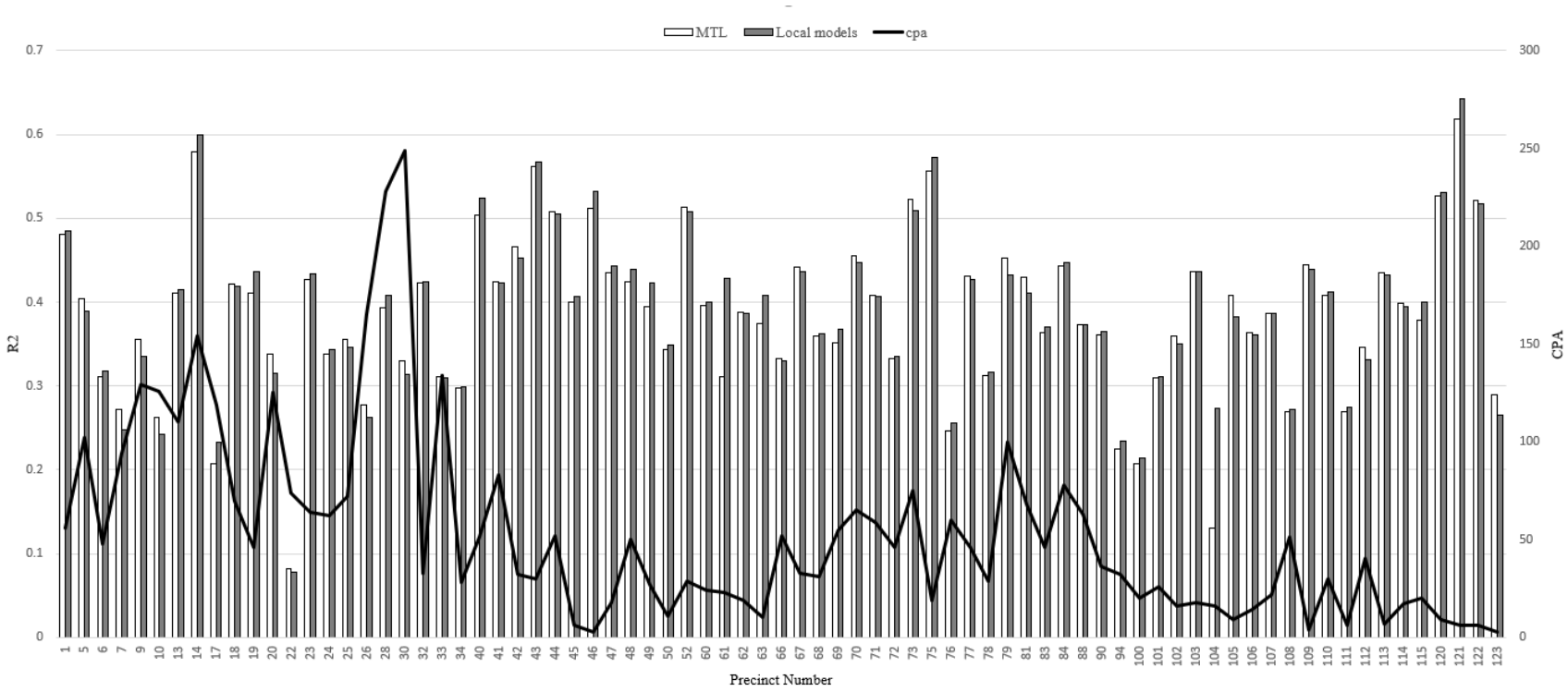


Figure 2-10 MTL and Local Models' Performances and CPA

2.3.6 Training Size Impact

MTL is supposed to be a better algorithm in case of limited training samples comparing to STL methods (Zhang and Yang 2017). Hence, we tested the performance of MTL against the local models using different data sizes by incrementally reducing the training samples. It should be noted that the global model with the same number of features, was not tested as it already showed the worst performance. The results of the test are presented in Table 2-3 which indicates that MTL modeling performs slightly better than the STL (i.e., local models build by ridge regression) models as the training size becomes more limited. However, the small difference between their performances may be due to still large data in our case study, even with using only 10 percent of training samples.

Table 2-3 Comparison of the performance of local Modeling and MTL with different training sizes

Training (percent)	Testing (percent)	Local Modeling Mean R^2	MTL Modeling Mean R^2
75	25	0.39	0.39
60	40	0.39	0.39
50	50	0.38	0.39
30	70	0.38	0.39
10	90	0.37	0.38

2.4 Conclusions

In this study, we examined the spatiotemporal crime-prediction performance of a multi-task learning method against linear local models and a global model. The result showed that the multi-task learning model outperformed the global model in prediction, however, its prediction performance on average stood equal to the performance of the local models. Two important findings of this study were first, the more negative effect of dependent variables skewness on MTL models as opposed to STL models. Second, the effect of training size which in case of limited samples, MTL performed better than the local modeling. Despite the equal performance of the MTL and the local models, several limitations were observed that they may have utility for future crime prediction.

Spatial and temporal scale is a vital subject in crime prediction. From a practical point of view, we need to address them adequately, however, finer spatial and temporal resolutions significantly influence the prediction results. If the used resolutions are too fine, then there will be too small historical crime data which results in data sparsity. Consequently, the models are not able to acquire a good estimate of crime rate. On the other hand, too coarse resolutions do not make the predictions sufficient for police preemptive actions. Besides that, an optimized spatial and temporal scale for one specific area, is not certainly ideal for other areas (Liu 2017).

Considering all crime types under one single crime class to prevent more data sparsity presented another challenge for this study. We could improve the results by designing a more accommodating framework by either choosing a coarser spatial scale (e.g., borough size instead of precinct size) or a coarser temporal resolution (e.g., daily basis) to avoid data sparsity while increasing the dependent variable range. This was a challenging task, not only because of efforts to save the practicality of the predictive models, but also due to the structure of the available crime

data and its supplementary information. Data acquisition from diverse sources, cleansing and processing data from a collected pool of unstructured data, aggregation and integration of the cleaned data were all the influential factors in making this study more difficult.

Demographic factors have been referred to as the most important determinants of crime rates. However, the demographic data that precisely overlapped the precinct borders were available only at a course resolution, borough level. As a result, the valuable information from demographics could not be used in the local models at the precinct level. Demographic data from census tract could be used, however, the more significantly challenging task was to obtain and preprocess the data to exactly match the precinct borders. As was discussed in the first chapter, one challenging aspect of spatiotemporal data analysis involves the difficulty of joining and relating different resolutions and sources of data. Furthermore, we studied the correlation of crime rate per area, but with the availability of demographic information at precinct resolution, we could have included a study of the crime rate correlation with population density, income level and economic condition in relation to different zonings. Additional improvement that could potentially be made for this study involves including more accurate zoning description of the areas within each precinct. The makeup of New York City's zones is broken down into four main categories: residential, commercial, manufacturing and parks. However, each major zone is also divided into multiple subzones. For example, residential areas include ten basic residence districts - R1 through R10 – and each zone differs in population density and required parking. In this study, to simplify the data processing and modeling, only the major zones in each block were used in the models. By using more detailed zoning information, there was a possibility to improve the modeling tasks. Finally, the available weather information at daily temporal resolution and solely from three stations in the

area could not show an ideal effectiveness at precinct level spatial resolution, when 77 precincts were receiving relatively similar information.

Other important factors such as economy and income could be employed in this study, however, the data acquisition and processing for the scope of this study was challenging. Certainly, there is room for improvement of the results if one were to consider the aforementioned angles, however, the primary purpose of this study to establish a basis for future crime studies by introducing MTL to the research community. We believe that appropriate variable selection at proper spatial resolution correlated to crime occurrence is critical to the success of the models, and in the future, adding other sources of data could be another step forward of such framework.

BIBLIOGRAPHY

BIBLIOGRAPHY

Argyriou, A., T. Evgeniou & M. Pontil. 2007. Multi-task feature learning. In *Advances in neural information processing systems*, 41-48.

Arulanandam, R., B. T. R. Savarimuthu & M. A. Purvis. 2014. Extracting crime information from online newspaper articles. In *Proceedings of the Second Australasian Web Conference-Volume 155*, 31-38. Australian Computer Society, Inc.

Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi & A. Pentland. 2014. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, 427-434. ACM.

Brown, D., J. Dalton & H. Hoyle. 2004. Spatial forecast methods for terrorist events in urban environments. In *International Conference on Intelligence and Security Informatics*, 426-435. Springer.

Carter, S. P., S. L. Carter & A. L. Dannenberg (2003) Zoning out crime and improving community health in Sarasota, Florida:"crime prevention through environmental design". *American Journal of Public Health,* 93**,** 1442-1445.

Caruana, R. (1997) Multitask learning. *Machine learning,* 28**,** 41-75.

Chainey, S. & J. Ratcliffe. 2013. *GIS and crime mapping*. John Wiley & Sons.

Chainey, S., L. Tompson & S. Uhlig (2008) The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal,* 21**,** 4-28.

Chen, J., J. Liu & J. Ye (2012) Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 5**,** 22.

Chen, J., L. Tang, J. Liu & J. Ye. 2009. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 137-144. ACM.

Chen, X., Y. Cho & S. Y. Jang. 2015. Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*, 63-68. IEEE.

Clinard, M. B. (1942) The process of urbanization and criminal behavior. *American Journal of Sociology,* 48**,** 202-213.

Cohn, E. G. (1990) Weather and crime. *The British Journal of Criminology,* 30**,** 51-64.

Das, S. & M. R. Choudhury (2016) A Geo-Statistical Approach for Crime hot spot Prediction. *International Journal of Criminology and Sociological Theory*, 9.

Evgeniou, T. & M. Pontil. 2004. Regularized multi--task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 109-117. ACM.

Flowers, R. B. 1989. *Demographics and criminality: The characteristics of crime in America*. Greenwood Press New York.

Gerber, M. S. (2014) Predicting crime using Twitter and kernel density estimation. *Decision Support Systems,* 61**,** 115-125.

Gorr, W. & R. Harries (2003) Introduction to crime forecasting. *International Journal of Forecasting,* 19**,** 551-555.

Groff, E. R. & N. G. La Vigne (2002) Forecasting the future of predictive crime mapping. *Crime Prevention Studies,* 13**,** 29-58.

Gruenewald, P. J., B. Freisthler, L. Remer, E. A. LaScala & A. Treno (2006) Ecological models of alcohol outlets and violent assaults: crime potentials and geospatial analysis. *Addiction,* 101**,** 666-677.

Hoerl, A. E. & R. W. Kennard (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics,* 12**,** 55-67.

Horrocks, J. & A. K. Menclova (2011) The effects of weather on crime. *New Zealand Economic Papers,* 45**,** 231-254.

Jalali, A., S. Sanghavi, C. Ruan & P. K. Ravikumar. 2010. A dirty model for multi-task learning. In *Advances in neural information processing systems*, 964-972.

Ji, S. & J. Ye. 2009. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, 457-464. ACM.

Kadar, C. & I. Pletikosa (2018) Mining large-scale human mobility data for long-term crime prediction. *arXiv preprint arXiv:1806.01400*.

Liu, H. & D. E. Brown (2003) Criminal incident prediction using a point-pattern-based density model. *International journal of forecasting,* 19**,** 603-622.

Liu, X. 2017. Temporal and Spatiotemporal Models for Short-Term Crime Prediction. Illinois Institute of Technology.

Malik, A. A. (2016) Urbanization and Crime: A Relational Analysis. *J. HUMAN. & Soc. Scl.*, 21**,** 68, 69-70.

Mookiah, L., W. Eberle & A. Siraj. 2015. Survey of Crime Analysis and Prediction. In *FLAIRS Conference*, 440-443.

Morrow-Howell, N. (1994) The M word: Multicollinearity in multiple regression. *Social Work Research*.

Poulsen, E. & L. W. Kennedy (2004) Using dasymetric mapping for spatially aggregated crime data. *Journal of Quantitative Criminology*, 20**,** 243-262.

Sassen, S. 2016. The Global City: Strategic Site, New Frontier. In *Managing Urban Futures*, 89-104. Routledge.

Smith, C., D. Quercia & L. Capra. 2013. Finger on the pulse: identifying deprivation using transit flow analysis. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 683-692. ACM.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 267-288.

Wang, B., D. Zhang, D. Zhang, P. J. Brantingham & A. L. Bertozzi (2017) Deep learning for real time crime forecasting. *arXiv preprint arXiv:1707.03340*.

Wang, H., D. Kifer, C. Graif & Z. Li. 2016. Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 635-644. ACM.

Wang, M. & M. S. Gerber. 2015. Using Twitter for Next-Place Prediction, with an Application to Crime Prediction. In *Computational Intelligence, 2015 IEEE Symposium Series on*, 941-948. IEEE.

Weatherburn, D. (2001) What causes crime? *BOCSAR NSW Crime and Justice Bulletins***,** 11.

Wirth, L. (1938) Urbanism as a Way of Life. *American journal of sociology*, 44**,** 1-24.

--- (1964) LOUIS WIRTH ON CITIES AND SOCIAL LIFE; SELECTED PAPERS.

Xue, Y. & D. E. Brown (2006) Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decision support systems*, 41**,** 560-573.

Yang, D., T. Heaney, A. Tonon, L. Wang & P. Cudré-Mauroux (2017) CrimeTelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web***,** 1-25.

Yu, C.-H., W. Ding, P. Chen & M. Morabito. 2014. Crime forecasting using spatio-temporal pattern with ensemble learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 174-185. Springer.

Zhang, Y. & Q. Yang (2017) A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhao, X. & J. Tang. 2017. Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 497-506. ACM.

--- (2018) Crime in Urban Areas:: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 20**,** 1-12.

Zhou, J., J. Chen & J. Ye. 2011. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*, 702-710.

--- (2011) Malsar: Multi-task learning via structural regularization. *Arizona State University*, 21.

Zhou, J., J. Liu, V. A. Narayan & J. Ye. 2012. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1095-1103. ACM.

Zhou, J., L. Yuan, J. Liu & J. Ye. 2011. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 814-822. ACM.

Chapter 3 AUTOMATED ANALYSIS OF THE US DROUGHT MONITOR MAPS WITH MACHINE LEARNING AND MULTIPLE DROUGHT INDICATORS

3.1 Introduction

Drought is a common, periodic and one of the costliest natural disasters that has direct and indirect economic, environmental and social impacts (Wilhite et al., 2007). These impacts become even more serious with the potential increase of drought occurrence and severity caused by climate change (Dai, 2011). A systematic and effective drought monitoring, prediction and planning system is thus crucial for drought mitigations (Boken, 2005). However, as discussed in Hao et al. (2017), drought analysis is not an easy task for a number of reasons. To start with, there is a lack of an explicit and universally accepted definition for drought since it is a multi-faceted phenomenon. Based on the variables in consideration, there are four general types of droughts, namely meteorological, agricultural, hydrological and socioeconomic drought, for each of which different combinations of drought indices are used to characterize them (Keyantash & Dracup, 2002). Yet, there is no agreement on typical indices and their thresholds for those drought types (M. Hayes et al., 2011) since they do not work for all circumstances (Wilhite, 2000).

Although developing and choosing a proper set of physical drought indices is the basis of drought monitoring to capture the complexity and describe the consequences of drought, a composite index method has been proved to bring more success to the analysis (Hao et al., 2017). The U.S. Drought Monitor (USDM) was developed as the landmark tool in this regard as it not only uses physical drought indices, but also relies on experts' knowledge in the information interpretation (Anderson et al., 2011). This type of composite drought monitoring, which transforms an abundant set of indicators into a sole product, is called the "hybrid monitoring approach" (M. J. Hayes et al., 2012).

The USDM was established in 1999 aiming at presenting current drought severity magnitude in the categorical means across the U.S. in a weekly map published every Thursday. In the USDM maps, drought is categorized into five categories starting from D0 (abnormally dry), to D1 (moderate drought), D2 (severe drought), D3 (extreme drought) and D4 (exceptional drought). The categories are based on a percentile approach which allows the users to interpret the drought intensity concerning the odds of event occurrence in 100 years (Svoboda, 2000). For example, D0 corresponds to a 20-30% chance for the drought to occur in ranges from 20 to 30 while for D4 it is less than 2%. A USDM map for the week of August 2018 is shown in Figure 3-1 as an example. The map shows areas of the U.S. that are experiencing drought in various severities as well as its impact levels.

To date, there are six main physical indicators in USDM to define the intensity of the categories: Palmer Drought Severity Index (PDSI) (Palmer, 1965), Climate Prediction Center (CPC) Soil Moisture Model Percentiles, U.S. Geological Survey (USGS) Daily Streamflow Percentiles, Percent of Normal Precipitation and Standardized Precipitation Index (SPI), and remotely sensed Satellite Vegetation Health Index (VT) along with many other supplementary indices such as the Keetch-Bryam Drought Index (KBDI) for fire, Surface Water Supply and snowpack (Svoboda et al., 2002), etc. These indices merged with other in situ data are jointly analyzed by experts to depict the drought categories across the country (M. J. Hayes et al., 2012).

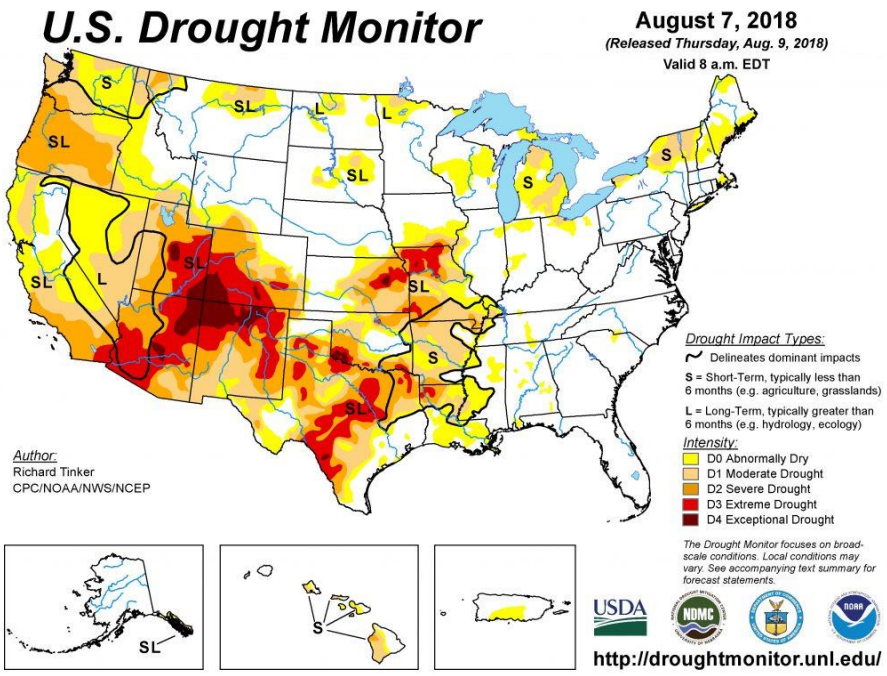


Figure 3-1- USDM map for the Week of August 7, 2018 (droughtmonitor.unl.edu, 2018)

The characteristics of the USDM which makes it a distinct effort in terms of drought monitoring are provided in Table 3-1. Since the uniqueness of the USDM has made it extremely popular, much attention is drawn to it from media, policy makers and managers (USDA, 2018) as a benchmark in their drought related communications and interpretations. Similarly, researchers have started using the USDM product as a reference observation to compare and validate their proposed drought monitoring and prediction methods (Anderson et al., 2013; Anderson et al., 2011; Brown et al., 2008; Gu et al., 2007; Hao & AghaKouchak, 2014; Lorenz et al., 2017a, 2017b; Otkin et al., 2016; Quiring, 2009). Although it is desirable to predict the USDM drought conditions which are in categorical format, it would not be an easy task due to the subjectivity included in the production process by the experts. A few studies (Hao, Hao, et al., 2016; Hao, Hong, et al., 2016) predicted the monthly average USDM drought categories using ordinal regression by integrating multiple drought indices. However, there has been no previous study using machine learning approaches to

predict the USDM drought categories specifically in the original weekly format as the USDM publishes the maps.

Table 3-1 Uniqueness of the US Drought Monitoring (droughtmonitor.unl.edu, 2019; Svoboda et al., 2002)

Characteristics	Details
The first nationwide unifying drought monitoring of multiple entities	- Authors from National Drought Mitigation Center (NDMC), United States Department of Agriculture (USDA), Climate Prediction Center (CPC), and National Climatic Data Center (NCDC) have the responsibility of drawing the maps who take turns for two weeks - The authors blend the best available data from various resources for interpretation
Receives local observers' collaboration	- More than 425 local observers such as state climatologists, National Weather Service staff, agricultural and water resources managers, and hydrologists - They provide drought impacts for the products using their familiarity and knowledge of the region so that the experts can depict the most accurate classification on the map
Simple and effective	- The classification system for droughts is easy to understand for public - Drought spatial extent, intensity, and duration are all considered - Flexibility with new technologies and data incorporation
Timely	- It is a weekly product which illustrates drought conditions and impacts in a timely manner

In this study, we aim to reproduce the same USDM drought analysis map over conterminous United States (CONUS) based on meteorological observations and land surface model simulated hydrological quantities through a machine learning approach and using multiple drought indicators. We apply linear and nonlinear machine learning approaches using multiple combinations of drought indices against a persistence model serving as the baseline model. The developed framework basically mimics the map synthesizing process executed by the USDM

authors. This will not only test the suitability of machine learning methods in drought monitoring and prediction, but also helps us to develop tools that can translate predictions with numerical models to easy-to-understand categorical drought forecasts.

The rest of this paper is organized as follows. Section 3-2 elaborates the study area, data and describes the methodology. Section 3-3 presents the results and discussions. Finally, in the last section we summarize and conclude the findings of this study.

3.2 Data and Methodology

In this section the framework of the study to reproduce the USDM drought maps is explained. The process of developing the framework is presented in Figure 3-2, starting from data preparation including data collection and simulation followed by data preprocessing prior to inputting into the models. Each task is explained in the following sub-sections.

As the drought indices used in our study were derived from land-surface model outputs forced by the North American Land Data Assimilation System Phase 2 (NLDAS2)'s meteorological forcing fields, in this study, we deliberately designed our modeling domain to be consistent with the NLDAS2 grids. Thus, the modeling grids span the entire CONUS from 25.0625 to 52.9375 degree latitude and -67.0625 to -124.9375 degree longitude, at $1/8^{\circ}$ latitude-longitude degree resolution which forms a meshed area with 224 rows and 464 columns (Mitchell et al., 2004).

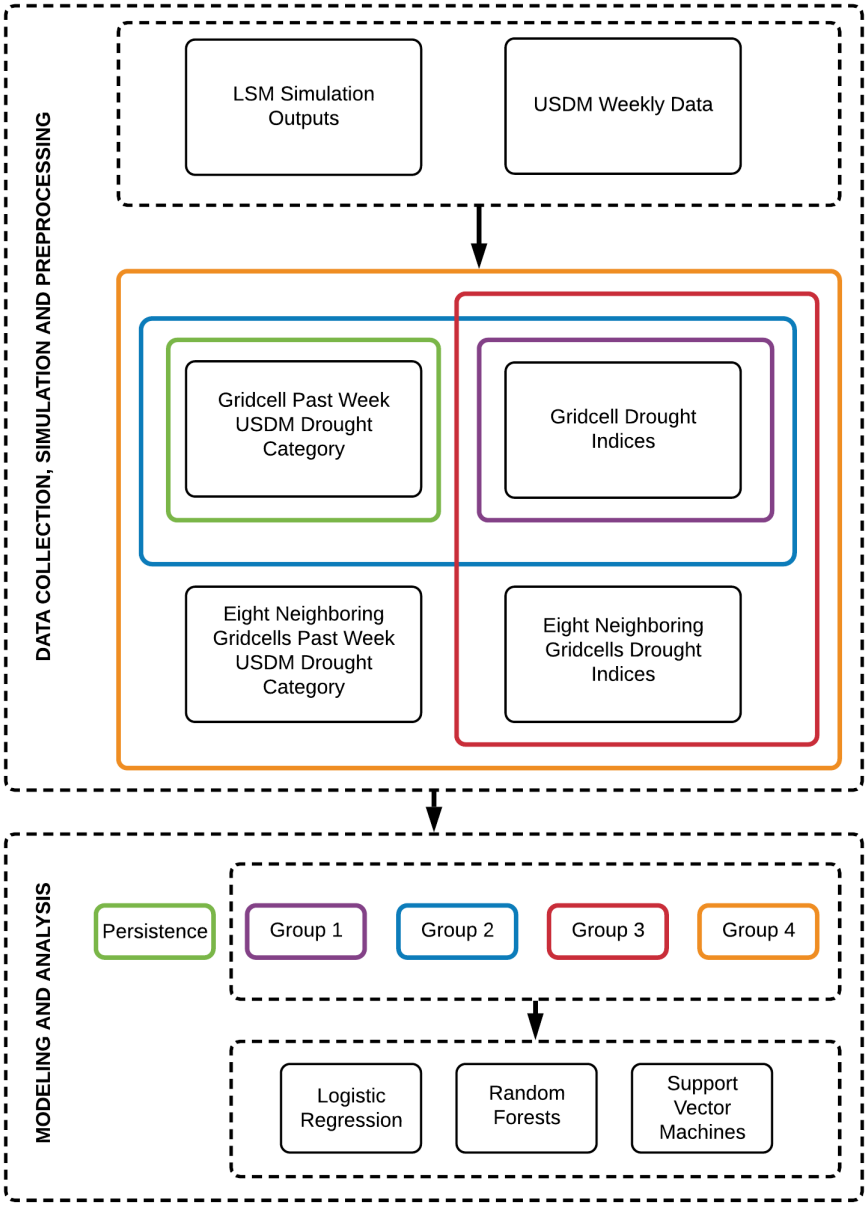


Figure 3-2 Flowchart of the proposed framework for USDM drought categories prediction

3.2.1 Data Collection, Simulation and Preprocessing

To reproduce the USDM maps, a collection of predictor variables, which correspond to drought indices were needed to predict the USDM categories. In the following paragraphs, the process of data collection, simulation and preprocessing are described. We also explain the rationale behind the selection of each variable and how we obtained, calculated, and resampled the values for each of them prior to modeling.

3.2.1.1 USDM Data

The USDM weekly drought maps were retrieved from the USDM archived data at https://droughtmonitor.unl.edu/Data/GISData.aspx for the years of 2000 through 2013, starting on January 4 of 2000 and ending on December 31 of 2013, creating a total of 731 weeks of data. The USDM drought maps are vector data that outline the regions in each drought category. As the goal of this study is to reproduce the weekly USDM drought condition across CONUS, each weekly map has to be rasterized to 1/8 degree NLDAS2 grid. Then for every week, each grid cell is labeled as one of the five USDM drought categories or "No Drought" which makes an overall of six possible states. In the resterization process, any grid cell covering two or more different drought categories is labeled with the drought category which occupied the largest area.

3.2.1.2 Land Surface Model Outputs and Drought Indices

As the input variables of the actual USDM weekly report vary widely, we selected the frequently used indices in forecasting and monitoring drought. These indices are also the ones that benefit the USDM weekly map production (Anderson et al., 2011). Standardized Precipitation Index (SPI), Standard Runoff Index (SRI), Soil Moisture Percentile (SMP) and Palmer Drought Severity Index (PDSI) are the employed indices in this study which are summarized in Table 3-2 and are used as the predictors of the models to predict the USDM drought categories.

Table 3-2 Summary of the drought indices

Drought Index	Definition	Used in This Study	Source(s) of Data in This Study	Reference
Standardized Precipitation Index (SPI)	The number of standard deviations that the cumulative precipitation deficit would deviate from the long-term normalized mean. SPI value can be calculated for multiple time scales, covering the last 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 18, 24, 30, 36, 48, 60, and 72 months.	SPI for the 30, 60, and 90 days prior to the day of forecast	NLDAS-2 Forcing File A Precipitation	McKee et al. (1993)
Standardized Runoff Index (SRI)	Defined the same as SPI, except for runoff, the number of standard deviation that the percentile of cumulative hydrologic runoff would deviate over a particular duration.	SRI for the 30, 60, and 90 days prior to the day of forecast	Noah3.6, NoahMP3.6, clsmf2.5 Surface and Subsurface Outputs	Shukla and Wood (2008)
Soil Moisture Percentile (SMP)	The quantile of the current day top 1-m total soil moisture among all the data pools from the historical period over a particular onward and backward time window.	Top 1-m soil moisture 29 Days Time Window	Noah3.6, NoahMP3.6, clsmf2.5 Soil Moisture Outputs	Sheffield et al. (2004)
Palmer Drought Severity Index (PDSI)	A standardized index in which the inputs of monthly temperature, precipitation and the available water capacity (AWC) of the soil are used for estimation of dryness.	PDSI	Obtained from Abatzoglou (2019)	Palmer (1965)

SRI and SPI are typically calculated based on monthly data, and can be calculated for up to 72 month historical time periods. In this study, as we try to predict USDM weekly maps, we calculate the SPI and SRI based on daily data (Table 3-2) at 30-day, 60-day, and 90-day periods. These are the periods that prior to the day of forecast. For convenience, we still call them SPI1, SPI2 and SPI3, just to be consistent with other literature as their time scales are roughly equivalent to one months, two months and three months. In order to create the indices, we first gathered the outputs of both NLDAS-2 Forcing precipitation data and the Land Information System (LIS) models (Noah-3.6, Noah-MP3.6, CLSM-F2.5) runoff and soil moisture from 1979 to 2013. The NLDAS-2 Precipitation, and LIS models hydrological runoff and soil moisture were used to calculate SPI, SRI and SMP, respectively. It is notable that the calculations of SRI and SMP were based on the average value of three LIS models outputs. SMP values were calculated at the top 1 meter for 29-days time window. More specifically, the soil moisture data of two weeks backward and onward

time window were added to the data of the target day soil moisture to form the data pool for this date in order to compute SMP. Lastly, PDSI data was obtained from Abatzoglou (2019) in 1/24 degree, which then was projected and resampled to the NLDAS extents. Altogether, throughout the entire domain, every grid cell holds 731 values for each index where each was calculated for the dates that the USDM weekly maps between 2000 to 2013 were published.

3.2.1.3 Predictors Grouping

Different groups of predictors were used to fit the models so that the impact of different combinations of predictors on the model prediction abilities could be assessed. One of the commonly used terms in this study is the past week USDM drought category or $USDM_{t-1}$. Here, *t* is time with weekly intervals, so *t-1* takes place a week lagging from the current time. As drought phenomena is a slow-moving process, the likelihood of switching the drought condition from current week to next week is usually low. Considering that fact, we attempt to examine the proposed models performances with inclusion or exclusion of the $USDM_{t-1}$ as an input feature, in order to find this feature importance in the prediction tasks. Moreover, by the idea of using past week drought condition as a predictor, we aim to discover how the USDM experts, aside from the use of all the physical indicators in quantifying the drought categories, would also reflect the past week drought condition as a basis in producing the current week drought map.

Toward this purpose, we defined five groups of predictors which are presented in Figure 2. It shows how different combinations of inputs (in color) supply each group of predictors. Group 1 consists the eight drought indices while Group 2 includes the past week USDM drought condition in addition to Group 1 data as one more extra predictor. In contrast to Groups 1 and 2 which solely use the target grid cell information, Groups 3 and 4 include the information of the eight neighboring grid cells as supplementary data. In other words, these Groups of data contains a three

by three matrix of grid cells, centered on the target grid cell with nine times more data points. Similar to Group 1, Group 3 includes only the eight drought indices while Group 4 includes the past week USDM data of the grid cells as an additional predictor. Accordingly, one of the five groups of data is imported in the persistence model. This group of data only takes $USDM_{t-1}$ data and contributes in the baseline model. The baseline model is explained in the modeling section thoroughly.

After grouping the data, standardization of the drought indices values as well as encoding the categorical variable (i.e. $USDM_{t-1}$) were completed prior to inputting them in the models. Altogether, we attempt to predict the USDM drought condition labels for each grid cell by five different groups of input in the modeling. The schematic of the prepared data for the modeling in the entire domain is presented in Figure 3-3.

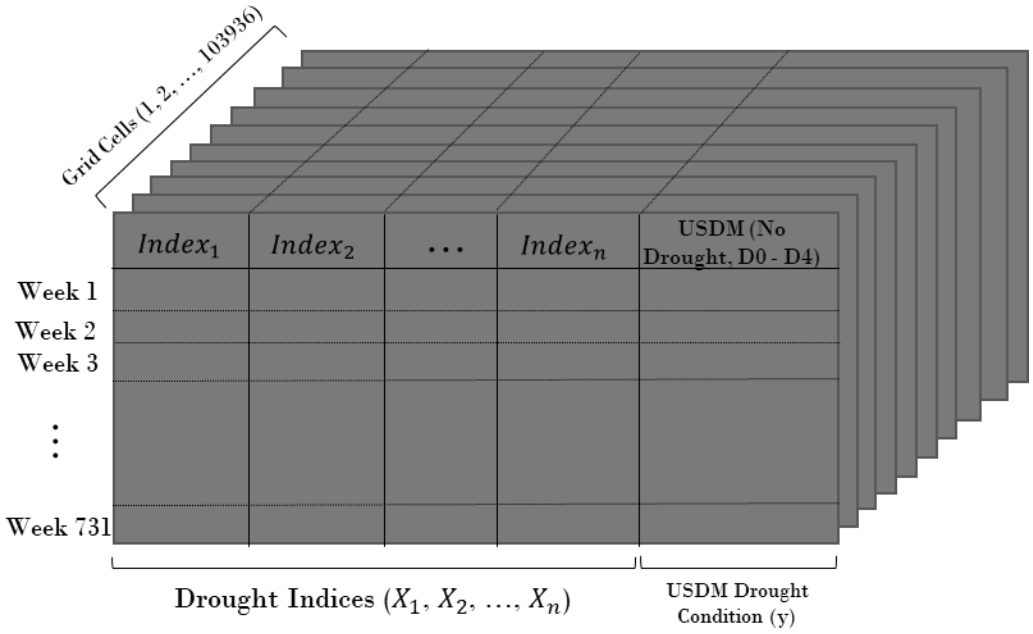


Figure 3-3 Schematic of the produced data domain

3.2.2 Modeling

3.2.2.1 Persistence Model

In machine learning context, generally the performance of an algorithm is compared against a simple and basic method called a baseline model. The performance metric (e.g. accuracy) will then become a benchmark to compare any other machine learning algorithm against. In this study a persistence model plays the role as the baseline model. We define a persistence model as a model which assumes the current week drought condition persists in next week. In other words, the model predicts the USDM drought category of an area for a specific week as its past week drought category. In this study, the rationale for using a persistence model as the baseline model is the slow-moving nature of drought, hence the probability of a drought (or wetness) condition persisting in the next weeks could have a relatively high likelihood. Obviously, the persistence modeling for the areas with more weekly variation in drought category is subject to more prediction error. Figure 3-2 shows how the corresponding input data is being carried over to the persistence model.

3.2.2.2 Machine Learning Models

Prediction of the USDM categories is an ordinal classification problem, as it is a forced choice for the models to predict six discrete responses, No Drought, D0, D1, D2, D3 and D4. Toward this purpose, three machine learning algorithms, logistic regression, random forest classifier and support vector machines (SVM) are selected to be examined for classification prediction.

The logistic regression model is used as a linear classification algorithm which uses the sigmoid function to limit the output of a linear equation between 0 and 1 as the probability outcome of the default class (Hosmer Jr et al., 2013). The estimation of the algorithm coefficient must be done on

training data using maximum likelihood. Logistic regression is a widely used classification technique due it its computational efficiency and being easily interpretable.

Random Forest (Ho, 1995) have successfully been implemented in various classification problems (banking, image classification, stock market, medicine and ecology) and is one the most accurate classification algorithms that works well with large datasets. The Random Forest classifier is a nonlinear classifier which consists an ensemble of decision tree classifiers. Each classifier is generated by a random set of features sampled independently from the input features, and each tree deposits a unit vote for the most suitable class to classify an input vector (Breiman, 2001). There are not many hyperparamters and they are easy to understand. Although, one of the major challenges in machine learning is overfitting, but the majority of the time this will not occur to a Random Forest classifier as there are sufficient trees in the forest.

SVM are broadly used as a classification tool in a variety of areas. They aim to determine the position of decision boundaries that produce the most optimum class separation (Cristianini & Shawe-Taylor, 2000). In classification, a maximal margin hyper-plane separates a specified set of binary labeled training data. However, if there is no possible linear features separation, SVM employ the techniques of kernels to make them linearly separable after they are mapped to a high dimensional feature space. The two standard kernel choices are polynomial and Radial Basis Function (RBF). In this study, we use an RBF kernel in SVM classifiers since RBF kernel is more capable compared to polynomial in representing the complex relationships in data especially the synergic complexities associated with growing data.

3.2.2.3 5-fold Cross-Validation

With the use of each machine learning algorithm and group of input variables, for each grid cell in the domain we build its own specific models. In all the three modeling algorithms, choosing the

optimal learning parameter(s) of the models known as “hyperparameter tuning” was performed by splitting the data to 80% training and 20% testing and executing 5-fold cross-validation on the training to select the best model. The logistic regression has only one hyperparameter, C with an $L2$ regularization (i.e. squared degree of coefficient as penalty term to loss function) in this study. For the Random Forest we set 200 trees as the number of estimators and then search in the parameter grid of maximum features, maximum depth, minimum samples leaf to find the optimum combination. Lastly, the hyperparameters in SVM with RBF kernel are C and γ, where C is the penalty of the objective function for misclassification and γ is the parameter of the kernel which controls the tradeoff between error of bias and variance in the model.

3.2.2.4 Metric of Performance Assessment

In this study, F_1 Score and Heidke Skill Score (HSS) are selected as the metric to evaluate the model performance. F_1 Score is usually more useful than Accuracy, especially when there are uneven class distributions. This is the case in our study as in general the number of weeks that the area of a grid cell may experience the extreme (D3) or the exceptional (D4) drought is far less than the rest of the drought categories, while the number of No Drought, D0, D1 and D3 are not equal either. F_1 Score is defined as the harmonic average of precision and recall (Goutte & Gaussier, 2005):

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

The Heidke skill score (Heidke, 1926) also known as kappa Index of Agreement (KIA) is a skill score for categorical prediction which presents the improvement of the prediction over the standard forecast which is usually a chance forecast. The range of the HSS is -∞ to 1 where a negative score suggests that the chance forecast is better, 0 means no skill, and a perfect forecast is equal to 1. For a multi-category event HSS is defined as (Barnston, 1992):

Table 3-3 Heidke Skill Score

	Observed Category				
Forecast Category	1	2	…	n	$\sum$ Forecast
1	X_{11}	X_{12}	…	X_{1n}	$\sum$ f1
2	X_{21}	X_{21}	…	X_{2n}	$\sum$ f2
…	…	…	…	…	…
n	X_{n1}	X_{n2}	…	X_{nn}	$\sum$ fn
$\sum$ Observation	$\sum$ O1	$\sum$ O2	…	$\sum$ On	Total

$$\mathrm{HSS} = \frac{\sum X_{ii} - \sum(X_{in}X_{ni})/Total}{Total - \sum(X_{in}X_{ni})/Total} \qquad (4)$$

Depending on the aforementioned five different groups of data, the number of the indices (i.e., X) can be one for the persistence model, eight (Groups 1 and 3) or nine (Groups 2 and 4) for the machine learning models. Then the models would predict the dependent variable (y) which is the USDM drought categories. Once the Groups 1, 2, 3 and 4 input features of each grid cell are modeled and tested using the three proposed classifiers, twelve different accuracy outcomes are produced. Ultimately, all the outcomes are compared against the persistence model accuracy one by one (grid cell by grid cell) across the entire domain so that the best general combination in terms of group of features and algorithm performance is determined.

3.3 Results and Discussion

Our introductory analysis to the data was to explore the spread of different drought conditions across the domain during the 731 weeks. By utilizing the outcome, we can better perceive the contribution associated to the number of data points with the models prediction performance. Out of 103,936 grid cells within the domain, 51,997 grid cells never experienced any USDM drought condition during the 14 years of data which means they were always labeled as No Drought or were not in the USDM weekly maps CONUS domain. The remaining 51,939 grid cells have experienced both D0 and D1 drought categories at least once during that time period. Therefore, in our classification task, there were at least three different classes, No Drought, D0 and D1 which are to be predicted. However, for the grid cells experiencing more of the drought conditions other than D0 and D1, the prediction is a multi-class classification task of four or more classes. During 731 weeks of the USDM data, there were 50,546 grid cells experiencing D2 (as well as No Drought, D0 and D1), 44,203 grid cells experiencing D3 (in addition to No Drought, D0, D1, and D2) and 24,210 grid cells experiencing D4 (along with No Drought, D0, D1, D2 and D3) at least once. Figure 3-4 presents the histograms of each drought category throughout the entire domain. The included grid cells in the histograms are out of those 51,939 which have experienced more than one type of USDM drought condition. From the histograms we can observe as the drought conditions become more severe (from No Drought to D4), the grid cell mean count of the categories decrease from 369.51 for No Drought down to 25.01 for D4.

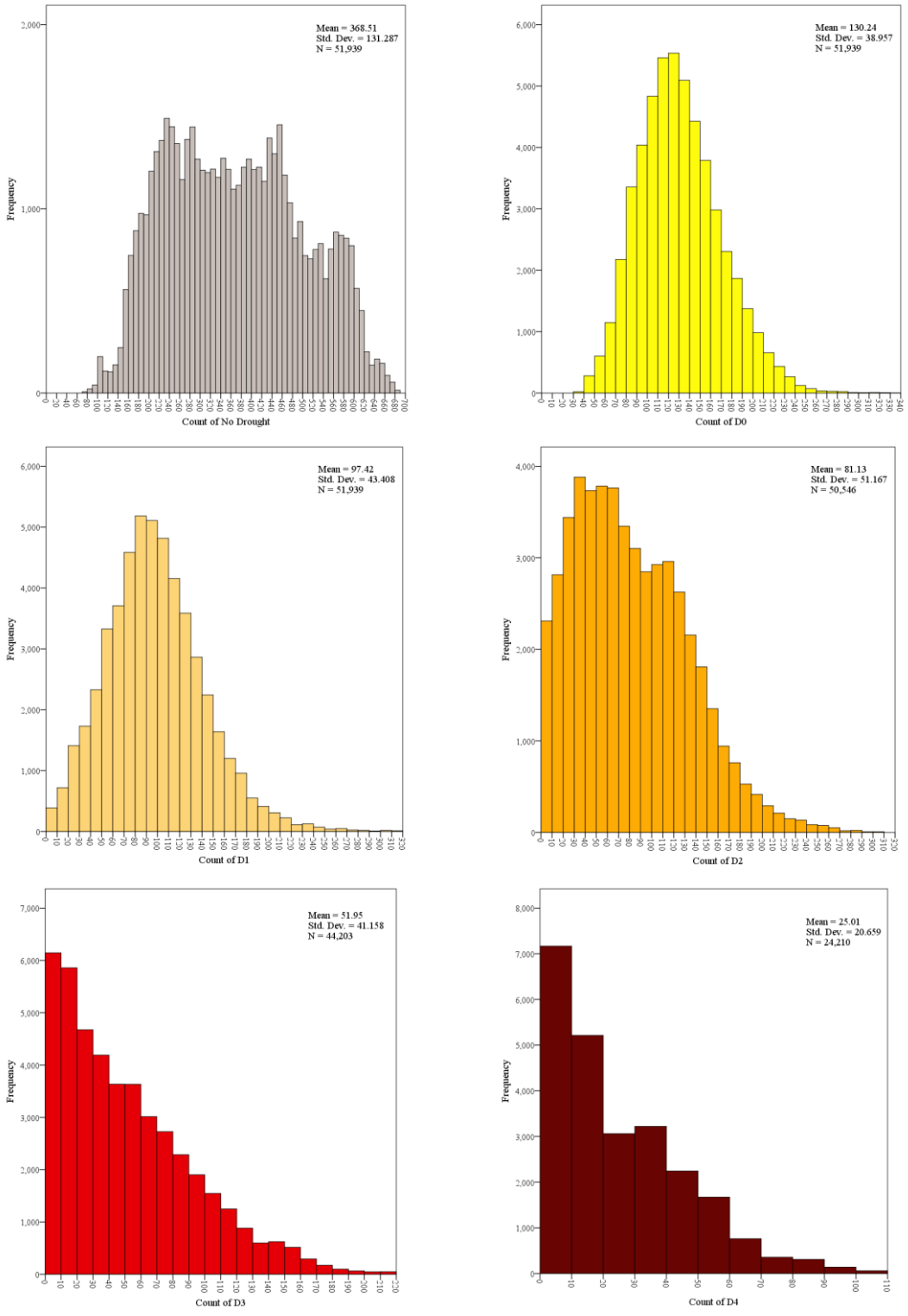


Figure 3-4 Histograms of the USDM drought categories counts across the domain in 14 years

3.3.1 Persistence Model

As we discussed earlier, the group of data which solely contained the $USDM_{t-1}$ was the input of the baseline predictive model known as the persistence model. The persistence model overall performance is presented in Tables 3-4 and 3-5, including the minimum, maximum and mean prediction F_1 scores and HSS for every USDM drought category as well as the weighted average F_1 score.

Table 3-4 Persistence model descriptive statistics over the entire domain

	Min F_1	Max F_1	Mean F_1	Std. Dev
No Drought	0.90	0.99	0.96	0.01
D0	0.42	0.96	0.81	0.08
D1	0	0.98	0.83	0.09
D2	0	0.99	0.84	0.11
D3	0	0.99	0.85	0.14
D4	0	0.99	0.83	0.20
Weighted Average	0.81	0.97	0.91	0.03

Table 3-5 Persistence model Heidke Skill Score

	Min HSS	Max HSS	Mean HSS	Std. Dev
No Drought	-0.01	1	0.88	0.17
D0	0.22	1	0.78	0.11
D1	-0.02	1	0.80	0.12
D2	-0.02	1	0.84	0.14
D3	-0.02	1	0.86	0.18
D4	-0.02	1	0.83	0.26

The results in Tables 3-4 and 3-5, show that the persistence model prediction score for all the classes and the weighted average is relatively high. This is basically an endorsement for the slow-moving nature of drought so that a persistence model achieves such high scores at all levels.

The persistence model performs worse in the areas with more drought weekly fluctuations since an alteration in the drought condition from the current week to the next corresponds to one prediction error for the model. Furthermore, the standard deviation of the accuracies from No Drought to D4 constantly increases, yet the Weighted Average standard deviation (in Table 3-4) stays as small as 0.03 because of the larger weights of the less severe drought conditions in contrast to D3 and D4 categories.

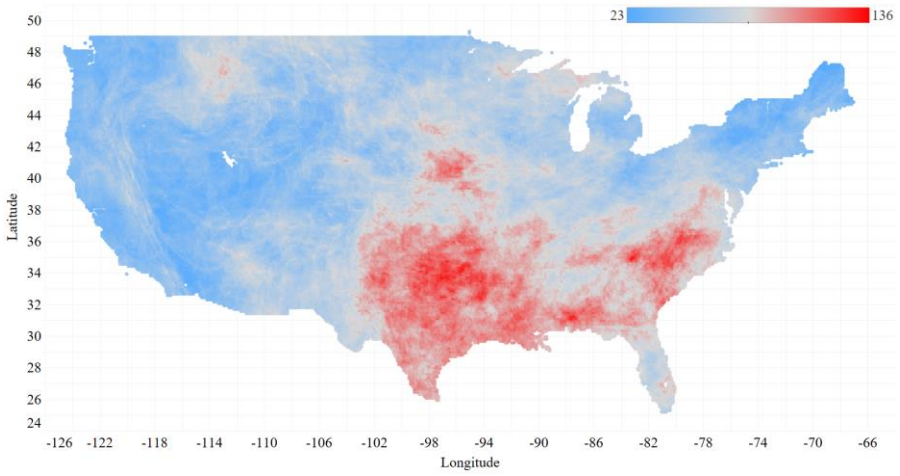


Figure 3-5 Spatial presentation of the number of weekly fluctuations for each grid cell during 731 weeks

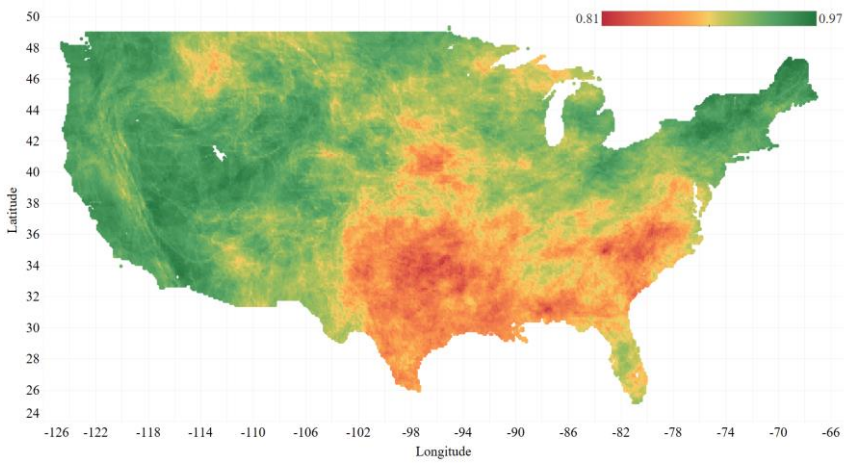


Figure 3-6 Spatial distribution of the persistence model weighted average F1 Score across the domain of study

The spatial distribution of the grid cells weekly fluctuation is presented in Figure 3-5 showing the lowest variation between the USDM drought categories during 731 weeks of data is 23, while the largest is 136. As we can see, the highest weekly fluctuations are located in Southeast and Plains areas where the climate is warm temperature, humid with hot summers (Kottek et al., 2006). Figure 3-6, on the other hand, displays the persistence model weighted average F_1 score across the domain. As it is noticeable from the Figures, the spread patterns of colors look similar, however, in the opposite direction displaying the message that the areas with more weekly fluctuations achieve less prediction accuracy by the persistence model and vice versa.

3.3.2 Machine Learning Models

3.3.2.1 Results for Using Group 1

In this section, we present and discuss the results of the logistic regression, Random Forest and SVM using four different Groups of input data. Tables 3-6 and 3-7 contain the summary of the obtained scores for entire domain by three models by running on the Group 1 data. As we can see, the nonlinear models (i.e., Random Forest and SVM) substantially perform better than the linear model (i.e. logistic regression), while the highest scores as well as the average score are obtained by SVM for all the drought categories. However, none of the models can reach the scores that were obtained by the persistence model by any means, neither for any of the six drought classes, nor on average. Moreover, the scores standard deviations of all three models are more than the baseline model so the prediction accuracies are also less consistent.

Table 3-6 Descriptive statistics of the models performances using Group 1 input features over the entire domain

	Logistic Regression				Random Forest				SVM			
	Min F_1	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**
No Drought	0.00	1.00	0.76	0.17	0.00	1.00	0.85	0.12	0.00	1.00	0.85	0.15
D0	0.00	0.78	0.20	0.18	0.00	0.92	0.55	0.13	0.00	0.94	0.60	0.12
D1	0.00	1.00	0.18	0.22	0.00	1.00	0.56	0.17	0.00	1.00	0.60	0.16
D2	0.00	1.00	0.24	0.26	0.00	1.00	0.60	0.20	0.00	1.00	0.63	0.19
D3	0.00	1.00	0.28	0.31	0.00	1.00	0.59	0.27	0.00	1.00	0.65	0.24
D4	0.00	1.00	0.44	0.39	0.00	1.00	0.63	0.34	0.00	1.00	0.70	0.29
Weighted Average	0.11	0.95	0.54	0.15	0.43	0.97	0.75	0.07	0.47	0.98	0.77	0.07

Table 3-7 Heidke Skill Score descriptive statistics of the models performances using Group 1

	Logistic Regression				Random Forest				SVM			
	Min HSS	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**
No Drought	-0.25	1	0.47	0.18	-0.04	1.00	0.70	0.13	-0.04	1.00	0.70	0.14
D0	-0.28	0.74	0.11	0.14	-0.07	0.94	0.48	0.14	-0.08	0.92	0.52	0.13
D1	-0.25	1	0.12	0.17	-0.07	1.00	0.51	0.17	-0.10	1.00	0.55	0.16
D2	-0.22	1	0.19	0.23	-0.06	1.00	0.56	0.21	-0.06	1.00	0.59	0.19
D3	-0.15	1	0.27	0.31	-0.05	1.00	0.60	0.27	-0.05	1.00	0.65	0.24
D4	-0.06	1	0.43	0.38	-0.05	1.00	0.62	0.35	-0.03	1.00	0.67	0.31

3.3.2.2 Results for Using Group 2

The results of the modeling with the Group 2 data set are presented in Tables 3-8 and 3-9. All three models especially the logistic regression demonstrate a great improvement over the Group 1 input feature just by adding $USDM_{t-1}$ as another variable. This indicates that the models learned to put a great weight on the extra added variable which is revealed to play an important role in terms of

improving the model's prediction capabilities. Also, the models can surpass the persistence model performance in D4 prediction score, however, on average all of them achieved an equal F_1 score of 0.91. However, the HSS for all three models when compared to the persistence model, we realize that the scores are still lower or at the best equal which means no superiority proven by the machine learning models.

Table 3-8 Descriptive statistics of the models performances uing Group 2 input features over the entire domain

	Logistic Regression				**Random Forest**				**SVM**			
	Min F_1	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**
No Drought	0.00	1.00	0.94	0.13	0.00	1.00	0.94	0.13	0.00	1.00	0.94	0.14
D0	0.00	1.00	0.81	0.10	0.00	1.00	0.80	0.11	0.00	1.00	0.81	0.10
D1	0.00	1.00	0.82	0.14	0.00	1.00	0.81	0.15	0.00	1.00	0.82	0.15
D2	0.00	1.00	0.83	0.18	0.00	1.00	0.83	0.17	0.00	1.00	0.83	0.17
D3	0.00	1.00	0.83	0.23	0.00	1.00	0.83	0.22	0.00	1.00	0.84	0.22
D4	0.00	1.00	0.84	0.26	0.00	1.00	0.85	0.25	0.00	1.00	0.85	0.26
Weighted Average	0.77	0.99	0.91	0.03	0.75	0.99	0.91	0.03	0.75	0.99	0.91	0.03

Table 3-9 Heidke Skill Score of models performances using Group 2

	Logistic Regression				**Random Forest**				**SVM**			
	Min HSS	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**
No Drought	-0.01	1.00	0.89	0.14	-0.01	1.00	0.89	0.14	-0.01	1.00	0.89	0.15
D0	-0.01	1.00	0.78	0.11	0.10	1.00	0.77	0.12	0.12	1.00	0.78	0.11
D1	-0.02	1.00	0.80	0.14	-0.04	1.00	0.79	0.14	-0.03	1.00	0.80	0.14
D2	-0.02	1.00	0.83	0.18	-0.02	1.00	0.82	0.18	-0.02	1.00	0.83	0.17
D3	-0.02	1.00	0.84	0.21	-0.02	1.00	0.84	0.21	-0.02	1.00	0.85	0.20
D4	-0.01	1.00	0.80	0.29	-0.01	1.00	0.80	0.29	-0.01	1.00	0.81	0.29

During the model training with Group 1 and 2 data, the range of the F_1 scores of the Random Forest and SVM varied from 0.99 to 1 which was almost perfect. However, on the testing data, the scores dropped down to 0.91 to show an overfitting problem. This type of challenge sometimes happens in nonlinear models when the number of data points compared to the number of features are small even though a cross validation is used. By observing the histograms in Figure 3-4, we could find the reason nested in the average count of D3 and D4 drought categories which are usually low. With using 80% of data in training even though randomly selected, the chances of a model seeing fewer of those categories during learning process become higher. This causes the models memorize instead of learn so while testing, the scores are not as promising as training.

With the use of Group 2 data in the modeling, on average in 31732 grid cells (61% of the domain) logistic regression performed better than or equal to the persistence model. This is the case for the Random Forest model in 27139 grid cells (52% of the domain) and in 31085 grid cells (60% of the domain) for the SVMs. Adding the past week information to the data, helped the models to improve their prediction accuracy, however, it was still challenging to be assertive about outperforming the baseline model. With the presumption that lack of data point may be the cause of underperformance, we tried the Groups 3 and 4 in the models so that we could possibly find out whether there would be any improvement in prediction accuracy.

3.3.2.3 Results for Using Group 3

The performance of the machine learning models without $USDM_{t-1}$ label as the predictor, yet with borrowing the neighboring grid cells which created Group 3 data were examined and are summarized in Tables 3-10 and 3-11.

Table 3-10 Descriptive statistics of the models performances using Group 3 input features over the entire domain

	Logistic Regression				Random Forest				SVM			
	Min F_1	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**
No Drought	0.00	0.98	0.78	0.14	0.00	1.00	0.93	0.06	0.00	1.00	0.94	0.08
D0	0.00	0.72	0.24	0.16	0.07	0.96	0.79	0.07	0.00	0.99	0.83	0.06
D1	0.00	1.00	0.21	0.20	0.00	1.00	0.79	0.08	0.00	1.00	0.83	0.07
D2	0.00	1.00	0.28	0.24	0.00	1.00	0.80	0.11	0.00	1.00	0.84	0.09
D3	0.00	1.00	0.31	0.30	0.00	1.00	0.80	0.15	0.00	1.00	0.84	0.13
D4	0.00	1.00	0.44	0.36	0.00	1.00	0.78	0.24	0.00	1.00	0.83	0.20
Weighted Average	0.17	0.93	0.55	0.14	0.61	0.99	0.87	0.05	0.63	1.00	0.90	0.05

Table 3-11 Heidke Skill Score of the models using Group 3

	Logistic Regression				Random Forest				SVM			
	Min HSS	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**
No Drought	-0.17	1.00	0.50	0.16	0.00	1.00	0.90	0.12	-0.01	1.00	0.92	0.14
D0	-0.21	0.80	0.14	0.14	0.00	1.00	0.83	0.10	0.06	1.00	0.87	0.08
D1	-0.21	1.00	0.15	0.18	-0.02	1.00	0.84	0.11	-0.01	1.00	0.88	0.10
D2	-0.20	1.00	0.22	0.24	-0.02	1.00	0.86	0.14	-0.03	1.00	0.89	0.11
D3	-0.15	1.00	0.30	0.33	-0.02	1.00	0.87	0.17	-0.02	1.00	0.90	0.15
D4	-0.05	1.00	0.48	0.39	-0.02	1.00	0.87	0.24	-0.02	1.00	0.90	0.20

The weighted average accuracy of the logistic regression dropped significantly once again when the past week information predictor was eliminated. Despite the importance of the eliminated predictor, the nonlinear models, Random Forest and SVM could sustain fairly close to the persistence model on average but still lower, with 0.87 and 0.90 F_1 prediction score, respectively. The results showed that the SVM model with Group 3 data could predict better than the persistence model for D0, while it had an equal score but less standard deviation for D1 and D2, and an equal

score and standard deviation for D4. However, comparing the HSS shows that our nonlinear models could outperform the persistence model in six classes in terms of mean score with less deviations in prediction score.

When compared the weighted average F_1 score across the entire domain, the logistic regression could not defeat the persistence model prediction scores in any of the grid cells, while the Random Forest and SVM were successful in 17,385 (33% of the grid cells) and 27,743 (53% of the grid cells), respectively. The results of modeling with Group 3 indicates that by employing the neighboring grid cells data and consequently a larger training set, we could improve the models, particularly the nonlinear ones, to capture the relationships between the variables and drought categories more precisely. However, the feature $USDM_{t-1}$ still illustrates a stronger impact than the size of training data when Group 2 and 3 are compared side by side.

3.3.2.4 Results for Using Group 4

The results of using Group 4 dataset in the modeling is presented in Tables 3-12 and 3-13. Compared to the Groups 1, 2 and 3 results, there is a noticeable improvement in both F_1 score and HSS, for the Random Forests and SVM. The logistic regression performs slightly better than the persistence model in the prediction score for the categories, however the F_1 weighted average scores are equal. On the other hand, the Random Forest and SVM outperform the persistence model in all the categories and F_1 weighted average scores with the highest scores equal to 0.96 achieved by the SVM. Using Group 4 of data, indicates that borrowing the neighboring grid cells information and including $USDM_{t-1}$, could certainly and significantly help the models learning curve to improve. Clearly, the lack of data points was preventing the models to capture a more comprehensive pattern while just using one single grid cell data.

Table 3-12 Descriptive statistics of the models performances using Group 4 input features over the entire domain

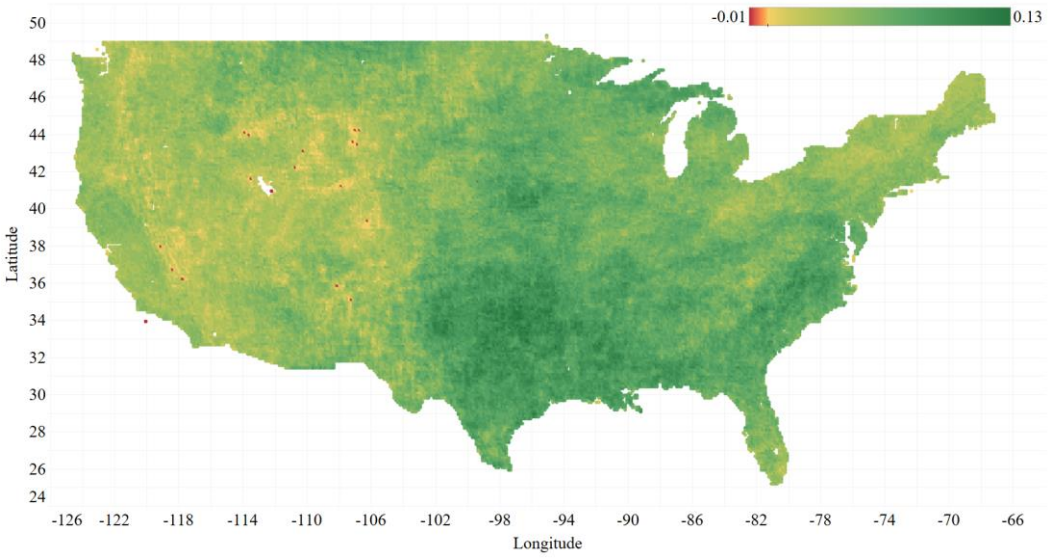
	Logistic Regression				Random Forest				SVM			
	Min F_1	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**	**Min F_1**	**Max F_1**	**Mean**	**Std. Dev**
No Drought	0.00	1.00	0.96	0.05	0.00	1.00	0.98	0.04	0.00	1.00	0.98	0.06
D0	0.36	0.96	0.81	0.08	0.51	0.99	0.90	0.04	0.61	1.00	0.93	0.03
D1	0.00	1.00	0.83	0.09	0.00	1.00	0.91	0.05	0.00	1.00	0.93	0.04
D2	0.00	1.00	0.85	0.11	0.00	1.00	0.92	0.08	0.00	1.00	0.94	0.07
D3	0.00	1.00	0.87	0.14	0.00	1.00	0.92	0.11	0.00	1.00	0.93	0.10
D4	0.00	1.00	0.85	0.20	0.00	1.00	0.90	0.18	0.00	1.00	0.91	0.17
Weighted Average	0.82	0.98	0.91	0.03	0.86	1.00	0.95	0.02	0.87	1.00	0.96	0.01

Table 3-13 Heidke Skill Score of the models using Group 4 data

	Logistic Regression				Random Forest				SVM			
	Min HSS	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**	**Min HSS**	**Max HSS**	**Mean**	**Std. Dev**
No Drought	-0.01	1.00	0.90	0.13	0.00	1.00	0.96	0.08	-0.01	1.00	0.97	0.09
D0	0.22	1.00	0.79	0.11	0.46	1.00	0.91	0.06	0.34	1.00	0.95	0.05
D1	-0.01	1.00	0.81	0.12	-0.01	1.00	0.93	0.07	-0.01	1.00	0.95	0.05
D2	-0.02	1.00	0.85	0.14	-0.01	1.00	0.94	0.09	-0.02	1.00	0.96	0.07
D3	-0.02	1.00	0.87	0.17	-0.02	1.00	0.94	0.12	-0.01	1.00	0.96	0.10
D4	-0.02	1.00	0.86	0.23	-0.01	1.00	0.93	0.18	-0.01	1.00	0.95	0.16

By looking into the one by one obtained F_1 weighted average scores for each grid cell across the domain, on average in 10794 grid cells (21% of the domain) the logistic regression performs worse than the baseline model, whereas in only 419 grid cells (0.8% of the domain) the random forest performs worse than the persistence. The SVM with the best results, misclassified just 18 grid cells with 1 percent difference weighted average score compared to the persistence model. Figure 3-7 shows the color map of the spatial distribution of the difference between the SVM and persistence model average score all over the domain. Aside from those 18 points with -0.01 accuracy

difference, the rest of them vary between 0 to 0.13. From Figures 3-6 and 3-7, it can be observed that in the Southeast and Plains areas which the persistence model performs worse (i.e. higher weekly fluctuation) the SVM model showed a larger difference in the prediction accuracy. We will discuss more about the comparison of the models predictions against the persistence model in boxplots later on in this section, although the purpose of outperforming the persistence model by the machine learning models has been met.



1)

Figure 3-7 Spatial distribution of the weighted average F1 Score difference between the Group4-SVM and persistence model

3.3.2.5 Side-by-Side Boxplot Comparison of the Model Performance Using Different Groups of Data

In this section, we present and discuss the performances of all the 13 different types of modeling in this study, next to each other in the format of boxplots. Figures 3-8 provides a side-by-side overall performance of the models while Figure 3-9 contains the results of the models for each USDM category. In the boxplots, the box middle line, bottom line and top line are the median, 25th percentile and 75th percentile, respectively. The whiskers extend 1.5 times the height of the box

(Interquartile range or IQR), and the points are extreme outliers which are three times greater than the IQR. From Figure 3-8, we could clearly find out that the USDM drought labels were better predicted by the nonlinear functions in terms of accuracy and deviation. The linear model fulfilled a meaningfully better prediction with the presence of the $USDM_{t-1}$ information as a predictor (Groups 2 and 4). The importance of this predictor can be observed by comparing the Groups 2 and 3 results, while modeling with Group 2 could obtain better results than Group 3, even with using fewer number of data points. The detected pattern in Figure 3-8 can be observed in all the six categories of Figure 3-9 where Group 4 performs the best, followed by Group 2, then Groups 3 and lastly Group 1.

In terms of feature importance in the models, both the logistic regression and random forest commonly recognized PDSI as the most important predictor in Group 1 and Group 3, while in Groups 2 and 4 $USDM_{t-1}$ received the largest coefficient, followed by PDSI as the second most important feature. The importance of the rest of the features in the models were relatively close to each other. Unfortunately, as the RBF kernel in SVM transforms the features into a high dimensional space, the implicit transformation does not allow us to obtain the feature importance.

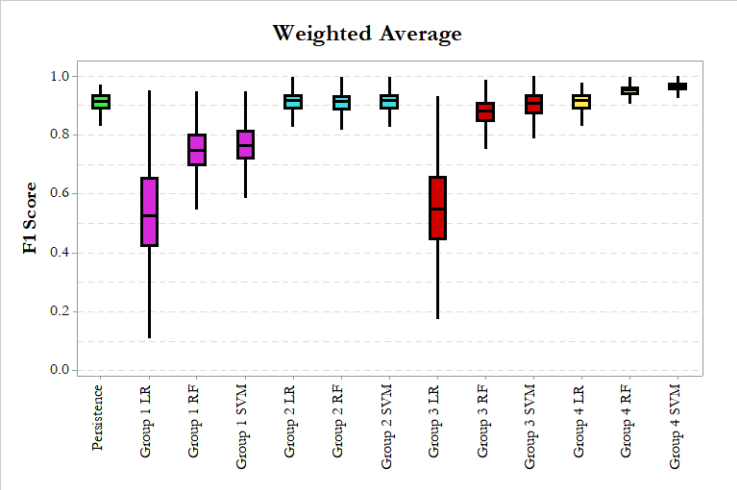


Figure 3-8 Side by side models' overall performances comparison

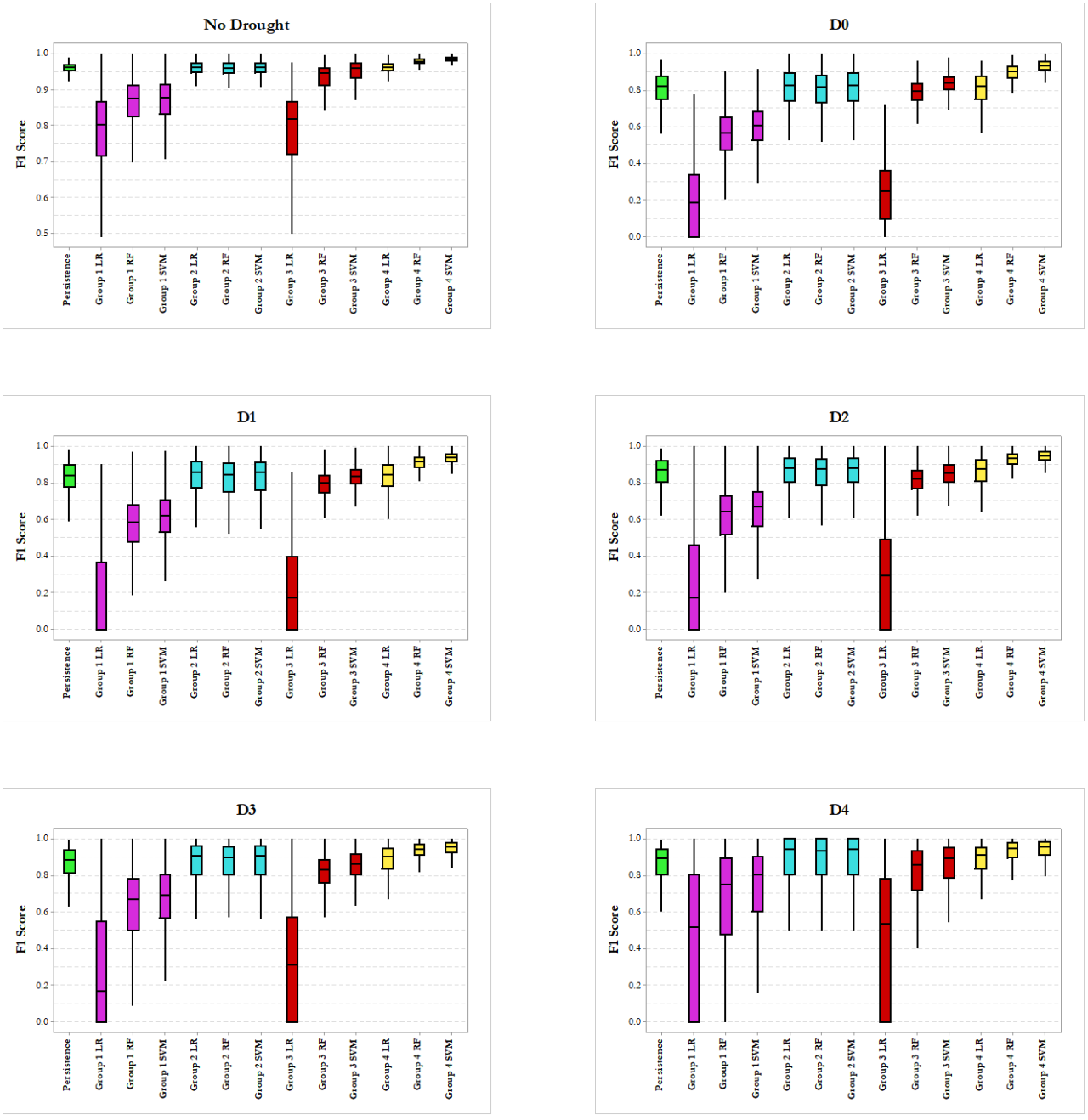


Figure 3-9 Side by side models' performances in prediction of each USDM drought category

For a better illustration in comparing the reproduced maps by the models and the actual USDM map, we also selected three random dates. In Figure 3-10 the actual USDM map is not experiencing any D4 but two spots of D3 in Midwest and Northwest regions, whereas Figure 3-11 shows the southern part of Plains experiencing D4 as well as D3. Figure 3-12, however, shows larger and more scattered areas of D3 and D4 across the domain.

3.3.2.6 Visual Comparison of the Models Over CONUS

In Figure 3-10, the persistence model can closely catch D3 areas however, it does not perform well in predicting the D0, D1 and D2 while the large areas of D0 are replaced with D1 and D2. This is possibly due to precipitations during the past week generated map date (9/27/2005) and the date of this map (10/04/2005) in which has made those areas drought severity one category less extreme. The generated maps from Groups 1 and 3 models do not look well reproduced except Group 3 RF and SVM, however, both still are not as smooth as expected. The entire Group 2 map plus Group 4 LR are very similar to the persistence model map which means the models are heavily relying on the $USDM_{t-1}$ as their predictors. Finally, the best performing model, Group 4 SVM is able to generate very similar map to the actual USDM map followed by Group 4 RF as the second-best model. If pay a closer attention, there is a slight difference between Group 4 RF and SVM in which RF is still mispredicting few spots on the map.

In Figure 3-11, the models produced more similar maps to the actual USDM compared to Figure 3-10 especially for Group 2. The similarity is due to less change from past week to this week which the persistence model is showing clearly. In other words, as it was discussed earlier, the models are using $USDM_{t-1}$ as the most important feature so once we have a more accurate persistence model (i.e. less change from past week) for a week, the rest of the models would be likewise more accurate. However, Group 4 RF and SVM could still generate the closest map to the actual

specifically in the regions with D3 and D4 as they are inherently capable of capturing nonlinear relationships while using more data for training.

Similar to Figure 3-11, Figure 3-12 has also a relatively similar persistence model to the actual USDM map except a few small areas such as not being able to recognize an D3 area in California and replacing a No Drought region in Indiana with D0. As it can be seen, the models in Groups 1 are not doing well, however, there is a significant improvement once the models are fed with $USDM_{t-1}$ as another feature in Group 2. The maps of Group 3 models are not as smooth, but we can see the above-mentioned areas that the persistence model was not able to catch are relatively being recognized by them especially by the SVM model. Lastly, the best performing model is Group 4 SVM which was able to produce almost as similar as the actual USDM map.

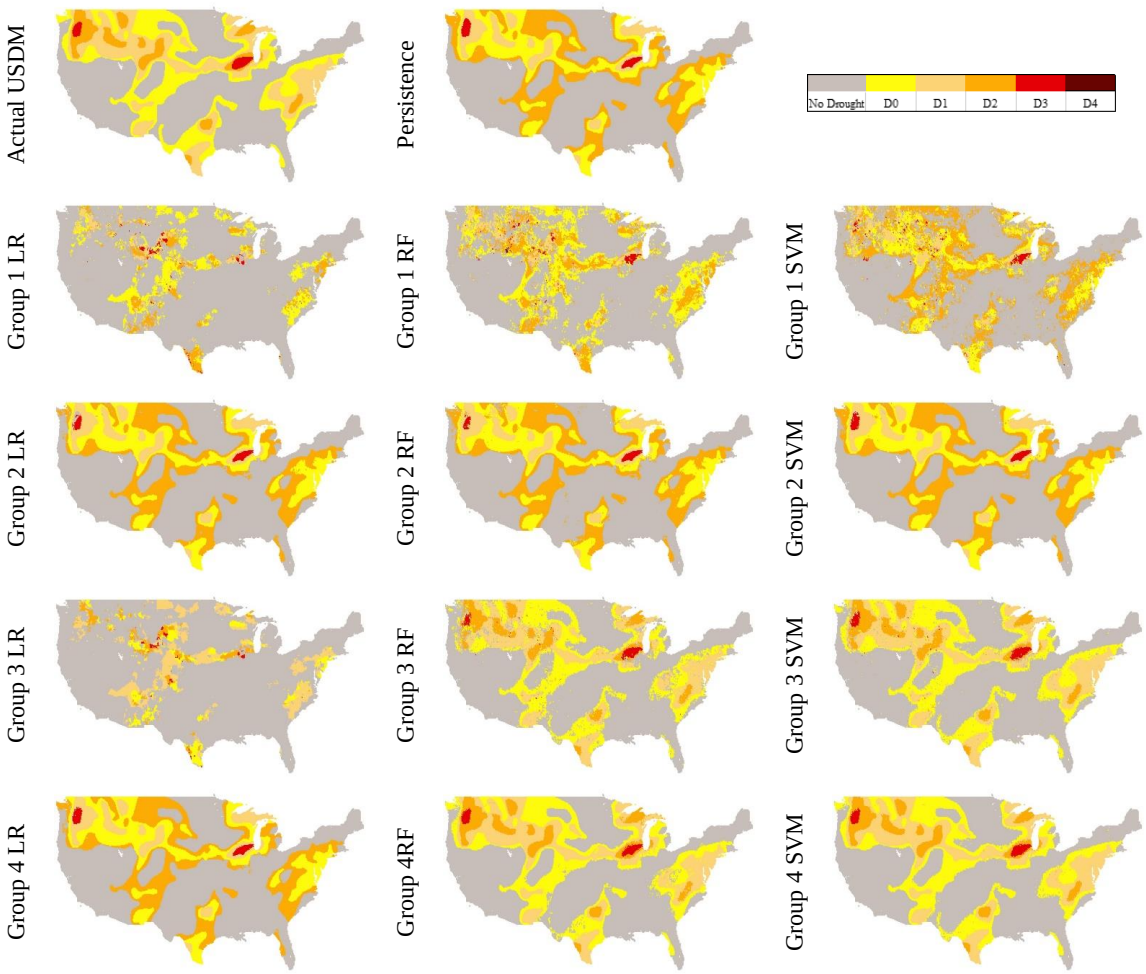


Figure 3-10 Produced maps 10/04/2005 by each model

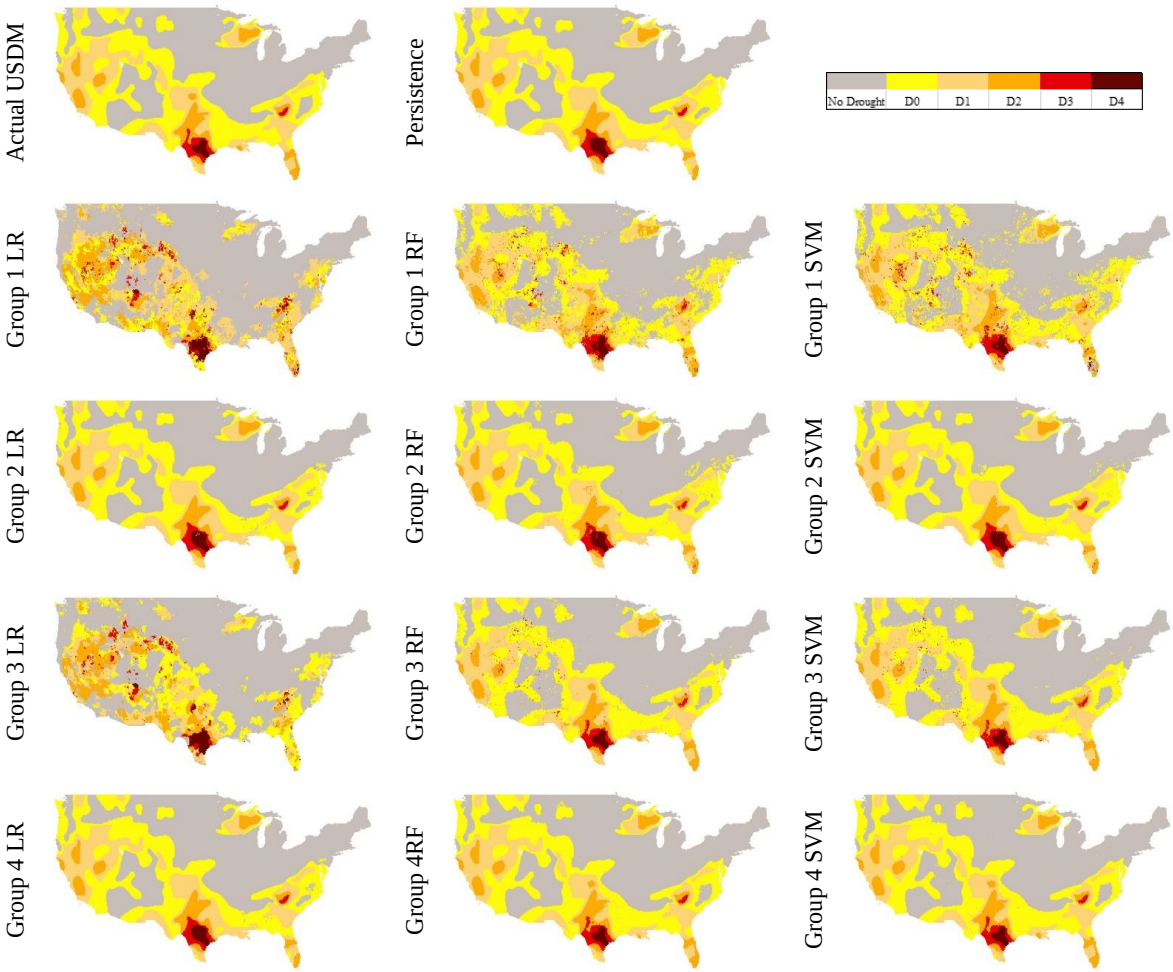


Figure 3-11- Produced maps of 03/17/2009 by each model

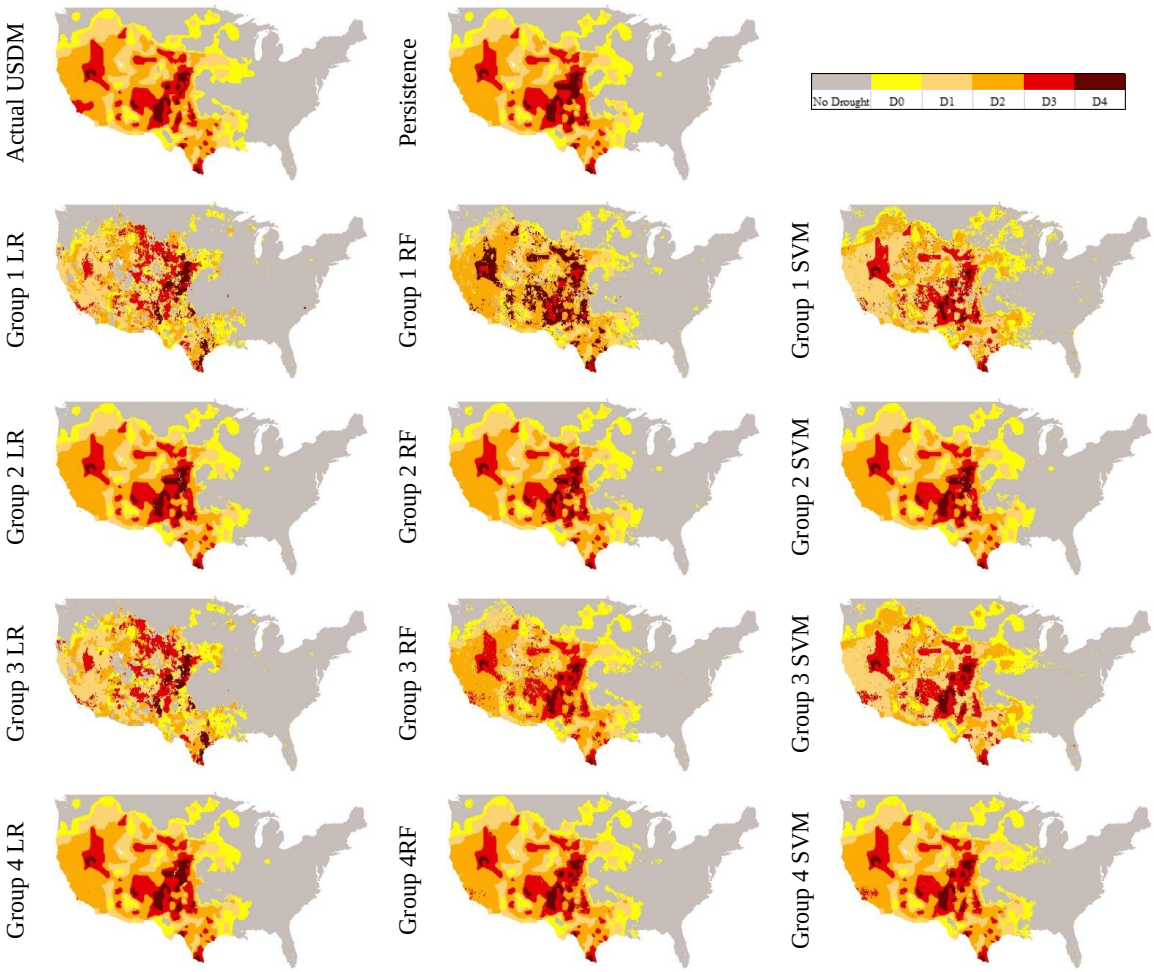


Figure 3-12 Produced maps of 08/13/2013 by each model

As the last step, in order to show a more in detail models comparison instead of the overall average performance, we selected a random sample grid cell located at the latitude of 35.0630, and longitude - 105.3130 (appeared to be in New Mexico) and put the test data from the years 2010 to 2013 in a time series graph. Figure 3-12 presents the actual test data into the models and each model prediction. It is notable that the graph is an ordered time series of the test data points, but the dates are not consecutive due to random selection of training and test set, while the weeks in between were used as training data for the models. Similar to the above generated maps, here the Group 2 models are significantly relying on the $USDM_{t-1}$ feature as whenever the persistence does

or does not predict correctly, Group 2 models are predicting accordingly. Group 1 models and Group 3 LR are the least consistent models, while Group 3 RF and SVM showed a fairly good performance even though they were not using $USDM_{t-1}$. Group 4 models show the best modeling results especially Group 4 SVM by being able to predict the date 8/23/2011 correctly where the majority of the models failed.

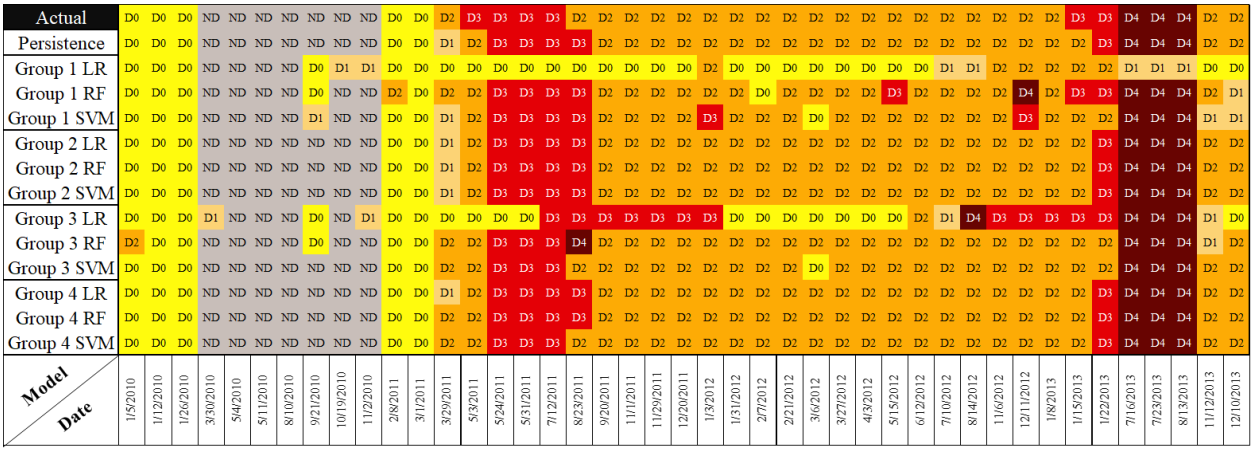


Figure 3-13 Time series of test data of grid cell located in (35.0629, -105.3130) New Mexico

3.4 Conclusions

Our proposed framework successfully reproduced the USDM drought categories using multiple drought indices and machine learning algorithms, logistic regression, Random Forest and SVM. The framework was compared to a persistence model as the baseline model in which it was assumed that current week drought condition would persist in next week. As this study was a classification task, the machine learning models were evaluated by their overall prediction scores (only for F1 score) as well as each class prediction score. Although, in terms of prediction accuracy, there was not much room left for improvement by the baseline model, our proposed framework could outperform it by testing different scenarios of the data inputs and machine learning algorithms to find the best combination.

We found out that employing the past week drought data as a predictor in the models played an important role in achieving high prediction scores especially for the logistic regression. The nonlinear models, Random Forest and SVM suffered less without the use of that predictor in terms of prediction score. Furthermore, taking the neighboring grid cells information into account, could compensate the lack of data points for training the models. It was essentially rectification of the temporal shortage of the available USDM data (731 weeks) by increasing it spatially. Training the models faced the lack of data problem particularly for the categories D3 and D4. In some grid cells when the number of D3 and D4 were smaller than the number of the folds in cross validation (i.e. 5 in this study) as well as random selection of training and test splits, technically some folds could not contain those categories during the learning process which resulted in poor predictive skill.

Future works could be the examination of a multi-task learning approach which works well with limited data by leveraging information from nearby locations. Also, since we have been successful in being close to mimicking the USDM experts drought categories synthesizing, this methodology could be used in an automated system in generating the weekly maps. The system would be using LSMs to produces the outputs which are needed to calculate the drought indices which represent meteorological, agricultural, and hydrologic drought. Thereafter by creating the indices for the target day that the map is going to be published and using the past week drought condition as another variable, the SVM model as the best performing model in this study would predict the drought conditions across the entire United States.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abatzoglou, J. T. (2019). GRIDMET. Retrieved from http://www.climatologylab.org/gridmet.html

Anderson, M. C., Hain, C., Otkin, J., Zhan, X., Mo, K., Svoboda, M., . . . Pimstein, A. (2013). An intercomparison of drought indicators based on thermal remote sensing and NLDAS-2 simulations with US Drought Monitor classifications. *Journal of Hydrometeorology, 14*(4), 1035-1056.

Anderson, M. C., Hain, C., Wardlow, B., Pimstein, A., Mecikalski, J. R., & Kustas, W. P. (2011). Evaluation of drought indices based on thermal remote sensing of evapotranspiration over the continental United States. *Journal of Climate, 24*(8), 2025-2044.

Barnston, A. G. (1992). Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting, 7*(4), 699-709.

Boken, V. K. (2005). Agricultural drought and its monitoring and prediction: some concepts. *Monitoring and predicting agricultural drought: A global study*, 3-10.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brown, J. F., Wardlow, B. D., Tadesse, T., Hayes, M. J., & Reed, B. C. (2008). The Vegetation Drought Response Index (VegDRI): A new integrated approach for monitoring drought stress in vegetation. *GIScience & Remote Sensing, 45*(1), 16-46.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press.

Dai, A. (2011). Drought under global warming: a review. *Wiley Interdisciplinary Reviews: Climate Change, 2*(1), 45-65.

droughtmonitor.unl.edu. (2018). USDM Map August 7, 2018. Retrieved from https://droughtmonitor.unl.edu/

droughtmonitor.unl.edu. (2019). What is the USDM? Retrieved from https://droughtmonitor.unl.edu/AboutUSDM/WhatIsTheUSDM.aspx

Goutte, C., & Gaussier, E. (2005). *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation.* Paper presented at the European Conference on Information Retrieval.

Gu, Y., Brown, J. F., Verdin, J. P., & Wardlow, B. (2007). A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States. *Geophysical research letters, 34*(6).

Hao, Z., & AghaKouchak, A. (2014). A nonparametric multivariate multi-index drought monitoring framework. *Journal of Hydrometeorology, 15*(1), 89-101.

Hao, Z., Hao, F., Xia, Y., Singh, V. P., Hong, Y., Shen, X., & Ouyang, W. (2016). A statistical method for categorical drought prediction based on NLDAS-2. *Journal of Applied Meteorology and Climatology, 55*(4), 1049-1061.

Hao, Z., Hong, Y., Xia, Y., Singh, V. P., Hao, F., & Cheng, H. (2016). Probabilistic drought characterization in the categorical form using ordinal regression. *Journal of hydrology, 535*, 331-339.

Hao, Z., Yuan, X., Xia, Y., Hao, F., & Singh, V. P. (2017). An overview of drought monitoring and prediction systems at regional and global scales. *Bulletin of the American Meteorological Society, 98*(9), 1879-1896.

Hayes, M., Svoboda, M., Wall, N., & Widhalm, M. (2011). The Lincoln declaration on drought indices: universal meteorological drought index recommended. *Bulletin of the American Meteorological Society, 92*(4), 485-488.

Hayes, M. J., Svoboda, M. D., Wardlow, B. D., Anderson, M. C., & Kogan, F. (2012). Drought monitoring: Historical and current perspectives.

Heidke, P. (1926). Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler, 8*(4), 301-349.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the Proceedings of 3rd international conference on document analysis and recognition.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.

Keyantash, J., & Dracup, J. A. (2002). The quantification of drought: an evaluation of drought indices. *Bulletin of the American Meteorological Society, 83*(8), 1167-1180.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift, 15*(3), 259-263.

Lorenz, D. J., Otkin, J. A., Svoboda, M., Hain, C. R., Anderson, M. C., & Zhong, Y. (2017a). Predicting the US Drought Monitor using precipitation, soil moisture, and evapotranspiration anomalies. Part II: Intraseasonal drought intensification forecasts. *Journal of Hydrometeorology, 18*(7), 1963-1982.

Lorenz, D. J., Otkin, J. A., Svoboda, M., Hain, C. R., Anderson, M. C., & Zhong, Y. (2017b). Predicting US Drought Monitor states using precipitation, soil moisture, and evapotranspiration anomalies. Part I: Development of a nondiscrete USDM index. *Journal of Hydrometeorology, 18*(7), 1943-1962.

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). *The relationship of drought frequency and duration to time scales.* Paper presented at the Proceedings of the 8th Conference on Applied Climatology.

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., . . . Luo, L. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research: Atmospheres, 109*(D7).

Otkin, J. A., Anderson, M. C., Hain, C., Svoboda, M., Johnson, D., Mueller, R., . . . Brown, J. (2016). Assessing the evolution of soil moisture and vegetation conditions during the 2012 United States flash drought. *Agricultural and Forest Meteorology, 218*, 230-242.

Palmer, W. (1965). Meteorological drought. US Weather Bureau Research Paper 45, 58 pp. *Office of Climatology, US Department of Commerce, Washington DC*.

Quiring, S. M. (2009). Developing objective operational definitions for monitoring drought. *Journal of Applied Meteorology and Climatology, 48*(6), 1217-1229.

Sheffield, J., Goteti, G., Wen, F., & Wood, E. F. (2004). A simulated soil moisture based drought analysis for the United States. *Journal of Geophysical Research: Atmospheres, 109*(D24).

Shukla, S., & Wood, A. W. (2008). Use of a standardized runoff index for characterizing hydrologic drought. *Geophysical research letters, 35*(2).

Svoboda, M. (2000). An introduction to the drought monitor.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., . . . Stooksbury, D. (2002). The drought monitor. *Bulletin of the American Meteorological Society, 83*(8), 1181-1190.

USDA. (2018). The U.S. Drought Monitor: A Resource for Farmers, Ranchers and Foresters. Retrieved from https://www.usda.gov/media/blog/2018/04/19/us-drought-monitor-resource-farmers-ranchers-and-foresters

Wilhite, D. A. (2000). Drought as a natural hazard: concepts and definitions.

Wilhite, D. A., Svoboda, M. D., & Hayes, M. J. (2007). Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water resources management, 21*(5), 763-774.

-

Chapter 4 DOWNSCALING SMAP SATELLITE RETRIEVED SOIL MOISTURE USING MACHINE LEARNING APPROACHES WITH AN UNCERTAINTY PERSPECTIVE

4.1 Introduction

Soil moisture (SM) is a crucial variable within the Earth's system, and plays a key role in regulating various processes in water, energy, and carbon fluxes among the land surface and the atmosphere (Ochsner et al., 2013; Robock et al., 2000; Seneviratne et al., 2010). As a result, soil moisture becomes important for various geoscience models, such as hydrology, meteorology, and Earth thermodynamics (Vereecken et al., 2008). Soil moisture is defined and referred to as the quantity of water contained by the upper soil sector, also known as the unsaturated zone (Hillel, 1998).

The advancement of remote sensing technologies has increased the accessibility of soil moisture, to the point that it is possible to obtain an exceptional volume of remotely measured soil moisture spatially and temporally, a task that is not feasible from ground observation networks (Kerr, 2007). Several remote sensing satellite systems for worldwide soil moisture measurements are METOP-A/B, Advanced Scatterometer (ASCAT), the National Aeronautics and Space Administration's (NASA) Soil Moisture Active Passive (SMAP), the Advanced Microwave Scanning Radiometer for Earth Observing System (AMSR-E), the Advance Microwave Scanning Radiometer 2 (AMSR2), and the European Space Agency (ESA) Soil Moisture and Ocean Salinity (SMOS) (Entekhabi, Njoku, et al., 2010; Kerr et al., 2016; Qu et al., 2021). Each entity delivers significant global soil moisture retrievals at 25–50 km spatial resolution every 2–3 days (Qu et al., 2021; Senyurek et al., 2020).

NASA's SMAP was launched on January 31, 2015 as an environmental monitoring satellite, and offers soil moisture on a global level. It is equipped with an L-band (active) radar and an L-band (passive) radiometer (Entekhabi, Njoku, et al., 2010). The active and passive instruments obtain

soil moisture measurements with a spatial resolution of 3 and 36 km, respectively (Chan et al., 2016). In July 2015, the active sensor failed to operate correctly, and since then, SMAP soil moisture has been retrieved solely by the passive instrument. Figure 4-1 shows the global soil moisture obtained by the radar-based instrument for an 8-day cycle in June 2015.

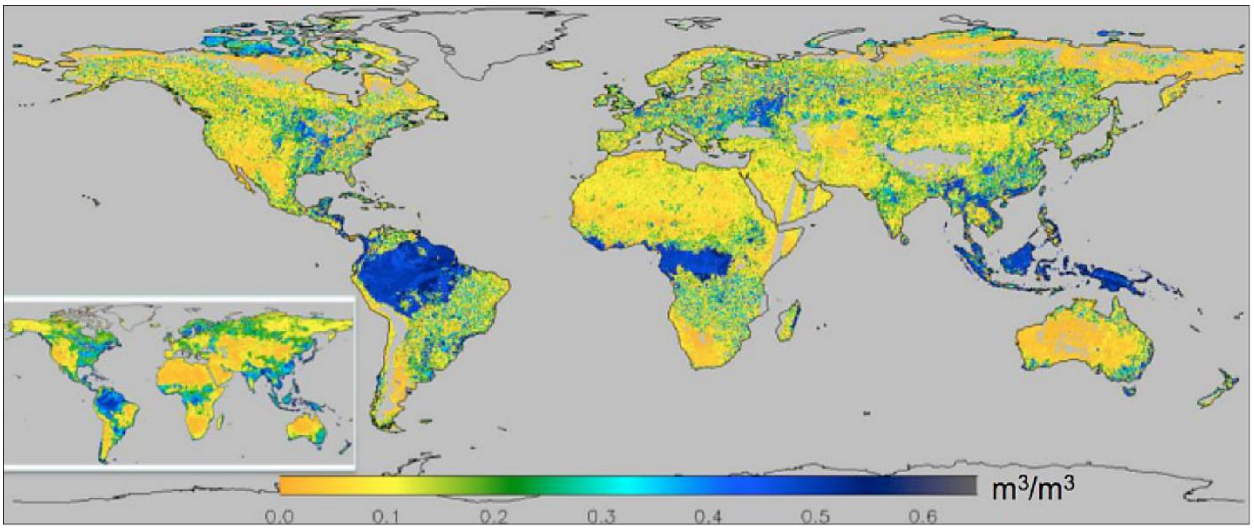


Figure 4-1 SMAP radar-based soil moisture for one 8-day cycle of June 19 to 26, 2015 (Retrieved from NASA (2015))

The SMAP soil moisture observation provides a suitable spatial resolution for global scale uses. However, the measurements cannot be utilized effectively for regional or local applications, such as agricultural purposes (e.g., yield estimation), drought, and flood monitoring. To avoid this problem, there is a need to obtain soil moisture at finer resolutions, from a multiple kilometer to less than one kilometer spatial resolution. As a result, a spatial downscaling is necessary for hydrological and agricultural applications (Peng et al., 2017). Soil moisture downscaling can be employed using different tactics, including satellite-based techniques (Active and Passive Microwave Data Fusion Methods), downscaling utilizing geoinformation data, and model-based approaches, which can be divided into statistical or land surface model downscaling (Peng et al., 2017). Numerous model-based soil moisture downscaling techniques have been proposed with

varying considerations of the effect of several environmental variables. The supporting theory of these techniques is to create either a statistical association or a physics-based model among satellite soil moisture retrieval and fine-scale ancillary variables (e.g., soil type, temperature, soil depth, topography) (Peng et al., 2017). Machine learning (ML) approaches in soil moisture downscaling fall within the category of statistical models where the model illustrates the spatial statistics of the soil moisture area to connect the spatial variability to the spatial average, or to disclose in what manner statistics vary throughout scales.

In recent years, downscaling large-scale satellite soil moisture with ML techniques has attracted significant interest due to their exceptional accuracy and stability in all aspects when compared to the other downscaling techniques (Kim et al., 2018; Qu et al., 2021). Techniques including Support Vector Machines (Jin et al., 2020; Kim et al., 2018), neural networks (Alemohammad et al., 2018), Random Forest (RF) (Abbaszadeh et al., 2019; Qu et al., 2021) are among the most popular in downscaling soil moisture measurement. Their results showed that Random Forest, which is an ensemble decision tree algorithm, seems to perform better in downscaling remotely sensed soil moisture when compared to other ML techniques (Abbaszadeh et al., 2019; Im et al., 2016; Jing et al., 2016; Pelletier et al., 2016; Qu et al., 2021; Teluguntla et al., 2018).

Soil moisture spatial variability is controlled by a multitude of land-atmosphere components, such as precipitation, temperature, soil type, vegetation and topography, and the combined effects of these variables' consequences in high soil moisture spatial heterogeneity. As a result, a soil moisture downscaling method that can take complex and nonlinear relationships into account is necessary to achieve accurate and fine spatiotemporal soil moisture.

Abbaszadeh et al. (2019) successfully downscaled SMAP soil moisture by using RF over CONUS from April 2015 to December 2015. This study attempts to replicate their study using similar variables, but using three different machine learning algorithms and different subsets of data to quantify the uncertainty of the process. The machine learning algorithms, RF, XGBoost, and a deep learning algorithm are employed to downscale the SMAP soil moisture passive (radiometer) measurements from 36 to 1-km resolution over the Contiguous United States (CONUS).

4.2 Dataset

In this study, SMAP soil moisture data, the ancillary data consisting of NDVI (to capture the effect of vegetation dynamics on soil moisture), land surface temperature and precipitation (i.e., atmospheric variables to catch the temporal dynamics), topography and soil texture (i.e., geophysical variables to maintain spatial variability), and ground truth data (i.e., in-situ soil moisture measurements) were obtained for CONUS over the course of 45 months, from April 2015 to December 2018 to be used in the proposed downscaling framework. The ancillary data are anticipated to enhance satisfactory explanatory power on the soil moisture profile on various scales. The data and sources are explained in the following subsections:

4.2.1 SMAP Radiometer Soil Moisture

SMAP satellite measures daily global soil moisture at a depth of 5 centimeters at AM (descending) and PM (ascending) overpasses (Entekhabi et al., 2008). In this study, using the proposed framework, the level 3 descending SMAP measured soil moisture from the passive sensor (radiometer) with 36 km resolution is downscaled throughout the CONUS (USGS, 2020).

4.2.2 Ancillary Data

4.2.2.1 Vegetation

Vegetation is an essential component of soil moisture variability that has a profound influence on runoff. Additionally, vegetation is strongly linked to soil, water, and atmosphere, so variation in vegetation can be a good indicator of soil moisture content dynamics (Engstrom et al., 2008). Because of its significance, vegetation has frequently been used as a supplementary variable in satellite soil moisture downscaling (Fang & Lakshmi, 2014; Peng et al., 2015). The normalized difference vegetation index (NDVI) is a practical indicator for quantifying vegetation coverage, which can assess vegetation dynamics (Zhang et al., 2018). For the time span of the study, NDVI was obtained from MODIS Terra at 1 km resolution generated every 16 days (MOD13Q1).

4.2.2.2 Land Surface Temperature

Land surface temperature (LST) is a key climate variable that regulates and substantially impacts soil moisture (Pablos et al., 2016). As a result, land surface temperature has been widely used as a predictive variable in satellite soil moisture measurements (Fang et al., 2018; Zhao et al., 2018). MODIS Terra produces daily LST (MOD11A1) at 1 km resolution with local equatorial crossing time of approximately 10:30 a.m. in descending node (Wan, 2006).

4.2.2.3 Precipitation

Soil moisture dynamics across space and time are markedly dependent on precipitation variation. Its correlation with soil moisture has been studied on different geographical scales (Hohenegger et al., 2009; Wei & Dirmeyer, 2012). As a result, precipitation can play a significant role as an ancillary variable in downscaling SMAP soil moisture. The precipitation data for this study was obtained from NASA's Daymet Version 3 model output data. The Daymet dataset provides daily surface weather data, such as minimum temperature, maximum temperature, vapor pressure and

precipitation at 1 km resolution in North America and Hawaii (DAAC, 2020; Thornton et al., 2014).

4.2.2.4 Topography

Studies have identified associations between soil moisture and topography as surface variables, particularly throughout wet cycles when precipitation is occurs more frequently than evaporation (Nyberg, 1996; Tromp-van Meerveld & McDonnell, 2006). Elevation has been proven to be a crucial element used in topography to improve downscaling satellite soil moisture (Colliander et al., 2017; Im et al., 2016). Therefore, elevation is selected to be another ancillary variable in this study. The elevation data source was obtained from GTOPO30, a global digital elevation model (DEM) with an approximately 1 km resolution provided by the USGS Earth Resources Observation and Science (EROS) archive (EROS, 2020).

4.2.2.5 Soil Texture

Soil texture (or type) refers to what the proportion of a soil mass is composed of regarding the quantity of small (clay), medium (silt), and large (sand) particles. By gaining an understanding of the soil texture and its physical properties, we can learn more about its relationship to soil moisture content (e.g., infiltration rate and permeability). Soil texture information has been exploited utilized as an effective source of information for improved downscaling satellite soil moisture measurements (Abbaszadeh et al., 2019; Kim & Barros, 2002; Montzka et al., 2018). In this study, the top 5 cm of soil type data were collected from Soil Datasets at Pennsylvania State University available at 1 km spatial resolution (PSU, 2020).

4.2.3 Ground soil moisture observation

To validate the results of the proposed downscaling framework, there is a need for an in-situ soil moisture observation known as ground-truth. The U.S. Climate Reference Network (USCRN) and

Soil Climate Analysis Network (SCAN) are two systematic and persistent networks of climate monitoring with stations throughout CONUS, Alaska and Hawaii. Their sites utilize excellent sensors to measure variables such as temperature, precipitation, wind speed, and soil conditions (Coopersmith et al., 2016; Schaefer & Paetzold, 2001). Both networks offer soil moisture measurements at different depths (i.e., 5, 10, 20, 50 and 100 cm) and time scales. To conform with SMAP measurement however, the daily soil moisture data was obtained at a depth of 5 cm. The number of stations during the selected 45 months of the study were equal to 191 SCAN and 132 USCRN sites that are shown in Figure 4-2. The Figure suggests that SCAN and USCRN stations are well distributed throughout CONUS to involve various climates and soil textures.

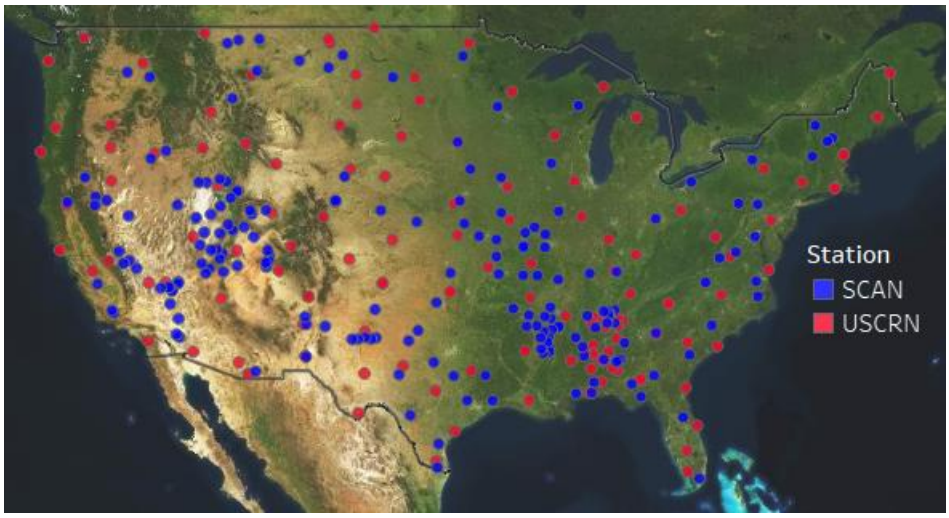


Figure 4-2 SCAN and USCRN stations networks across CONUS

4.3 Methodology

4.3.1 Data Arrangement and Modeling Schemes

To downscale SMAP radiometer soil moisture from a 36 km resolution to 1 km resolution, the fine resolution (1 km) auxiliary data including NDVI, precipitation, LST, elevation and soil type, which are known to be significantly correlated in capturing soil moisture, spatial and temporal dynamics are used. It should be noted that the input features are obtained for the location coordinates of the in-situ stations. The in-situ soil moisture measurements from USCRN and SCAN networks are considered as the predictand which the SMAP retrievals are validated against. The main assumption of downscaling is that the measured in-situ soil moisture is the representative value for the whole 1 km grid cell where the station is located.

The proposed framework incorporates two modeling schemes: local and global models. In local modeling, the stations are categorized based on their soil texture properties, where SMAP, as well as the rest of the ancillary variables (except soil type), are used as predictors in the model. Once all the data from each station for each soil type are combined, a local model for each soil type is developed to predict the in-situ soil moisture measurements. In global modeling, unlike the local model, soil texture is similarly managed to be employed as another predictor in addition to the rest of the input feature. Figure 4-3 shows all 15 soil textures and their covered area percentage across CONUS. Approximately 72 percent of the CONUS surface layer is covered with loam, silty loam, and sandy loam soil textures. Out of the remainder, water and bedrock are not considered soil layers, and for silt, no measurement station is available. Thus, there are 12 soil textures encompassing 98.74% of CONUS to be input into the downscaling framework.

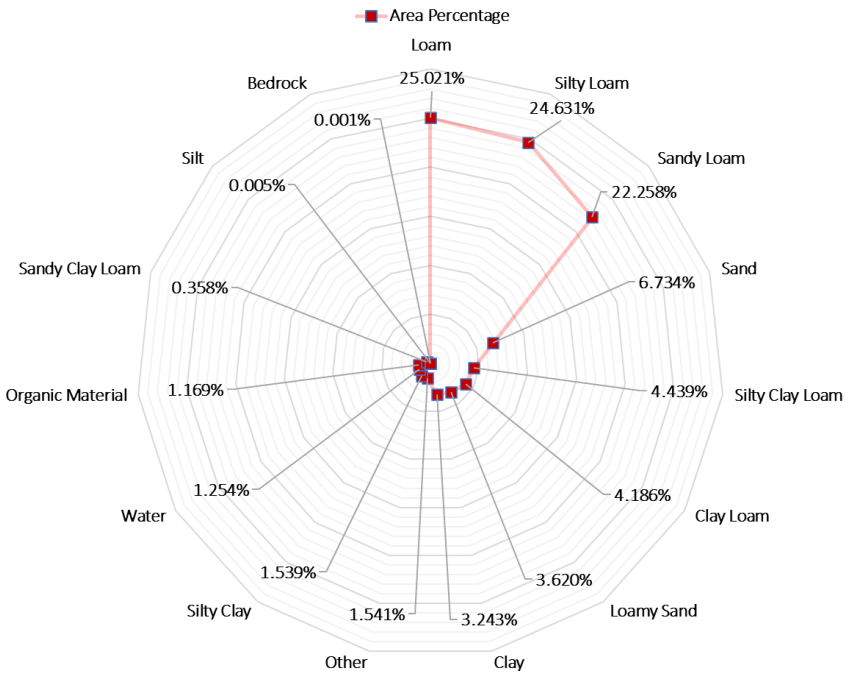


Figure 4-3 Soil textures and the covered area percentage across CONUS

The structure of the framework is illustrated in Figure 4-4. As can be seen, the data from the time span of the study is divided into three different temporal arrangements: cumulative, yearly, and quarterly, where each consists of four different data subsets. Cumulative data begins with data from the year 2015, and in succeeding years, data are added incrementally to previous combinations to create four different subsets: 2015, 2015 to 2016, 2015 to 2017, and 2015 to 2018. The motivation for employing this data arrangement is to recognize the impact of the data size on the accuracy of the prediction against the computational expenses. Yearly data isolates each year's data into four distinct years: 2015, 2016, 2017 and 2018. With the use of this yearly data, with a nearly identical number of data points, any present inconsistency and irregularity in data sources can be identified. Quarterly data contains each season's data with the specific purpose of gaining insight into the effect of seasonality and repeating patterns during soil moisture downscaling.

4.3.2 Machine Learning Algorithms, Model Selection and Metrics of Performance

In this study, three machine learning regression algorithms, RF, XGBoost and an artificial neural network (ANN), are utilized to implement the modeling of the framework. RF is a robust ensemble decision tree with a bagging algorithm. Bagging in RF strengthens the model to reduce variance and avoids overfitting by creating various models by resampling the data (Breiman, 2001). RF has been a highly successful machine learning algorithm in satellite soil moisture downscaling. XGBoost is termed as Extreme Gradient Boosting Algorithm which is also an ensemble technique that operates with boosting trees (Chen et al., 2015). XGBoost utilizes a gradient descent algorithm to remedy the preceding error created by the model by learning from it to improve next step performance. The previous results are rectified, and performance is enhanced. RF generates many trees, all with equivalent weight for leaves within the model, whereas XGBoost initiates leaf weighting to correct the ones that do not enhance the model predictability. XGBoost has gained popularity among data scientists, especially in machine learning competitions due to its speed and scalability.

The ANN algorithm in this study is chosen to be a Fully Connected Neural Network (FCNN). FCNNs are a form of ANN where the architecture is comprised of a sequence of fully connected layers, such that all their nodes (i.e., neurons) in one layer are connected to all the neurons in the following layer. The network architecture requires an input layer, one or multiple hidden layers and an output layer. The key advantage of FCNNs is that there are no specific assumptions required concerning the input. The ability of ANNs in learning complex nonlinear relationships between inputs and objective data is the reason for their popularity in geoscience. Over the past decade, ANNs have constantly been regularly applied to downscale soil moisture retrieval (Aires et al., 2017; Alemohammad et al., 2018; Jimenez et al., 2009).

Model hyperparameters tuning, evaluation and selection are attained by 10-fold cross-validation on 80% of training data. The best estimator is thereafter tested with 20% of the data. The performance metrics used to test the best trained models are $\boldsymbol{R^2}$ and unbiased Root Mean Square Error (ubRMSE). $\boldsymbol{R^2}$ is a measure of goodness of fit and the explanatory power of the model to the dependent variable. ubRMSE defined by Entekhabi, Reichle, et al. (2010), is a metric that SMAP employs to determine the measurement accuracy. As opposed to RMSE, ubRMSE is not harshly affected in presence of biases in the mean of the magnitude of variations in the retrievals. The ubRMSE is assumed to indicate the RMSE of soil moisture anomalies that are calculated by eliminating the mean seasonal cycle.

While RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \qquad (1)$$

Where N is the number of data points, $\boldsymbol{x_i}$ is the actual observation and $\boldsymbol{\hat{x}_i}$ is the estimated value.

ubRMSE is defined as below:

$$ubRMSE = \sqrt{\frac{\sum_{i=1}^{N}(\left(\hat{x}_i - \sum_{i=1}^{N}\frac{\hat{x}_i}{N}\right) - \left(x_i - \sum_{i=1}^{N}\frac{x_i}{N}\right))^2}{N}} \qquad (2)$$

and the relationship between RMSE and ubRMSE is:

$$RMSE^2 = ubRMSE^2 + b^2 \qquad (3)$$

Where b is the mean-bias.

Ultimately, with 11 data subsets and 12 soil types, there are 132 local models to be developed by each algorithm, resulting in a total of 396 local models. In contrast, given that global models

receive all data as input, and not for each soil type (soil type is used as a predictive feature), the number of global models is equal to 33, where the three aforementioned algorithms are supplied by 12 different data subsets. Overall, the framework develops 429 different local and global models. The major contributions of this research are to propose: 1) a soil moisture downscaling framework using machine learning, 2) a comparison between local and global modeling, 3) an assessment of three machine learning algorithm in soil moisture downscaling and 4) the uncertainty associated with the proposed models.

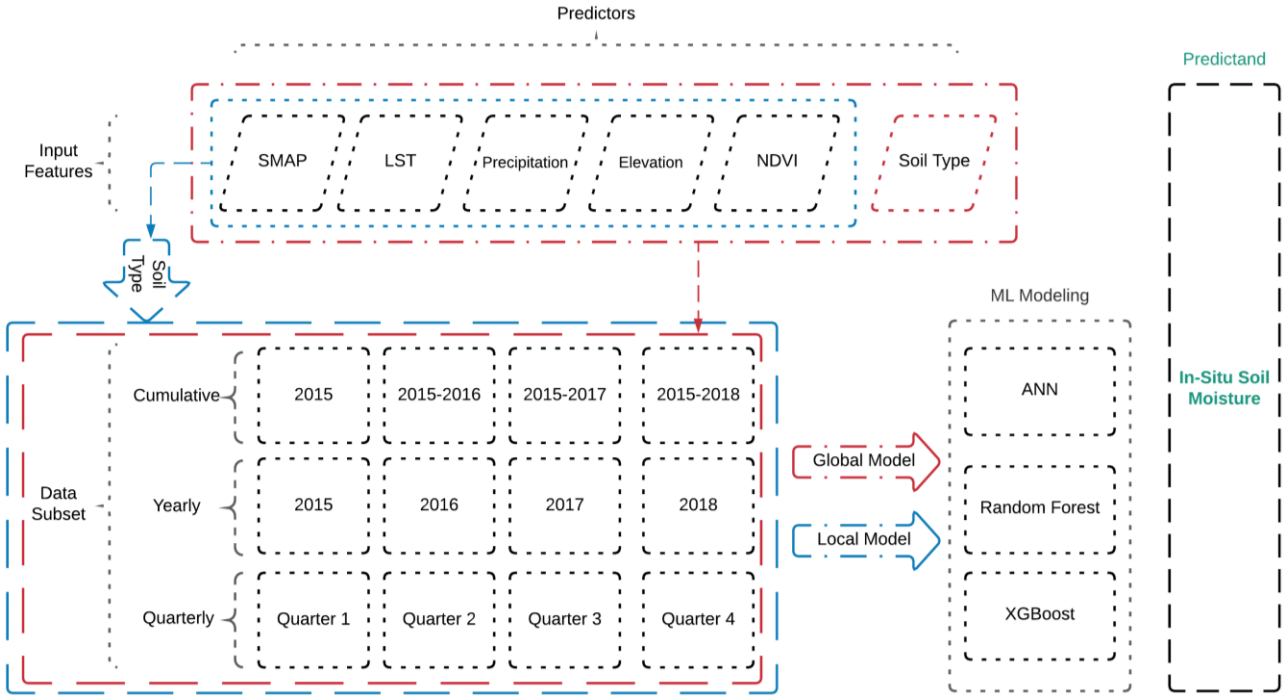


Figure 4-4 Flowchart of the proposed soil moisture downscaling framework

4.4 Results and Discussion

In this section, the results are discussed in three parts. An overview of the preprocessed data is presented in the first section. The following section includes the performance of the local models, and in the third section, the global model's results are presented. The highest performing model results are discussed in the fourth section, and in the final section, a deeper look into the overall downscaling results with an uncertainty perspective are presented.

4.4.1 Data Preprocessing

In this section, a summary of the preprocessed data before modeling is explained. In the first step, missing and outlier values were handled. Some in-situ stations did not have any reported value for the time of the study. Additionally, SMAP values for some stations during the entire four years of the data remained the same throughout the entire four years of data collection. After removing all those stations, out of 323 primary SCAN and USCRN stations, 296 stations remained to be used in the analysis. Table 4-1 shows the number of stations from both networks located on each soil texture, the percentage of the number of stations on each soil texture, as well as and the percentage area over CONUS covered with that soil texture. As can be seen in Table 4-1, the percentage of the number of stations from both networks together on each soil texture, and the soil texture covered area percentage, are very close. Organic material and sandy clay loam have only one in-situ station, and loam with 62, silty loam with 71 and sandy loam with 77 have the greatest number of stations (~ 71% of the stations). This insight contributes to the understanding that the data allocated to each soil texture has a fair spatial distribution when compared to the proportion they cover the area of across CONUS. Once the data of the stations for each soil texture are combined, the data subsets Quarterly, Yearly and Cumulative are created for the modeling task. Table 4-2 contains the number of data points in each data subset for every soil texture. Quarter 1 has the

fewest data points compared to the other three quarters. This is because data collection began in April 2015, and as a result, 2015 includes only nine months of data, while 2016, 2017, and 2018 have a full year of data. For organic material and sandy clay loam, the number of data points are small, and this is because only one station for each soil texture measuring soil moisture exists.

Table 4-1 Number of in-situ soil moisture stations on each soil texture

Soil Texture	Number of Stations	Percentage of Station	Area in CONUS %
Clay	8	2.70	3.243
Clay Loam	14	4.73	4.186
Loam	62	20.95	25.021
Loamy Sand	14	4.73	3.620
Other	3	1.01	1.541
Organic Material	1	0.34	1.169
Sand	19	6.42	6.734
Sandy Clay Loam	1	0.34	0.358
Silty Clay	3	1.01	1.539
Silty Clay Loam	23	7.77	4.439
Silty Loam	71	23.99	24.631
Sandy Loam	77	26.01	22.258

Table 4-2 - Number of data points in in each subset of data for different soil types

Soil Type	Clay	Clay Loam	Loam	Loamy Sand	Other	Organic Material	Sand	Sandy Clay Loam	Silty Clay	Silty Clay Loam	Silty Loam	Sandy Loam
Quarter 1	206	402	1630	465	62	4	556	40	73	671	1475	1774
Quarter 2	354	719	4012	852	159	33	1054	30	178	1170	3143	3214
Quarter 3	386	952	4226	821	167	37	1086	27	206	1269	3734	3336
Quarter 4	355	670	3386	797	134	21	967	34	156	1185	3037	3301
2015	317	562	2709	583	123	16	778	23	127	863	2327	2246
2016	389	756	3554	864	177	35	1041	21	173	1246	3185	3257
2017	351	706	3594	802	141	19	928	50	164	1104	3064	3122
2018	244	719	3397	686	81	25	916	37	149	1082	2813	3000
2015-16	706	1318	6263	1447	300	51	1819	44	300	2109	5512	5503
2015-17	1057	2024	9857	2249	441	70	2747	94	464	3213	8576	8625
2015-18	1301	2743	13254	2935	522	95	3663	131	613	4295	11389	11625

4.4.2 Local Models

In this section, the results of the local models are compared. Because there were 396 local models, only the results of the best-performing models are reported. The preprocessed data of each data subset for each soil texture were input into three algorithms, RF, XGBoost and FCNN. Among those three algorithms, FCNN did not perform well enough compared to RF and XGBoost. ANN models are data demanding algorithms, and consequently, they usually do not result in consistent findings in cases of insufficient data size. ANN algorithms do require large datasets to find hidden relationships between variables in a complex system. Because of this, only the results of the best-performing models (RF and XGBoost) are reported. Tables 4-3 and 4-4 present the $\boldsymbol{R^2}$ and ubRMSE of RF and XGBoost local models, respectively. As previously mentioned, hyperparameters for the models were executed through a 10-fold cross-validation of 80 percent of the data (i.e., training) and the remaining 20 percent were used for testing the trained model. It should be noted that to have a fair comparison, the training and the testing data points of each data subset were the same in both the RF and XGBoost, as well as the global model. The local models for organic material and sandy clay loam were not able to settle on a stable result. An insufficient number of data points to train and test the data were the reason for the inconsistency in the results. Insufficient data increases bias and in case of underfitting, variance decreases and results in higher inconsistency of the model prediction for a given data point that the model has not seen before. The elevation variable also contributed to the weak performance of the local models for organic material and sandy clay loam local models. Elevation, unlike the other used variables in this study, is spatially and temporally static. As a result, the effect of elevation in the models becomes apparent when there is more than one station available in the model. As this was not the case for

organic material and sandy clay loam, the models were lacking one input feature compared to the other soil textures.

Table 4-3 Performance of the Random Forest models in downscaling the SMAP soil moisture for different soil types across CONUS; ubRMSE is in m3/m3; NC = Not Consistent

Soil Type	Clay		Clay Loam		Loam		Loamy Sand		Other		Organic Material		Sand		Sandy Clay Loam		Silty Clay		Silty Clay Loam		Silty Loam		Sandy Loam	
Metric	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**
Quarter 1	0.97	0.036	0.81	0.048	0.75	0.047	0.82	0.033	0.93	0.032	NC	NC	0.70	0.017	NC	NC	0.92	0.058	0.87	0.057	0.85	0.059	0.86	0.043
Quarter 2	0.89	0.058	0.76	0.059	0.79	0.036	0.87	0.023	0.79	0.049	NC	NC	0.75	0.022	NC	NC	0.90	0.056	0.89	0.051	0.86	0.051	0.84	0.040
Quarter 3	0.83	0.051	0.68	0.062	0.74	0.030	0.80	0.021	0.77	0.045	NC	NC	0.70	0.020	NC	NC	0.86	0.047	0.88	0.050	0.86	0.046	0.80	0.044
Quarter 4	0.89	0.059	0.78	0.060	0.83	0.029	0.85	0.025	0.91	0.034	NC	NC	0.76	0.016	NC	NC	0.76	0.076	0.91	0.044	0.90	0.039	0.82	0.040
2015	0.87	0.065	0.84	0.047	0.83	0.030	0.79	0.025	0.91	0.035	NC	NC	0.81	0.014	NC	NC	0.91	0.058	0.93	0.042	0.84	0.052	0.81	0.045

Table 4-3 (cont'd)

2016	0.88	0.059	0.87	0.042	0.76	0.036	0.85	0.027	0.87	0.039	NC	NC	0.71	0.019	NC	NC	0.81	0.077	0.88	0.049	0.90	0.042	0.84	0.039
2017	0.87	0.064	0.84	0.044	0.81	0.031	0.84	0.027	0.94	0.024	NC	NC	0.66	0.021	NC	NC	0.96	0.036	0.91	0.048	0.87	0.051	0.89	0.034
2018	0.92	0.041	0.79	0.046	0.80	0.031	0.83	0.026	0.88	0.030	NC	NC	0.70	0.020	NC	NC	0.84	0.069	0.86	0.061	0.86	0.050	0.84	0.041
2015 -2016	0.88	0.058	0.84	0.045	0.81	0.032	0.79	0.032	0.91	0.036	NC	NC	0.86	0.016	NC	NC	0.90	0.051	0.92	0.044	0.90	0.042	0.86	0.039
2015 -2017	0.92	0.053	0.84	0.046	0.80	0.033	0.75	0.035	0.87	0.042	NC	NC	0.73	0.020	NC	NC	0.88	0.062	0.88	0.053	0.87	0.047	0.87	0.038
2015 -2018	0.88	0.057	0.76	0.056	0.80	0.033	0.79	0.029	0.88	0.037	NC	NC	0.78	0.017	NC	NC	0.88	0.056	0.89	0.051	0.88	0.048	0.86	0.038

Table 4-4- Performance of the XGBoost models in downscaling the SMAP soil moisture for different soil types across CONUS; ubRMSE is in m3/m3; NC = Not Consistent

Soil Type	Clay		Clay Loam		Loam		Loamy Sand		Other		Organic Material		Sand		Sandy Clay Loam		Silty Clay		Silty Clay Loam		Silty Loam		Sandy Loam	
Metric	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**	**R^2**	**ubRMSE**
Quarter 1	0.93	0.053	0.75	0.056	0.81	0.041	0.83	0.031	0.92	0.032	NC	NC	0.56	0.020	NC	NC	0.91	0.057	0.89	0.057	0.87	0.054	0.88	0.039
Quarter 2	0.86	0.062	0.75	0.065	0.80	0.035	0.85	0.023	0.81	0.053	NC	NC	0.75	0.022	NC	NC	0.86	0.057	0.88	0.057	0.85	0.053	0.84	0.041
Quarter 3	0.86	0.056	0.66	0.064	0.76	0.029	0.79	0.020	0.78	0.047	NC	NC	0.67	0.021	NC	NC	0.88	0.053	0.87	0.053	0.85	0.047	0.80	0.043
Quarter 4	0.88	0.063	0.76	0.066	0.84	0.028	0.85	0.026	0.87	0.038	NC	NC	0.75	0.016	NC	NC	0.83	0.038	0.93	0.038	0.90	0.040	0.85	0.037
2015	0.87	0.069	0.84	0.048	0.84	0.029	0.78	0.025	0.89	0.038	NC	NC	0.79	0.015	NC	NC	0.88	0.047	0.91	0.047	0.88	0.045	0.84	0.041

Table 4-4 (cont'd)

2016	0.88	0.062	0.87	0.044	0.81	0.032	0.83	0.029	0.88	0.037	NC	NC	0.71	0.019	NC	NC	0.78	0.053	0.86	0.053	0.89	0.044	0.86	0.038
2017	0.85	0.067	0.86	0.044	0.84	0.028	0.81	0.030	0.93	0.025	NC	NC	0.69	0.020	NC	NC	0.95	0.044	0.92	0.044	0.89	0.047	0.89	0.034
2018	0.92	0.059	0.80	0.048	0.80	0.031	0.80	0.027	0.92	0.023	NC	NC	0.68	0.023	NC	NC	0.82	0.057	0.88	0.057	0.87	0.048	0.87	0.038
2015 -2016	0.87	0.063	0.83	0.047	0.82	0.031	0.79	0.033	0.93	0.032	NC	NC	0.84	0.017	NC	NC	0.89	0.047	0.90	0.047	0.90	0.042	0.87	0.037
2015 -2017	0.89	0.056	0.83	0.047	0.81	0.033	0.73	0.036	0.87	0.037	NC	NC	0.72	0.020	NC	NC	0.86	0.056	0.86	0.056	0.88	0.046	0.87	0.037
2015 -2018	0.88	0.059	0.74	0.059	0.80	0.033	0.77	0.031	0.89	0.034	NC	NC	0.76	0.018	NC	NC	0.88	0.053	0.88	0.053	0.88	0.047	0.86	0.037

The results of the local models for each soil texture using the data subsets shows that RF and XGBoost performed similarly. Table 4-5 shows the results of the best performance obtained by RF and XGBoost local models for each soil texture, indicating that both RF and XGBoost performed in remarkably similar ways to choose the best data subset, with the exception of sand and silty clay loam.

Table 4-5 The data subsets resulting in the best prediction accuracy of RF and XGBoost models for each soil texture; ubRMSE is in m3/m3; NA = Not Available; NC = Not Consistent

	Random Forest			XGBoost		
Soil Texture	**Best Data Subset**	**R^2**	**ubRMSE**	**Best Data Subset**	**R^2**	**ubRMSE**
Clay	Quarter 1	0.97	0.036	Quarter 1	0.93	0.053
Clay Loam	2016	0.87	0.042	2016	0.87	0.044
Loam	Quarter 4	0.83	0.029	Quarter 4 and 2017	0.84	0.028
Loamy Sand	Quarter 2	0.87	0.023	Quarter 2	0.85	0.023
Other	2017	0.94	0.024	2017	0.93	0.025
Organic Material	NA	NC	NC	NA	NC	NC
Sand	2015-2016	0.86	0.016	2015	0.79	0.015
Sandy Clay Loam	NA	NC	NC	NA	NC	NC
Silty Clay	2017	0.96	0.036	2017	0.95	0.044
Silty Clay Loam	2015	0.93	0.042	Quarter 4	0.93	0.038
Silty Loam	Quarter 4	0.90	0.039	Quarter 4	0.90	0.040
Sandy Loam	2017	0.89	0.034	2017	0.89	0.034

4.4.3 Global Model

In this section, the results of the global model are explained. Out of 33 developed global models, FCNN with the data from 2015 to 2018 performed the best. The model hyperparameters were tuned using a 10-fold cross-validation on 80 percent of the data and finally tested with the remaining 20 percent. The FCNN architecture achieved the best results when five hidden layers were employed with the overall 88% accuracy for training and 85% testing. The results of the global model are presented in Table 4-6. The global model performance for each soil texture was generally lower than the local models. However, the model exhibited a consistent result for organic material and sandy clay loam when the models were validated with different splits of data.

However, the accuracies were not yet close enough to the other soil textures. Given that the local models were not able to offer consistent generalizability, a global model highlights its advantage in gaining a relative consistent accuracy by learning from similar information in other soil textures. In the next section, a more comprehensive evaluation of the local and global models through ensemble averaging will be discussed.

Table 4-6 Performance of the FCNN model as the best performing global model with 2015 - 2018 data

Soil Texture	R^2	ubRMSE
Clay	0.88	0.055
Clay Loam	0.75	0.060
Loam	0.78	0.034
Loamy Sand	0.73	0.034
Other	0.92	0.038
Organic Material	0.60	0.060
Sand	0.75	0.020
Sandy Clay Loam	0.30	0.048
Silty Clay	0.89	0.055
Silty Clay Loam	0.78	0.076
Silty Loam	0.87	0.050
Sandy Loam	0.82	0.045

4.4.4 Ensemble Averaging

Ensemble averaging is a method used to learn from multiple models where the contribution of every member is equal to the final result. In this study, because the accuracy of the local models for each soil texture ranges within each data subset, the final accuracy of predictions for each algorithm (RF and XGBoost) regarding the soil textures can be calculated by the model averaging ensemble. Table 4-7 shows the results of this averaging. As indicated in the Table, the results are almost equal, with a slight outperformance for RF in six soil textures, as compared to XGBoost, which achieved better results in four. Aside from organic material and sandy clay loam, the global model outperformed the local models in Other, Sand and Silty Clay.

Table 4-7 Ensemble averaging results of RF and XGBoost local models versus FCNN global model

Soil Texture	Random Forest		XGBoost		FCNN	
	Average R^2	Average ubRMSE	Average R^2	Average ubRMSE	Average R^2	Average ubRMSE
Clay	0.89	0.055	0.88	0.059	0.88	0.055
Clay Loam	0.80	0.050	0.79	0.053	0.75	0.060
Loam	0.79	0.034	0.81	0.032	0.78	0.034
Loamy Sand	0.82	0.028	0.80	0.028	0.73	0.034
Other	0.88	0.037	0.88	0.036	0.92	0.038
Organic Material	NC	NC	NC	NC	0.60	0.006
Sand	0.74	0.018	0.72	0.019	0.75	0.020
Sandy Clay Loam	NC	NC	NC	NC	0.30	0.048
Silty Clay	0.87	0.059	0.87	0.062	0.89	0.055
Silty Clay Loam	0.89	0.050	0.89	0.051	0.78	0.076
Silty Loam	0.87	0.048	0.88	0.047	0.87	0.050
Sandy Loam	0.84	0.040	0.86	0.038	0.82	0.045

4.4.5 Local Models Ranking

From Table 4-5, it can be inferred that the 2017 data subset resulted in the highest accuracy for both RF and XGBoost in multiple soil textures, but inconsistently. By ranking the local models for each soil texture, it is noticeable that the local models in conjunction with any specific data arrangement could not indicate a consistent dominance over the rest (Tables 4-8 and 4-9). The median ranking of the between different data arrangements suggests the local models within the 2017 data subset could generate the overall highest accuracy. To obtain better comprehension of the models' performances in downscaling the soil moisture, the RF model with the data from the year 2017 was selected as an example with the best scoring data arrangement. Figure 4-5 presents the ranking of the RF for all the soil texture, excluding organic material and sandy clay loam. The RF models with 2017 data subsets could contain the best in three soil textures: Other, Silty Clay. and Sandy Loam. However, in Sand and Clay, the performances were among the lowest.

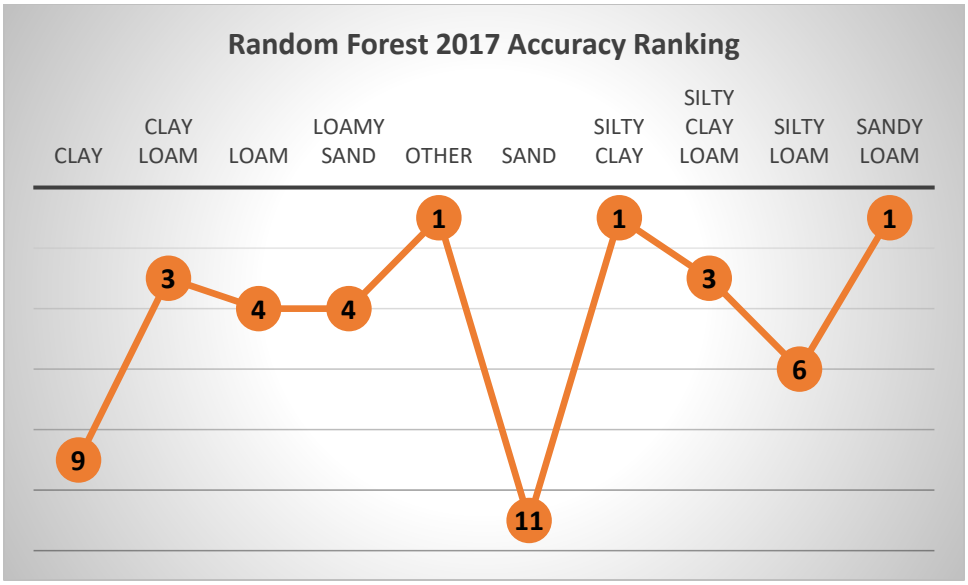


Figure 4-5 Random Forest 2017 models soil moisture prediction accuracy rankings for each soil type when compared to the rest of the models with different data subsets

The scatterplots of the RF model for each soil texture, excluding organic material and sandy clay loam are presented in Figure 4-6. The plots show the in-situ values on the X-axis, and downscaled soil moisture values on the first Y-axis, and SMAP values on the second Y-axis. The models were properly capable of capturing the trend, while the SMAP 36 km values are scattered without showing any clear correlation with the in-situ values.

Table 4-8 Ranking of the Random Forest models for each soil type and the median of the overall used data subset

	Clay	Clay Loam	Loam	Loamy Sand	Other	Organic Material	Sand	Sandy Clay Loam	Silty Clay	Silty Clay Loam	Silty Loam	Sandy Loam	Median Rank
Quarter 1	1	6	10	6	2	-	9	-	2	10	10	5	6
Quarter 2	4	9	8	1	10	-	5	-	5	6	9	8	7
Quarter 3	11	11	11	7	11	-	10	-	8	7	8	11	10.5
Quarter 4	5	8	1	2	5	-	4	-	11	4	1	9	4.5
2015	10	2	2	8	4	-	2	-	3	1	11	10	3.5
2016	6	1	9	3	8	-	7	-	10	8	2	6	6.5
2017	9	3	4	4	1	-	11	-	1	3	6	1	3.5
2018	3	7	5	5	6	-	8	-	9	11	7	7	7
2015 -16	7	4	3	9	3	-	1	-	4	2	3	3	3
2015 -17	2	5	6	11	9	-	6	-	6	9	5	2	6
2015 -18	8	10	7	10	7	-	3	-	7	5	4	4	7

Table 4-9 Ranking of the XGBoost models for each soil type and the median of the overall used data subset

	Clay	Clay Loam	Loam	Loamy Sand	Other	Organic Material	Sand	Sandy Clay Loam	Silty Clay	Silty Clay Loam	Silty Loam	Sandy Loam	Median Rank
Quarter 1	1	8	7	3	4	-	11	-	2	5	9	2	5
Quarter 2	9	9	10	1	10	-	5	-	8	8	11	9	9
Quarter 3	10	11	11	7	11	-	10	-	6	9	10	11	10
Quarter 4	6	7	2	2	8	-	4	-	9	1	1	8	4
2015	7	3	3	9	6	-	2	-	4	3	5	10	4
2016	5	1	5	4	7	-	7	-	11	11	3	7	5
2017	11	2	1	5	1	-	8	-	1	2	4	1	2
2018	2	6	9	6	3	-	9	-	10	7	8	5	7
2015 -16	8	4	4	8	2	-	1	-	3	4	2	4	4
2015 -17	3	5	6	11	9	-	6	-	7	10	7	3	7
2015 -18	4	10	8	10	5	-	3	-	5	6	6	6	6

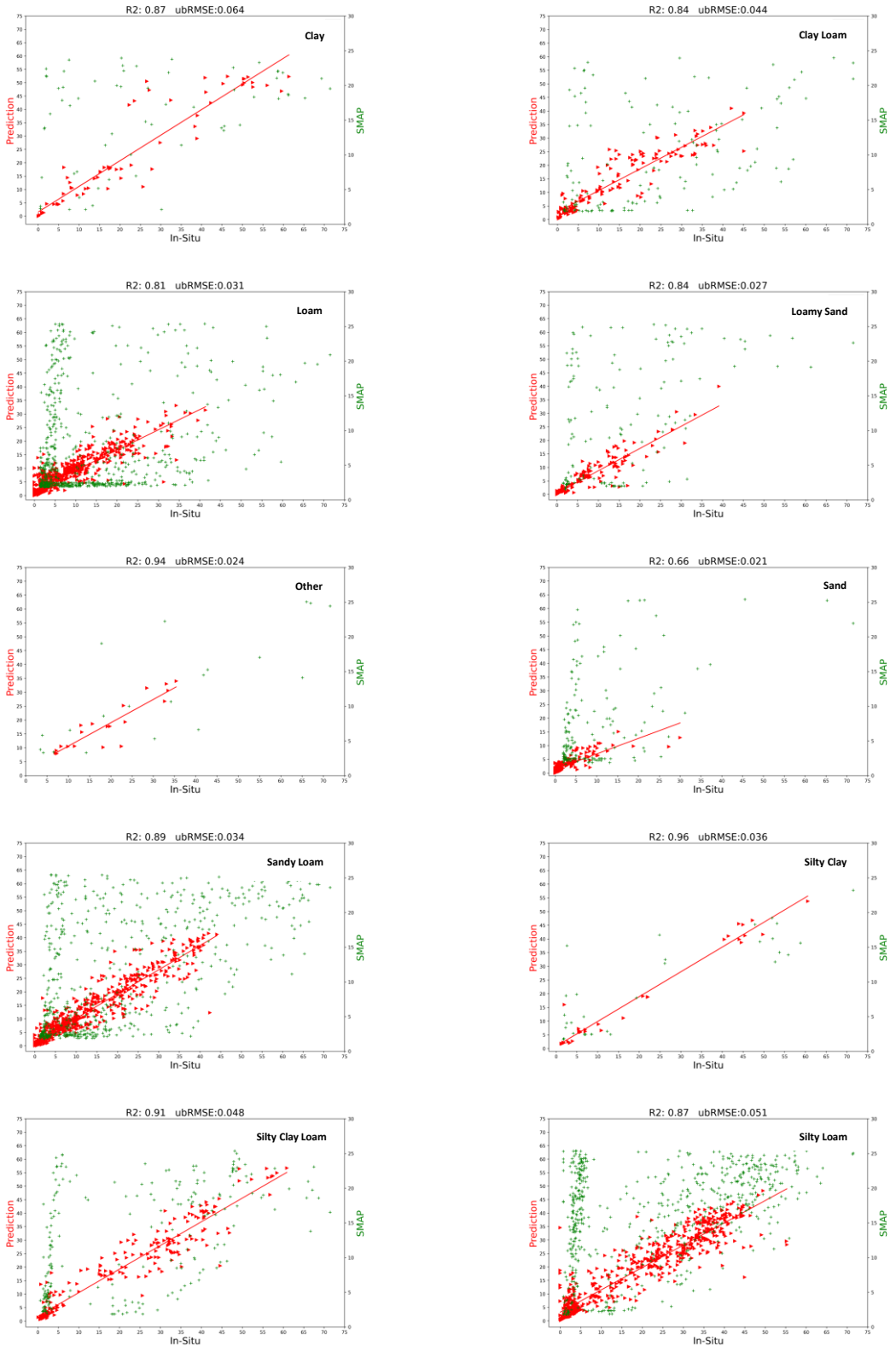


Figure 4-6 Scatterplot of the predicted soil moisture for each soil type by the Random Forest Models with the 2017 data

4.4.6 Uncertainty Analysis of Local Models

Although the accuracies of the local models were high, none of the data subsets suggested performed the best in all soil textures. The variability of the results in each model, necessitates an investigation of the underlying reason. Therefore, in this part, the uncertainty associated with the obtained results of the downscaling is discussed. It should be noted that since the RF and XGBoost performed almost identically, RF model results were selected in this part of the analysis. For this purpose, the ranges between the maximum and minimum accuracies of the 11 models for each soil type were calculated and were compared to the variabilities in spatial properties (NDVI, elevation) of each soil textures. The comparison was evaluated by a Pearson Correlation Coefficient test (Benesty et al., 2009). The results did not indicate any significant correlation between the range of accuracies and the ranges in NDVI and elevation. However, when the ranges of accuracies were tested against the data size properties, including the number of stations in each soil texture, the number of data points range in each soil texture, and in the covered area percentage of each soil texture, significant correlations were discovered. The correlation coefficient between the $\mathbf{R^2}$ range and the number of stations, number of data points range and area percentage across CONUS were equal to - 0.72, - 0.71, and - 0.68. Additionally, the correlations were significant at the 0.05 level with p-values equal to 0.017, 0.022 and 0.034. The results are included in Tables 4-10 and 4-11, and Figure 4-7.

Table 4-10 The accuracy range of the local models and the data size properties of each soil texture

Soil Texture	R^2 Range	Number of Stations	Number of Data Points Range	Area in CONUS
Clay	0.14	8	1095	3.243%
Clay Loam	0.19	14	2341	4.186%
Loam	0.09	62	11624	25.021%
Loamy Sand	0.12	14	2470	3.620%
Other	0.17	3	460	1.541%
Sand	0.20	19	3107	6.734%
Silty Clay	0.20	3	540	1.539%
Silty Clay Loam	0.07	23	3624	4.439%
Silty Loam	0.06	71	9914	24.631%
Sandy Loam	0.09	77	9851	22.258%

Table 4-11 Pearson correlation coefficient between R2 ranges and data size properties with significance at the 0.05 level

		R^2 Range	Number of Stations	Range in Number of Data Points	Area Coverage
R^2 Range	Pearson Correlation	1	- 0.72	- 0.71	- 0.68
	Significance (2-tailed)		0.017	0.022	0.034

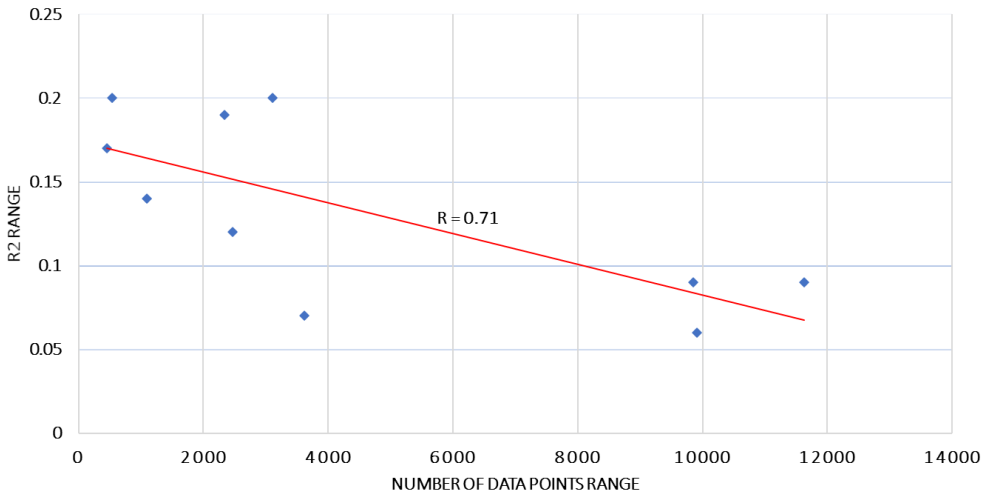


Figure 4-7 Relationship between the models R2 ranges and the range in the number of data points for each soil texture

The significant correlation coefficients imply that as the number of data points increases, the variability of the model explanatory power consequently tends to decrease. The outcome of the

test advises that in a downscaling task, a particular set of data in a model may result in high accuracies, but the results may vary significantly if the task is implemented using another time span of data. Using this finding, it was possible to define an uncertainty interval for the downscaling performance. Table 4-12 illustrates the uncertainty intervals for each soil texture. The accuracies are attained by the ensemble averaging of the local models created for each data subset in each soil texture.

Table 4-12 Downscaling accuracy uncertainty interval for each soil texture across CONUS

Soil Texture	Downscaling Accuracy Uncertainty Interval
Clay	89 ± 7%
Clay Loam	80 ± 9.5%
Loam	79 ± 4.5%
Loamy Sand	82 ± 6%
Other	88 ± 8.5%
Sand	74 ± 10%
Silty Clay	87 ± 10%
Silty Clay Loam	89 ± 3.5%
Silty Loam	87 ± 3%
Sandy Loam	84 ± 4.5%

A visual illustration of uncertainty intervals is shown in Figure 4-8 regarding the area percentage of each soil texture. It is noteworthy that the higher the percentage area, the less uncertainty was associated with the downscaling procedure. This signified that the achieved downscaling performances for most of the CONUS area was coupled with more confidence. In general, the models that received more input data were more skilled in generalizing the hidden relationships among them with less variance. For example, the uncertainty for loam, silty loam and sandy clay, which constitute 72% of the CONUS, varied between 3% to 4.5%, whereas for the remaining 25% area of CONUS the uncertainty differed in a range of 6% to 20%.

Figure 4-8 Downscaled soil moisture ensemble averaged accuracy band and the covered area percentage in each soil texture

4.5 Conclusions

In this study, SMAP soil moisture 36 km was downscaled to 1 km spatial resolution. SMAP soil moisture data, the ancillary data consisting of NDVI (to capture the effect of vegetation dynamics on soil moisture), land surface temperature and precipitation (i.e., atmospheric variables to catch the temporal dynamics), topography and soil texture (i.e., geophysical variables to maintain spatial variability), and ground truth data (i.e., in-situ soil moisture measurements) were obtained for CONUS over the course of 45 months, from April 2015 to December 2018 to be used in the propose downscaling framework. The proposed framework incorporated two modeling schemes: local and global models. In local modeling, the stations were categorized based on their soil texture properties, where SMAP, as well as the rest of the ancillary variables (except soil type), are used as predictors in the model. In global modeling, unlike the local model, soil texture was similarly managed to be employed as another predictor in addition to the rest of the input feature. Three machine learning regression algorithms, RF, XGBoost and fully connected neural networks were

utilized to implement the modeling of the framework. The time span of the study was divided into three different temporal arrangements: cumulative, yearly, and quarterly, where each consisted of four different data subsets. A total of 396 local models, and 33 global models, the framework developed 468 different models. The results suggested that RF and XGBoost local models performed almost equally, but significantly better than FCNN. Conversely, FCNN outperformed RF and XGBoost in global modeling. The advantage of the global scheme over local scheme was its capacity to offer a consistent result for two soil textures: organic material and sandy clay loam, even though the accuracies were not close enough to other soil textures. The proposed framework also managed to offer high downscaling accuracies by using an ensemble averaging of the local models. With the help of an uncertainty analysis, the results suggested that the accuracy of the models significantly depended on the temporality of the selected data. By ensemble averaging the results of the local models for each soil texture, and the range of the variability between the minimum and maximum accuracy, the proposed framework was able to offer a consistent result with an uncertainty interval. Another finding of this study was the significant correlation between the uncertainty intervals and the data size, where the soil texture with more in-situ stations had a lower degree of uncertainty. Future works could add more in-situ stations from different measurement networks, and potentially include more topographical data such as landforms to increase the spatial features of the grid cells.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abbaszadeh, P., Moradkhani, H., & Zhan, X. (2019). Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. *Water Resources Research*, *55*(1), 324-344.

Aires, F., Miolane, L., Prigent, C., Pham, B., Fluet-Chouinard, E., Lehner, B., & Papa, F. (2017). A global dynamic long-term inundation extent dataset at high spatial resolution derived through downscaling of satellite observations. *Journal of Hydrometeorology*, *18*(5), 1305-1325.

Alemohammad, S. H., Kolassa, J., Prigent, C., Aires, F., & Gentine, P. (2018). Global downscaling of remotely sensed soil moisture using neural networks. *Hydrology and Earth System Sciences*, *22*(10), 5341-5356.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chan, S. K., Bindlish, R., O'Neill, P. E., Njoku, E., Jackson, T., Colliander, A., Chen, F., Burgin, M., Dunbar, S., & Piepmeier, J. (2016). Assessment of the SMAP passive soil moisture product. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 4994-5007.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4).

Colliander, A., Fisher, J. B., Halverson, G., Merlin, O., Misra, S., Bindlish, R., Jackson, T. J., & Yueh, S. (2017). Spatial downscaling of SMAP soil moisture using MODIS land surface temperature and NDVI during SMAPVEX15. *IEEE Geoscience and Remote Sensing Letters*, *14*(11), 2107-2111.

Coopersmith, E., Cosh, M., Bell, J. E., & Crow, W. (2016). Multi-profile analysis of soil moisture within the US Climate Reference Network. *Vadose Zone Journal*, *15*(1), 1-8.

DAAC. (2020). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3*. https://daac.ornl.gov/DAYMET/guides/Daymet_V3_CFMosaics.html

Engstrom, R., Hope, A., Kwon, H., & Stow, D. (2008). The relationship between soil moisture and NDVI near Barrow, Alaska. *Physical Geography*, *29*(1), 38-53.

Entekhabi, D., Jackson, T., Njoku, E., O'neill, P., & Entin, J. (2008). Soil moisture active/passive (SMAP) mission concept. Atmospheric and Environmental Remote Sensing Data Processing and Utilization IV: Readiness for GEOSS II,

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., & Johnson, J. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, *98*(5), 704-716.

Entekhabi, D., Reichle, R. H., Koster, R. D., & Crow, W. T. (2010). Performance metrics for soil moisture retrievals and application requirements. *Journal of Hydrometeorology*, *11*(3), 832-840.

EROS, U. (2020). *Digital Elevation - Global 30 Arc-Second Elevation (GTOPO30)*. https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30?qt-science_center_objects=0#qt-science_center_objects

Fang, B., & Lakshmi, V. (2014). Soil moisture at watershed scale: Remote sensing techniques. *Journal of Hydrology*, *516*, 258-272.

Fang, B., Lakshmi, V., Bindlish, R., & Jackson, T. J. (2018). Downscaling of SMAP soil moisture using land surface temperature and vegetation data. *Vadose Zone Journal*, *17*(1), 1-15.

Hillel, D. (1998). *Environmental soil physics: Fundamentals, applications, and environmental considerations*. Elsevier.

Hohenegger, C., Brockhaus, P., Bretherton, C. S., & Schär, C. (2009). The soil moisture–precipitation feedback in simulations with explicit and parameterized convection. *Journal of Climate*, *22*(19), 5003-5020.

Im, J., Park, S., Rhee, J., Baik, J., & Choi, M. (2016). Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches. *Environmental Earth Sciences*, *75*(15), 1-19.

Jimenez, C., Prigent, C., & Aires, F. (2009). Toward an estimation of global land surface heat fluxes from multisatellite observations. *Journal of Geophysical Research: Atmospheres*, *114*(D6).

Jin, Y., Ge, Y., Liu, Y., Chen, Y., Zhang, H., & Heuvelink, G. B. (2020). A machine learning-based geostatistical downscaling method for coarse-resolution soil moisture products. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Jing, W., Yang, Y., Yue, X., & Zhao, X. (2016). A spatial downscaling algorithm for satellite-based precipitation over the Tibetan plateau based on NDVI, DEM, and land surface temperature. *Remote Sensing*, *8*(8), 655.

Kerr, Y. H. (2007). Soil moisture from space: Where are we? *Hydrogeology journal*, *15*(1), 117-120.

Kerr, Y. H., Al-Yaari, A., Rodriguez-Fernandez, N., Parrens, M., Molero, B., Leroux, D., Bircher, S., Mahmoodi, A., Mialon, A., & Richaume, P. (2016). Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation. *Remote Sensing of Environment*, *180*, 40-63.

Kim, D., Moon, H., Kim, H., Im, J., & Choi, M. (2018). Intercomparison of downscaling techniques for satellite soil moisture products. *Advances in Meteorology*, *2018*.

NASA. (2015). *https://directory.eoportal.org/web/eoportal/satellite-missions/content/-/article/smap*. 2015.

Nyberg, L. (1996). Spatial variability of soil water content in the covered catchment at Gårdsjön, Sweden. *Hydrological Processes*, *10*(1), 89-103.

Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. A., Draper, C. S., Hagimoto, Y., Kerr, Y. H., Larson, K. M., Njoku, E. G., & Small, E. E. (2013). State of the art in large-scale soil moisture monitoring. *Soil Science Society of America Journal*, *77*(6), 1888-1919.

Pablos, M., Martínez-Fernández, J., Piles, M., Sánchez, N., Vall-llossera, M., & Camps, A. (2016). Multi-temporal evaluation of soil moisture and land surface temperature dynamics using in situ and satellite observations. *Remote Sensing*, *8*(7), 587.

Pelletier, C., Valero, S., Inglada, J., Champion, N., & Dedieu, G. (2016). Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, *187*, 156-168.

Peng, J., Loew, A., Merlin, O., & Verhoest, N. E. (2017). A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics*, *55*(2), 341-366.

Peng, J., Niesel, J., & Loew, A. (2015). Evaluation of soil moisture downscaling using a simple thermal-based proxy–the REMEDHUS network (Spain) example. *Hydrology and Earth System Sciences*, *19*(12), 4765-4782.

Qu, Y., Zhu, Z., Montzka, C., Chai, L., Liu, S., Ge, Y., Liu, J., Lu, Z., He, X., & Zheng, J. (2021). Inter-comparison of several soil moisture downscaling methods over the Qinghai-Tibet Plateau, China. *Journal of Hydrology*, *592*, 125616.

Robock, A., Vinnikov, K. Y., Srinivasan, G., Entin, J. K., Hollinger, S. E., Speranskaya, N. A., Liu, S., & Namkhai, A. (2000). The global soil moisture data bank. *Bulletin of the American Meteorological Society*, *81*(6), 1281-1300.

Schaefer, G. L., & Paetzold, R. F. (2001). SNOTEL (SNOwpack TELemetry) and SCAN (soil climate analysis network). *Automated Weather Stations for Applications in Agriculture and Water Resources Management: Current Use and Future Perspectives*(1074), 187-194.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., & Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, *99*(3-4), 125-161.

Senyurek, V., Lei, F., Boyd, D., Kurum, M., Gurbuz, A. C., & Moorhead, R. (2020). Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS. *Remote Sensing*, *12*(7), 1168.

Teluguntla, P., Thenkabail, P. S., Oliphant, A., Xiong, J., Gumma, M. K., Congalton, R. G., Yadav, K., & Huete, A. (2018). A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform. *ISPRS Journal of Photogrammetry and Remote Sensing*, *144*, 325-340.

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. B. (2014). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2*.

Tromp-van Meerveld, H., & McDonnell, J. (2006). On the interrelations between topography, soil depth, soil moisture, transpiration rates and species distribution at the hillslope scale. *Advances in Water Resources*, *29*(2), 293-310.

USGS. (2020). *Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (MOD13Q1)*. https://lpdaac.usgs.gov/products/mod13q1v006/

Vereecken, H., Huisman, J., Bogena, H., Vanderborght, J., Vrugt, J., & Hopmans, J. (2008). On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research*, *44*(4).

Wan, Z. (2006). MODIS land surface temperature products users' guide. *ICESS, University of California*.

Wei, J., & Dirmeyer, P. A. (2012). Dissecting soil moisture-precipitation coupling. *Geophysical Research Letters*, *39*(19).

Zhang, H., Chang, J., Zhang, L., Wang, Y., Li, Y., & Wang, X. (2018). NDVI dynamic changes and their relationship with meteorological factors and soil moisture. *Environmental Earth Sciences*, *77*(16), 1-11.

Zhao, W., Sánchez, N., Lu, H., & Li, A. (2018). A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression. *Journal of Hydrology*, *563*, 1009-1024.

Chapter 5 CONCLUSIONS

In the modern realm of ubiquitous, large, frequent, and instant data capturing with the advancements in instrumentation, data generation and data gathering techniques, we are offered new prospects to comprehend and analyze the role of geography in everyday life. However, traditional geographic data analytics are now strictly challenged by the volume, velocity, variety and veracity of the data requiring analysis to extract value. Because of that, geographic data science has received a remarkable attention in the past two decades to tackle those challenges. However, considering that much of data science's success is formed outside of geography, there is an increased risk within such perspectives that location stays only as an additional column within a database, no more or less important than any other feature. Geographic data science combines the data with spatial and temporal components. The spatial and temporal dependence allow us to interpolate and extrapolate to fill gaps in the presence of inadequate data and infer reasonable approximations elsewhere by the incorporation of information of diverse kinds and sources. Although, within scientific communities, there exist arguments regarding whether geographic data science is a scientific discipline of its own. Since data science is still in its early adoption phases in geography and for the transformation from a practice to a discipline, geographic data science is required to develop its unique concepts, differentiating itself from other disciplines such as statistics or computer science. This becomes possible when geographers, within a community of practice, are enabled to first learn and connect the current tools, methods, and domain knowledge to address the existing challenges of geographic data analysis. To take a step toward that purpose, in this dissertation, knowledge-rich applications of data science in the analysis of geographic spatiotemporal big datasets inspired by the existing challenges were studied and examined. In the first chapter, the challenges and opportunities in the era of "big data" were reviewed and it was explained that data science has formed as an interdisciplinary method to transform large amounts

of data into information. However, despite being common in other fields of science, data science is still in its initial implementation phases in the geography discipline. Furthermore, the opportunity to bridge the gap between geography and data science, and to explore the opportunities and challenges facing machine learning encouraged this research.

This research tackled three different problems within geography; one within the subfield of human geography, and two within physical geography. In the second chapter, a fine resolution spatiotemporal crime prediction framework was proposed to evaluate the performance of multi-task learning methods against the commonly used single-task learning methods. Although, there existed many gaps and challenges due to the limited scope of the study, and the complexity of human dynamics prediction, several findings were discovered. In case of limited samples, MTL could perform better than the local modeling. Finer spatial and temporal resolutions significantly influenced the prediction results due to insufficient data causing sparsity in the dependent variable. On the other hand, by choosing larger spatiotemporal resolutions, the framework could not make the predictions practical for police preemptive actions. However, the purpose of this study was to establish a basis for future crime analytical studies by introducing MTL to the community for further research.

In the third chapter, a framework using machine learning and land surface model outputs was developed to reproduce the USDM weekly drought maps. The results showed that the proposed framework could reproduce the USDM maps to a near-perfect level. Although, in terms of prediction accuracy, there was not much room left for improvement by the baseline model, our proposed framework could outperform it by testing different scenarios of the data inputs and machine learning algorithms to find the best combination. It was found out that employing the past week drought data as a predictor in the models played an important role in achieving high

prediction scores especially in the logistic regression. Additionally, the drought classification task in this study was a nonlinear problem since Random Forest and SVM outperformed the logistic regression. One of the main challenges in this study was the lack of data points and imbalanced distribution of the extreme drought categories in the multi-class classification tasks across the domain of study which led to biased models to perform poorly for those categories. The issue was resolved by leveraging data from the neighboring grid cells to improve model performance for these categories that was essentially compensation of the temporal shortage of the available USDM data by increasing it spatially.

In the fourth chapter, a framework was proposed to downscale SMAP satellite soil moisture retrievals from 36 to 1 km spatial resolution. A group of ancillary data were utilized to improve the downscaling task. NDVI to capture the effect of vegetation dynamics on soil moisture, land surface temperature and precipitation as the atmospheric variables to catch the temporal dynamics, and topography and soil texture as the geophysical variables to maintain spatial variability, were the variables that have been proven to improve the process. Three different machine and different data subsets in the proposed framework, managed to offer high downscaling accuracies by using an ensemble averaging of the local models. With the help of an uncertainty analysis, the results suggested that the accuracy of the models significantly depended on the temporality of the selected data. By ensemble averaging the results of the local models for each soil texture, and the range of the variability between the minimum and maximum accuracy, the proposed framework was able to offer a consistent result with an uncertainty interval. One of the main challenges in this study was the insufficient in-situ validation data points, particularly in soil textures with very few ground stations.

The contributions made in this research offer machine learning methods to effectively and efficiently overcome the existing challenges facing traditional approaches to analyzing spatiotemporal data. Additionally, this research recognizes the challenges that naturally arise with spatiotemporal data analysis for machine learning methods and offers solutions along the way. Overall, with the use of domain knowledge, machine learning algorithms proved their ability to learn the essential behavior of a system from training datasets. Although, insufficient sample size and in-situ observations related to the selected spatial and temporal resolution were found to be yet the primary obstacle in this research. Other challenges were originating from multi-source and resolution data which limited the more detailed studies. Poor data quality (e.g., SMAP data in this research) was another challenge which undermined the overall data quality and size, which consequently affected the modeling tasks performance. In the future, using our understanding of the challenges from the data and the shortcomings of the existing machine learning methods in every specific topic, there will opportunities to outline geographic data science as a unique discipline with its own concepts and immediate solutions in the analysis of the data.