An Investigation of Fuzzy Modeling for Spatial Prediction with Sparsely Distributed Data

by

Robert Thomas
BSc, University of Victoria, 2015

A Master's Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Mechanical Engineering

© Robert Thomas, 2018
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.

Supervisory Committee

An Investigation of Fuzzy Modeling for Spatial Prediction with Sparsely Distributed Data

by

Robert Thomas
BSc, University of Victoria, 2015

Supervisory Committee

Dr. Caterina Valeo (Department of Mechanical Engineering)
Supervisor

Dr. Usman T. Khan (Department of Civil Engineering, Lassonde School of Engineering)
Co-Supervisor

Dr. Brad Buckham (Department of Mechanical Engineering)
Departmental Member

Abstract

Dr. Caterina Valeo (Department of Mechanical Engineering)
Supervisor
Dr. Usman T. Khan (Department of Civil Engineering, Lassonde School of Engineering)
Co-Supervisor
Dr. Brad Buckham (Department of Mechanical Engineering)
Departmental Member

Dioxins are highly toxic persistent environmental pollutants that occur in marine harbour sediments as the results of industrial practices around the world and pose a significant risk to human health. To adequately remediate contaminated sediments, the spatial extent of contamination must first be determined by spatial interpolation. The ability to lower sampling frequency and perform laboratory analysis on fewer samples, yet still produce an adequate pollutant distribution map, would reduce the initial cost of new remediation projects. Fuzzy Set Theory has been shown as a way to reduce uncertainty due to data sparsity and provides an advantageous way to quantify gradational changes like those of pollutant concentrations through fuzzing clustering based approaches; Fuzzy modelling has the ability to utilize these advantages for making spatial predictions. To assess the ability of fuzzy modeling to make spatial predictions using fewer sample points, its predictive ability was compared to Ordinary Kriging (OK) and Inverse Distance Weighting (IDW) under increasingly sparse data conditions. This research used a Takagi-Sugeno (T-S) fuzzy modelling approach with fuzzy c-means clustering to make spatial predictions of lead concentrations in soil to determine the efficacy of the fuzzy model for applications of modeling dioxins in marine sediment. The spatial density of the data used to make the predictions was incrementally reduced to simulate increasingly sparse spatial data conditions. To determine model performance, the data at each increment not used for making the spatial predictions was used as validation data, which the model attempted to predict

and the performance was analyzed. Initially, the parameters associated with the T-S fuzzy model were determined by the optimum observed performance, where the combination of parameters that produced the most accurate prediction of the validation data were retained as optimal for each increment of the data reduction. To determine performance Mean Absolute Error, the Coefficient of Determination, and Root Mean Squared Error were selected as metrics. To give each metric equal weighting a binned scoring system was developed where each metric received a score from 1 to 10, the average represented that methods score. The Akaike Information Criterion (AIC) was also employed to determine the effect of the varied validation set lengths on performance. For the T-S fuzzy model as the amount of data used to solve the respective validation set points was reduced the number of clusters was lower and the cluster centres were more spread out, the fuzzy overlap between clusters was larger, and the widths of the membership function in the T-S fuzzy model were wider. Although it was possible to determine an optimal number of clusters, fuzzy overlap, and membership function width that yielded an optimal prediction of the validation data, gain in performance was minor compared to many other combinations of parameters. Therefore, for the data used in this study the T-S fuzzy model was insensitive to parameter choice. For OK, as the data was reduced, the range of spatial dependence in the data from variography became lower, and for IDW the power parameters optimal value became lower to give a greater weighting to more widely spread points. For the T-S fuzzy model, OK, and IDW the increasingly sparse data conditions resulted in an increasingly poor model performance for all metrics. This was supported by AIC values for each method at each increment of the data reduction that were within 1 point of each other. The ability of the methods to predict outlier points and reproduce the variance in the validation sets was very similar and overall quite poor. Based on the scoring system IDW did exhibit a slight

outperformance of the T-S fuzzy model, which slightly outperformed OK. However, the scoring system employed in this research was overly sensitive and so was only useful for assessing relative performance. The performance of the T-S model was very dependent on the number of outliers in the respective validation set. For modeling under sparse data conditions, the T-S fuzzy modeling approach using FCM clustering and constant width Gaussian shaped membership functions used in this research did not show any advantages over IDW and OK for the type of data tested. Therefore, it was not possible to speculate on a possible reduction in sampling frequency for delineating the extent of contamination for new remediation projects.

Table of Contents

Acknowledgments

Firstly, I would like to thank my loving Wife for her never-ending support and encouragement, without her, I'm not sure where I would be. I would also like to extend a huge thank you to my Supervisor, Dr. Caterina Valeo, for giving me the opportunity to pursue an MASc under her guidance, thank-you for believing in me. Thanks to my Co-Supervisor, Dr. Usman T Khan, for sharing his breadth of Fuzzy Set Theory knowledge. Thank you to all my family and friends and to my Mom, Dad, and Sisters who have done so much for me. I would also like to acknowledge Rad Haghi, for always entertaining my crazy ideas and helping to solve my many coding problems. Finally, to my son Finnigan: Finn, if one day you read this, know that I finished it just after your second birthday, I love you.

List of Tables

List of Figures

1. Introduction

1.1 Background and motivation

The release of pollutants into the natural environment has been a problem of global concern since the beginning of the industrial revolution. Due to the nature of global commerce, port cities and harbours have been the focus of much industrial activity. As a result the bottom sediments in these environments have acted as sinks for persistent environmental pollutants; of major concern to human health are dioxins, furans, and dioxin-like polychlorinated biphenyls (PCBs) (hereby collectively referred to as dioxins) (Hites, 2011). The major contributor to exposure of humans to dioxins is through ingestion of contaminated food as a result of bio-accumulation in fish (Travis & Hattemer-Frey, 1991). Chronic exposure to dioxins can lead to infertility, birth defects, impaired child development, diabetes, damage to the immune system, disruption of hormone function, and cancer (Mitrou et al., 2001). Seafood in the most common exposure pathway of dioxins to humans, therefore for the protection of human health the remediation of these contaminants in aquatic environments is of the upmost importance (Kulkarni, Crespo, & Afonso, 2008). The first step required for remediation is an accurate assessment of the spatial distribution of contamination to ensure the most effective and cost effective remediation possible. Spatially continuous data are required to delineate the boundaries of unsafe levels of contamination and to determine the volume of contaminated material to be removed. However, in aquatic environments point samples are generally collected on predetermined grid spacing's and the contaminant concentrations are spatially interpolated to provide a continuous surface that must be as accurate as possible. Spatial interpolation methods are traditionally grouped into geostatistical and deterministic methods. The most commonly used deterministic method is Inverse Distance Weighting (IDW), which uses a deterministic function

that estimates values at un-sampled points by the linear combination of known sample values weighted by the distance between them (Li & Heap, 2008; Shepard, 1968). IDW is a simple method requiring minimal modeler input (Shahbeik, Afzal, Moarefvand, & Qumarsy, 2014); however, because it's based solely on distance, it often performs very poorly with sparsely distributed geospatial data (Li & Heap, 2011). The most commonly used geostatistical method for spatial interpolation is kriging (Li & Heap, 2011), which employs a semivariogram that plots the semivariance between points against the distance between them, to determine the variance between samples as a function of the distance between them (Matheron, 1963). From the semivariogram it is possible to determine the range of spatial dependence of sampled points and use them in solving a value at an unknown location (Burrough & McDonnell, 1998). However, accurate estimation of a semivariogram is complicated, computationally expensive, and can introduce modeler bias (Gedeon et al., 2003). Furthermore, kriging has been shown to have a significant smoothing affect, where areas of high pollutant concentration could be missed (Goovaerts, 1999). This is not ideal since underestimation of pollutant concentrations can lead to an increased risk to human health. Both geostatistical and deterministic methods have one major commonality, that is, the greater the sample density the greater the accuracy of the spatial interpolation (Stahl et al., 2006). However, the sampling cost of sediment in marine environments and the analytical assessment cost for dioxins are extremely high. Therefore, obtaining an adequate number of samples to achieve an acceptable resolution during interpolation may not be possible and this high cost may be prohibitive to new remediation projects (Chris Gill, Stantec, personal communication, Oct. 28th, 2016).

There is however a separate family of data-driven predictive methods that utilize Zadeh's (1965) fuzzy set theory, which have been proven as a suitable method for the prediction of sparse

non-linear data and have been used for many applications of spatial estimation in the geosciences (Kajornrit, Wong, & Fung, 2016; Muhammad & Glass, 2011; Collazo-Cuevas et al., 2010; Tutmez & Haptipoglu, 2010; Zhang, 2009; Bardossy & Fodor, 2004; Burrough, 1989;). Fuzzy set theory (Zadeh, 1965) provides a convenient way of describing the degree of belonging (membership μ) of an element to a set, between 0 (no belonging) and 1 (complete belonging). For example, let U be an ordinary set with elements $\{x_1, x_2, \ldots, x_n\}$ and $\tilde{A}$ be a fuzzy subset of U, in which the elements x_i have degrees of membership (belonging to $\tilde{A}$) given by a membership function $\mu_{\tilde{A}}(x_i) = \alpha$, which dictates that an element x_i has a degree of membership α to fuzzy set $\tilde{A}$, where $0 \leq \alpha \leq 1$ (Figure 1).

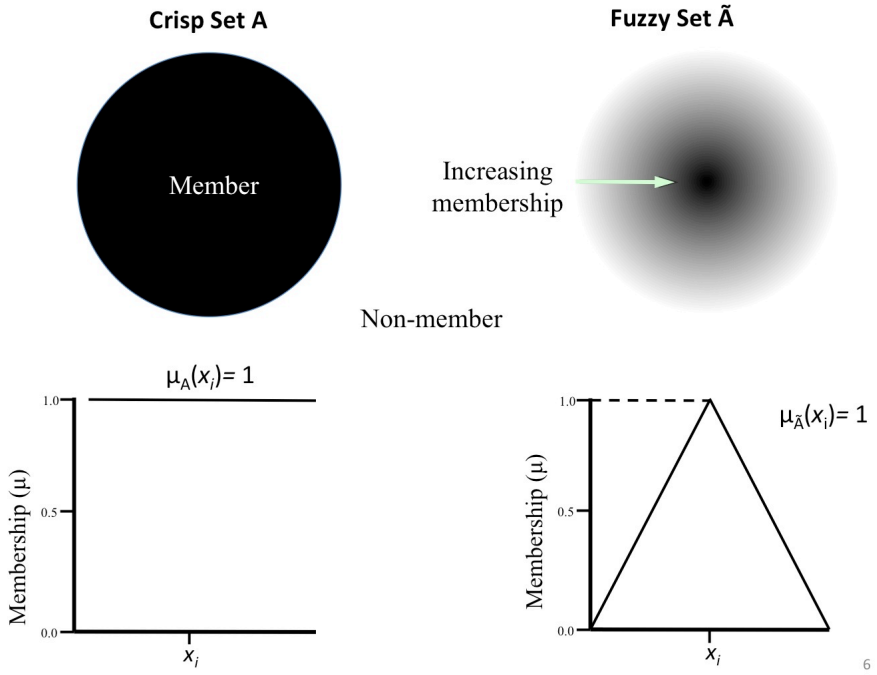


Figure 1-1, Example of a crisp and fuzzy set and their respective membership functions (modified from Sonmez, Gokceoglu, and Ulusay, 2004)

System modeling techniques that employ fuzzy set theory are commonly referred to as fuzzy modeling (Kajornrit, Wong, & Fung, 2016; Muhammad & Glass, 2011; Tutmez & Haptipoglu, 2010). The most commonly used fuzzy modeling technique for spatial estimation is the Takagi – Sugeno (T-S) method (Kajornrit et al., 2016, Muhammad & Glass, 2011; Kazemi & Hosseni, 2001; Tutmez & Hatipoglu, 2010; Tutmez, Tercan, & Kaymak, 2007). T-S fuzzy modeling breaks down the input data space into a number of fuzzy regions and creates a linear function for each region (Tagaki & Sugeno, 1985). It is advantageous for spatial modeling because of its transparency and interpretability (Pedrycz & Izakian, 2014). The degree of belonging that an un-sampled point has to each region in the data is used in predicting a data value at that location. The partitioning of the input space into fuzzy regions is achieved though fuzzy clustering, which is the foundation of fuzzy spatial modeling using the T-S method. Fuzzy clustering differs from traditional crisp data clustering in that in fuzzy methods each element (sample point) can have a degree of membership to multiple clusters within the data (Bezdek, 1981). Each cluster is defined by a cluster center, which has a value calculated from the membership-weighted average of the members of that cluster (Bezdek, 1981). For spatial modeling, data is clustered in the 3-dimensional product space defined by the Cartesian map coordinates (x, y) and the magnitude of a pollutant concentration (p) (Figure 1-2).

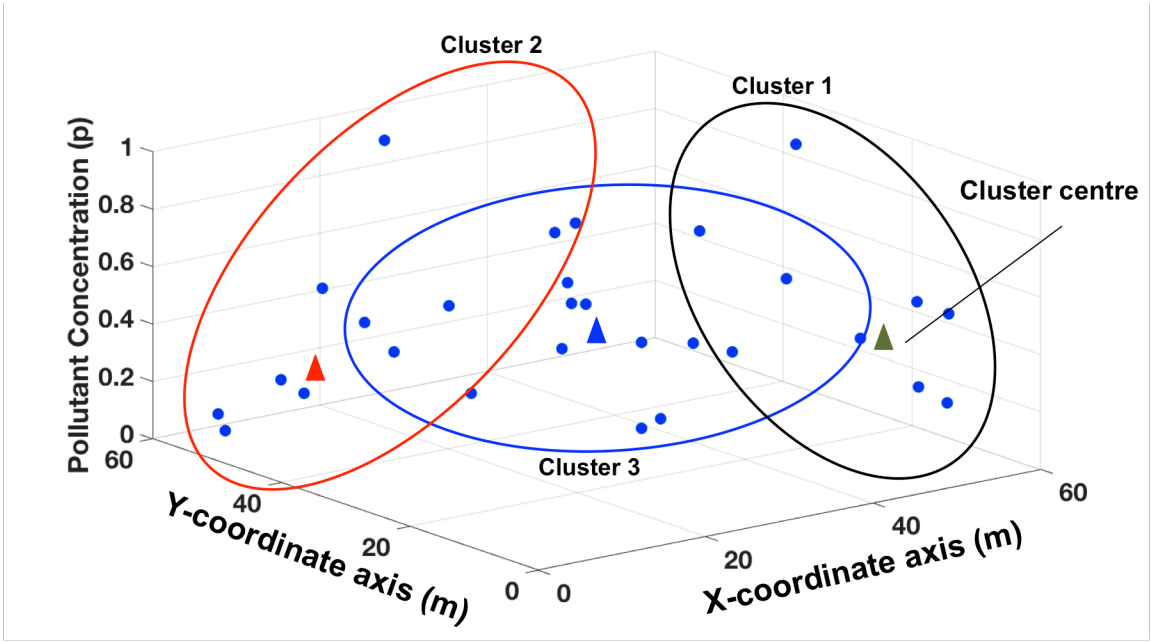


Figure 1-2, Theoretical visualization of fuzzy clustering in the 3-dimensional product space

After clustering, each data point has membership to one or more cluster centres that are in terms of pollutant concentration. The clusters are then partitioned onto the x and y Cartesian product space and a membership function generated for the x and y coordinate axis of each cluster. The membership functions from each cluster are then used to determine the degree of membership an un-sampled location has to the different clusters within the data (Figure 1-3).

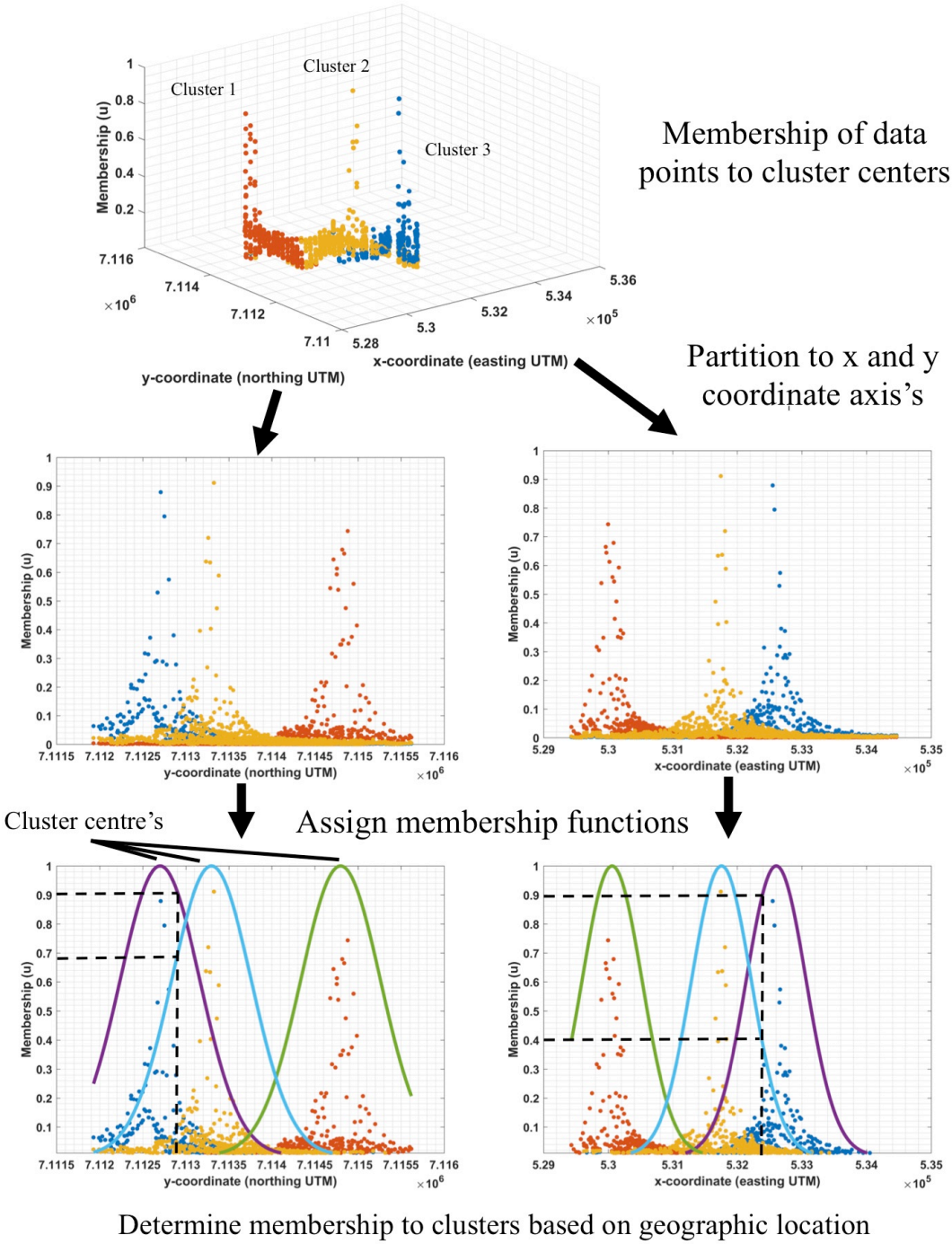


Figure 1-3, Example of partitioning membership to cluster centres to the map coordinate axis's and applying a membership function for 3 fuzzy clusters to determine membership based on geographic location

Ultimately, the combination of these degrees of membership is used in solving a pollutant concentration at an unknown location. Once the data has been clustered, it is subjected to a rule-based fuzzy inference system (FIS) that makes inferences about un-sampled geographic locations based on their membership to the clusters within the data. A single rule is introduced for each cluster using conditional IF-THEN statements makes on geographic position to determine each rules contribution. The FIS uses input variables referred to as antecedents for each rule, in the T-S fuzzy model, the antecedents are the fuzzy sets generated from the clustered data. The antecedents are subjected to IF-THEN statements to produces a weighting for the prediction from each rule based on membership to that rule (cluster). The output of each rule is referred to as a rule consequent. When a rule in the FIS is executed, if the antecedent is unaffected by the IF-THEN condition, that rule is skipped and the next rule is executed. If the IF-THEN condition produces a consequent then that rule is deemed to have "fired" or been executed. Each rule in the T-S fuzzy model uses *x* and *y* Cartesian coordinates as inputs to solve the consequent (output variable) for each rule in the form of a linear equation to produce a crisp value (Takagi & Sugeno, 1985). The form of the general first order TS model is

$$R_i = if\ x_1 is\ A_{i1}\ and\ x_2\ is\ A_{in}\ then\ p_i = a_{i1}x_1 + a_{i2}x_2 + b_i \qquad (1)$$

$$i = 1, \dots, K$$

where R_i is the *i*th rule, x_1 and x_2 are the antecedent variables (Cartesian coordinates *x,y*), A_{i1} and A_{i2} are fuzzy sets for the *i*th rule and respective coordinate axes, p_i is the *i*th rule output, K is the number of rules (clusters), and a_{i1}, a_{i2}, and b_i are unknown model parameters, which must be solved, this is accomplished by least squares regression using the data inherent in each cluster (Muhammad & Glass, 2011; Kazemi and Hosseini, 2011; Tutmez & Tercan, 2007; Tutmez and

Hatipoglu, 2010). The model output for a given input is obtained by aggregation of all utilized rule consequents weighted by their membership to the respective rules.

The T-S methods simplistic nature and interpretability make it advantageous compared to other FIS (Zhou & Gan, 2008). Applications using T-S fuzzy modeling for spatial estimation have shown its predictive capacity to outperform kriging (Kord and Moghaddam (2014); Tutmez & Hatipoglu, 2010; Tutmez et al., 2007). The advantage of T-S modeling is that it allows a complex data surface to be broken down into more easily modeled individual fuzzy surfaces by clustering, and using a rule base to estimate a value at unknown locations by solving the intersection of the contributing surfaces based on the degree of belonging an unknown location has to each surface (Zhou & Gan, 2008). Fuzzy set theory has been cited as a useful tool for modeling under sparse data conditions (Khan, 2015; Bárdossy and Fodor, 2001; McBratney & Odeh, 1997). Fuzzy clustering utilizes Fuzzy Set theory and allows the inclusion of more information by relating each data point to each theoretical cluster in the data by a certain degree of membership. Through the ability of fuzzy modeling to take advantage of this additional information it may be possible to produce an adequate contaminate distribution map for remediation planning using fewer point samples. Lowering the number of samples required would reduce the initial cost of new remediation projects and may lead to more remediation projects being undertaken in the future; ultimately leading to a cleaner environment and lower anthropogenic health risks.

1.2 Research Objectives

The objective of this research is to investigate the predictive ability of fuzzy modeling via the T-S method with fuzzy clustering for spatial estimation. The investigation will involve comparing T-S fuzzy modeling to the conventional spatial interpolation methods Kriging and

IDW. The focus of the investigation is on the effect of increasingly sparse data conditions on the accuracy of the aforementioned methods. This will help to determine if fuzzy modeling has the ability produce an equally accurate pollutant distribution map with a lower spatial sampling density. The practical application of this research is to attempt to provide a method by which fewer samples could be collected and analyzed and yet still provide a detailed enough pollutant distribution map to calculate masses of contaminated material to be disposed of and establish boundaries for areas to be remediated, all at a lower initial cost.

1.3 Overview

Chapter 1 has identified the difficulties with performing spatial interpolation under sparse spatial data conditions using traditional methods and proposed a possible solution using fuzzy modeling. Chapter 2 provides a summary of the relevant literature and quantifies the assertions made in Chapter 1. Chapter 3 begins with a description of the geospatial data used in this research, specifically, the type of data and its relevance to the problem statement. The remainder of Chapter 3 summarizes the methods used and how the relative predictive ability of the models used was compared under increasingly sparse spatial data conditions. Chapter 4 reviews the results of the study and discusses the implications of the results, including suggestions for future research. Chapter 5 provides a conclusion and reiterates the findings of this research.

2. Literature Review

2.1 Traditional Spatial Interpolation

Of the methods available for spatial interpolation, ordinary kriging (OK) and IDW are the most commonly used in the geosciences (Zarco-Perello & Simões, 2017). Many examples of comparisons between IDW and OK for spatial interpolation of different phenomena exist in the literature. Zarco-Perello & Simões (2017) compared OK to IDW for the spatial interpolation of coral reef parameters in the Gulf of Mexico. They found that for larger distances between samples both IDW and OK performed poorly, and that at shorter distances IDW showed the lowest prediction error and had a lesser smoothing effect than OK. Shahbeiket et al. (2014) compared IDW and OK for the interpolation of ore grades in an iron deposit. They found that predictions by OK had a lower over-all error and that OK was better for predicting regional highs within the deposit. Qiao et al. (2018) compared OK and IDW for the spatial interpolation of pollutants in soil, they found that IDW was better suited for their application because it required fewer data points and had a lesser smoothing effect than kriging. Mousavi et al. (2017) found that OK outperformed IDW for the spatial prediction of physical and chemical soil parameters and conclude that because these phenomena are highly spatially dependent they are more easily modeled by geostatistics. Mirzaei and Sakizadeh (2016) compared OK and IDW for the estimation of ground water contamination. They concluded that the results of the two methods are very similar, but that OK does have a significant smoothing effect. Li and Heap (2011) reviewed 51 comparative analyses of spatial interpolation methods. They found that OK and IDW are the most commonly compared and were able to make several observations about spatial sample density based on the studies observations. Most significantly, they observed that at higher spatial density of data points OK and IDW do not produce significantly different results and that

at very low sample densities OK performs very poorly, but overall OK does generally outperform IDW at higher spatial densities. Additionally, they observed that as range and variance of the input data increased the performance decreased for both methods. Liao, Li, and Zhang (2018) compared the effect of the spatial density and sample spatial distribution on the performance of OK and IDW for interpolation of heavy metals in soil. They determined that sampling methodology had a negligible impact on performance, that IDW had the greatest performance for large spatial scales, and that in cases where the total variance of the population was adequately sampled OK outperformed IDW. The literature is in general agreement that as sample density increases, as does performance regardless of the spatial interpolation method employed (Li & Heap, 2011). However, for determining performance under sparse spatial data conditions, a preferred method of spatial interpolation appears to be dependent on the data being modeled.

2.2 Fuzzy Modeling

Fuzzy modeling originates from modeling complex systems in control theory, its success in data driven modeling is more recent, but has shown great promise (Zhou & Ghan, 2008). In general two types of fuzzy inference system (FIS) frame works for spatial prediction exist in the literature; they are Mamdani (linguistic) and T-S based (data-driven). Mamdani based methods stem from the use of linguistic variable's where the inference systems inputs and outputs are both fuzzy sets (Mamdani and Assilian, 1973). Expert knowledge is used to create the input fuzzy sets in linguistic terms and the output fuzzy sets must be defuzzified to produce a crisp output (Mamdani & Assilian, 1973). The general form of the Mamdani FIS is:

$$R_i = if\ x_1 is\ A_{i1}\ and\ x_2\ is\ A_{in}\ then\ p_i\ is\ B_i \quad (2)$$

$$i = 1, \dots, K$$

where R_i is the *i*th rule, x_1 and x_2 are again the antecedent variables, A_{i1} and A_{i2} are fuzzy sets determined linguistically by expert knowledge, output p_i is represented by the fuzzy set B_i, and K is the number of rules. In the case of qualitative modelling of environmental variables the linguistic input fuzzy sets might describe levels of concentration, such as high and low, or contaminated and uncontaminated (Tutmez & Tercan, 2007). The major issue for using a Mamdani FIS for large and complex data sets, is that the number of rules required to describe the system becomes excessively large and even if the system is less complex expert knowledge is still required to determine the membership functions (Zhou & Ghan, 2008; Tutmez & Tercan, 2007). This may introduce modeler bias and can be very time consuming. Furthermore, Tutmez & Tercan (2007) found the T-S method to be more simplistic and outperform the Mandami FIS for spatial estimation. The most common T-S methodology used for fuzzy spatial estimation stems from a paper by Setnes, Babuška, & Verbruggen (1998) in which they outline a method using fuzzy clustering clustering. This methodology has been applied to several aspects of spatial prediction in the geosciences. It has been used for grade estimation in mineral deposits (Muhammad & Glass, 2011; Tutmez et al., 2007), prediction of environmental variables in ground water (Tutmez & Hatipoglu, 2010), rainfall prediction (Kanjornit et al., 2016), and prediction of mechanical rock properties (Tutmez & Tercan, 2007). An extension of the T-S methodology referred to as Adaptive-Neural FIS (ANFIS) has also been used for spatial prediction in the literature. This method inserts the structure of the T-S methodology into a neural network, such that the steps of the T-S FIS are adjusted iteratively until an optimum is reached (Jang, 1993). The steps of the T-S FIS that are optimized by ANFIS are the spread of the

membership functions and the model parameters of each rule. Therefore, ANFIS can be considered an optimized version of the T-S FIS. However, problems exist with model interpretability and loss of control of parameters when ANFIS is employed, as well as being extremely computationally expensive (Kanjornit et al., 2016; Nayak and Sudheer, 2007). Furthermore, when using ANFIS the number of clusters and membership function type must still be specified manually, these factors have been shown to greatly the effect the performance of the TS method (Kanjornit & Wong, 2013; Tutmez et al., 2007). Given the computational expense, increased complexity, and loss of interpretability and adjustability, using the ANFIS technique may not be justified for spatial interpolation depending on the type of data.

2.3 Membership Functions

In order to determine the membership a certain geographic coordinate has to any cluster within the data a membership function must be applied. Gaussian type membership functions are the most commonly used in spatial fuzzy modeling due to their simplicity and flexibility (Kajornrit et al., 2016; Kord & Moghaddam, 2014; Tutmez & Hatipoglu, 2010; Tutmez and Dag (2007); Tutmez et al., 2007; Setnes et al., 1998; Sugeno & Yasukawa, 1993; Pham, 1996). Muhammad and Glass (2011) and Tutmez (2007) did employ triangular and trapezoidal membership functions, respectively, for ore grade estimation. This was justified because they were modeling a small number of linguistic regions in the data. The use of Gaussian membership functions does require significant assumptions and may greatly over simplify and misrepresent the complexity inherent in the data (Muhammad & Glass, 2011). However, for modeling the term “close” in an FIS, Gaussian membership functions have been shown to exhibit excellent predictive ability (Tutmez and Dag, 2007; Pham, 1996).

2.4 Fuzzy Clustering

The most commonly used fuzzy clustering method employed in fuzzy spatial modeling is Bezdek's (1981) fuzzy c-means (FCM) algorithm (Kajornrit, Wong, & Fung, 2016; Tumez & Hatipaglu, 2007; Tutmez & Hatipoglu, 2010; Tutmez & Tercan, 2007; Tutmez et al., 2007). Amini et al. (2005) deemed FCM to be very useful for clustering soil pollution data; however, Nourzadeh et al. (2012) showed that an extension of the FCM algorithm known as Gustafson–Kessel (GK) clustering out performed FCM for clustering soil pollution data. This is due to the fact that FCM is only able to extract spherical clusters from a dataset, whereas the GK method recognizes clusters with different geometric shapes, which is more representative of real world pollutant distributions (Muhammad & Glass, 2011). Muhammad and Glass (2011) compared FCM and GK clustering for fuzzy modeling and determined GK to be superior to FCM. Furthermore, Nayak and Sudheer (2008) compared fuzzy subtractive clustering and GK clustering and found GK clustering to be a more robust method.

2.5 Cluster Validation

Determining the optimal number of clusters is a key step in fuzzy clustering because it can affect the predictive ability of the subsequent fuzzy model (Sugeno &Yasukawa, 1993). A Cluster validation method proposed by Tutmez et al. (2007) has been used for spatial prediction applications and has been shown to adequately identify the number of clusters, which produce an accurate fuzzy model result (Kajornrit et al., 2016; Kajornrit & Wong, 2013; Tutmez & Haptipoglu, 2007). This method was designed specifically for spatially dependent geoscientific data and respects the inherent variability in the data, however its use has not been widespread. The combined use of Partition Coefficient and Classification entropy method has been employed for cluster validation using spatial data (Muhammad & Glass, 2011). Kord and Moghaddam

(2014) used Minasny and McBrateny's (2002) FuzME software for clustering and it's built in validation methods to evaluate the optimal number of clusters and so the adequacy of these validation methods could not be quantified.

2.6 Comparisons of Fuzzy Modeling, IDW, and Kriging

Direct comparisons of T-S fuzzy modeling to traditional interpolation methods are scarce in the literature, but what does exist, show minor outperformance of T-S fuzzy modeling over kriging and IDW. Tutmez and Dag (2007) compared a T-S fuzzy model and kriging for the prediction of lignite thickness for mineral inventory estimation. They used Gaussian type membership functions and FCM clustering with the Tutmez et al. (2007) cluster validation method. They found the fuzzy model had a lower Root Mean Squared Error (RMSE) and higher coefficient of determination (R^2). Tutmez & Hatipoglu (2010) compared the T-S fuzzy model and kriging for the spatial estimation of nitrate in an aquifer. They also used Gaussian type membership functions, FCM clustering, and the Tutmez et al. (2007) cluster validation method. They found the fuzzy model had a lower RMSE and Mean Absolute Error (MAE) than kriging and the fuzzy model also had a higher Variance Account For (VAF), indicating a more accurate prediction result. Kord and Moghaddam (2014) compared the T-S method and kriging to spatially analyze drinking water quality in an aquifer. They used FCM clustering, Gaussian membership functions, and cluster validity measures from the FuzME clustering software and found the fuzzy modeling techniques to outperform Kriging for RMSE, MAE, and R^2. A downside to fuzzy modeling cited in the literature is that its predictive ability tends to break down near the margins of the data where less overlap of membership functions occurs (Kajornrit et al., 2016).

2.7 Gaps in knowledge

The spatial predictive ability of the T-S fuzzy model with fuzzy clustering has been tested for a variety of spatially dependent geoscientific data types. Some of these studies include no comparison to established spatial prediction methods, but simply show the method is viable for their particular application. In the few examples, which do compare the T-S fuzzy modeling approach to the geostatistical method Kriging, fuzzy modelling shows a minor outperformance of OK. There are no observed comparisons between fuzzy modeling and IDW. Fuzzy set theory has been cited as a useful way to account for the uncertainty of scarce data (Khan, 2015; Zhang, 2009; Hwang & Thill, 2004; Bárdossy & Fodor, 2001; McBratney & Odeh, 1997). However, there are no occurrences in the literature validating this claim in the realm of fuzzy spatial modeling. Pham (1997) did conclude that fuzzy modeling could be employed in situations where there is insufficient data for variography to be performed prior to Kriging. This conclusion is however not quantified. In order to best quantify the performance of fuzzy modeling, the most commonly used methods identified in the literature will be employed in this research. Specifically, the T-S fuzzy modeling framework, using FCM clustering with Gaussian shaped membership functions. Although there is no clear consensus in the literature on which cluster validation method is best, the optimal number of clusters will be selected based on the models performance in predicting the respective validation sets.

Specifically, this research seeks to address the following:

1. Determine the performance of the T-S fuzzy model with FCM clustering using incrementally less spatially distributed data.
2. Compare the predictive ability of the T-S fuzzy model to that of IDW and OK using increasingly less spatially distributed data.

3. Methodology

3.1 Data Collection and Description

Marine sediment samples are very expensive to obtain and their results are often not released to the public due to the negative connotations associated with their contents. Because of the inability to obtain dioxin concentrations in marine sediment it was necessary to test the viability of the T-S fuzzy model using a different data source. This research employs spatially distributed soil geochemical data to compare the predictive abilities of the T-S fuzzy model, IDW, and OK under increasingly sparse data conditions. The geochemical signatures present in the data are the result of naturally occurring mineralization and are not related to anthropogenic sources. The data is useful, because the indicator geochemical signatures associated with the mineralization are analogous with several pollutants of concern (POC) in urban settings. The spatial distribution of the contaminates in the data are very similar to that of an anthropogenic sources, since both generally originate from point sources and spread out into the surrounding environment. The data provides 1535 spatially distributed sample points (Figure 3-1) each consisting of a 51 element spectral array with notable POCs: arsenic, mercury, and lead. Although no spatial data set would have an identical statistical distribution, the challenges faced during spatial interpolation are similar regardless of the data type. Therefore, use of this data still provides an adequate initial test of the relative predicative ability of the T-S fuzzy model under increasingly sparse spatial data density.

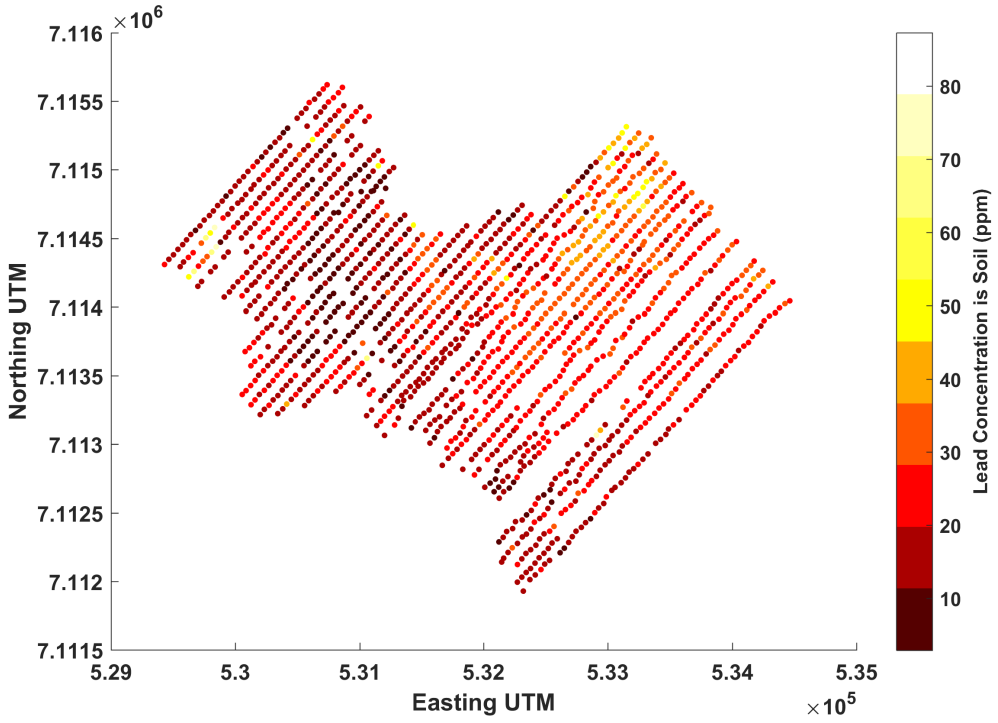


Figure 3-1, Spatial expression of data used for this research, Datum: NAD83, UTM Zone 8, Eastern Yukon Territory

The 1535 soil samples were collected on an approximate 50 by 50 m grid spacing at depths ranging from 0.1-1.2 m below ground surface. The samples have a composition ranging from well to poorly developed soils, to glacial sediment. The sample area is approximately 5000 x 4000 m in size and is located in Yukon Territory. The samples were collected between 2015/08/17 - 2015/08/22 and 2016/06/16 – 2016/06/25 by Archer, Cathro and Associates (1981) Ltd soil sampling team, including the author, under contract for ATAC Resources Limited. ALS Minerals in Vancouver BC analyzed the samples by metallurgical assay with inductively coupled plasma and atomic emission spectrometry. The spatial coordinates of each sample were recorded in the spatial datum NAD83 UTM Zone 8. Each data point is characterized by an easting and northing UTM coordinate in meters. Due to its relevance to human health and high rank as a POC in urban settings and its availability in the soil data, lead was selected to test the efficacy of

T-S fuzzy modeling against OK and IDW. Figure 3-2 summarizes the relevant statistical information about the utilized data. Most notable in the relevant statistics is the high Kurtosis, which indicates a large number of outliers exist in the data set. Further implications of the statistical distribution of the data with be discussed in Section 4.

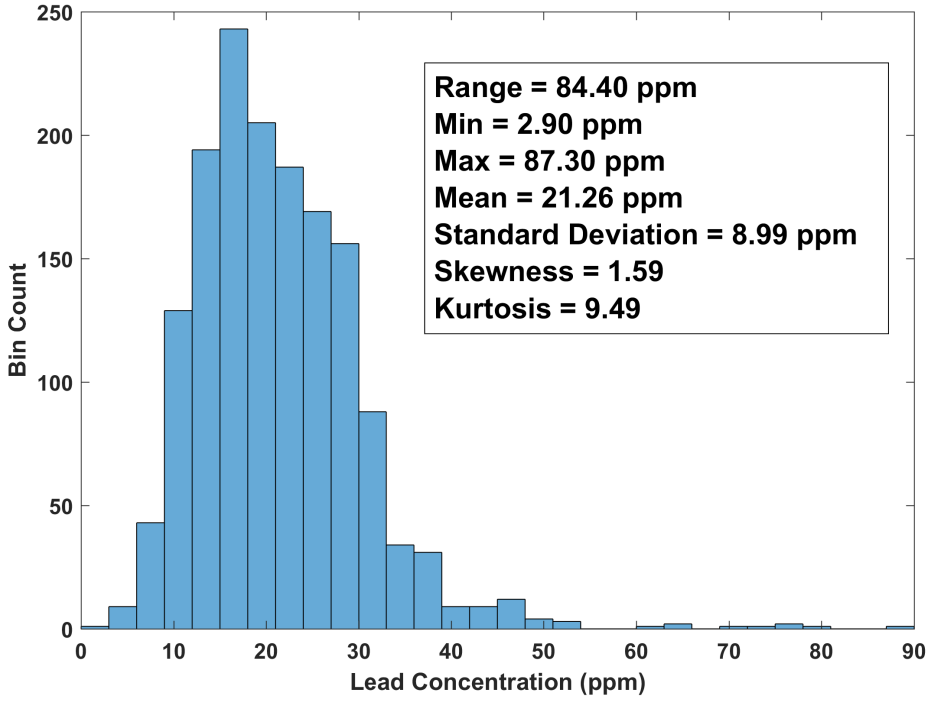


Figure 3-2, Histogram and summary statistics of Lead data

3.2 Traditional Spatial Interpolation

To test the relative performance of T-S fuzzy modeling, it will be compared to the most established non-geostatistical spatial interpolation method, IDW and to the most commonly used geostatistical spatial interpolation method, OK. IDW uses the linear combination of known sample values assuming greater weighting for nearer points and lower weighting for more distant points to solve unknown values (Li & Heap, 2008). This is calculated by the formula:

$$\hat{z}_i = \frac{\sum_{i=1}^{n} {z_i}/{d_i^p}}{\sum_{i=1}^{n} {1}/{d_i^p}} \quad (3)$$

where $\hat{z}_i$ is the estimated value, z_i is a known sample value, d_i is the distance between the point of interest and the i^{th} contributing point, n is the number of samples contributing to the estimation, and p is the power parameter, which dictates the relative weighting of points (Shepard, 1968). The neighborhood of the search area (n) is the number of samples contributing to the estimation and the power parameter is chosen to minimize mean absolute error of the interpolation. Most commonly, a power parameter of two (i.e. $p = 2$) is used and the size of n is chosen to minimize error (Li & Heap, 2008).

To preform spatial interpolation by kriging the semivariance (γ) of the dependent variable (z) between all sample pairs within the data must first be calculated to determine the spatial dependence of the data. The semivariance of z for distance (h) is calculated by the equation:

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n}(z(x_i) - z(x_i + h))^2 \quad (4)$$

where n is the number of sample pairs, $z(x_i)$ and $z(x_i + h)$ are the z values for the ith sample pair, the $\gamma(h)$ is then plotted against h to create an experimental semivariogram, which is used to determine important geostatistical features about the data. These important features include the range of spatial dependence defined by the distance (h) where the sill is reached, and the nugget, which is the y intercept of the model fitted to the experimental semivariogram (Burrough & McDonnell, 1998) (Figure 3-3).

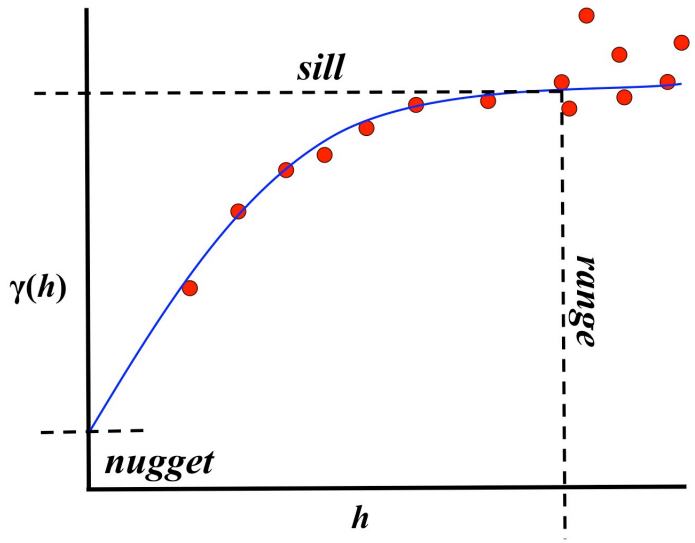


Figure 3-3, Theoretical experimental semivariogram illustrating range, nugget, and sill and model fit

The available models that can be fitted to the semivariogram include but are not limited to, linear, exponential, spherical, stable, and Gaussian. The variance of the data points on the semivariogram are then minimized to this fitted model in the least squares sense to produce kriging weights (λ) for the estimation of an unknown point within the data (Li & Heap, 2008). This accomplished by the core equation:

$$\hat{Z}(x_0) - \mu = \sum_{i=1}^{n} \lambda_i [Z(x_i) - \mu(x_0)] \qquad (5)$$

where $\hat{Z}$ is the estimated value at the point x_0 and μ is the mean within the search window, that is, within the predetermined range. Many more complex variants of kriging have been used, however for practicality, in this research OK is employed. Since variography and OK, and IDW are well-developed tools, this research will employ the GIS platform ArcMap 10.4.1 (ESRI, 2011) and use the Geostatistical Analyst Package to perform OK and IDW using optimized

parameters. Using the spatial predictions from ArcMap, model validation will be performed using MATLAB R2017a (MathWorks, 2017).

3.3 T-S Fuzzy Modeling

T-S Fuzzy modelling has 5 distinct steps: Fuzzy clustering (i), partitioning of data onto the coordinate axes to establish membership functions (ii), definition of FIS rule base (iii), Least Squares Estimation (LSE) of model parameters for each cluster (iv), and estimation of a final pollutant concentration (v) (Figure 3-4).

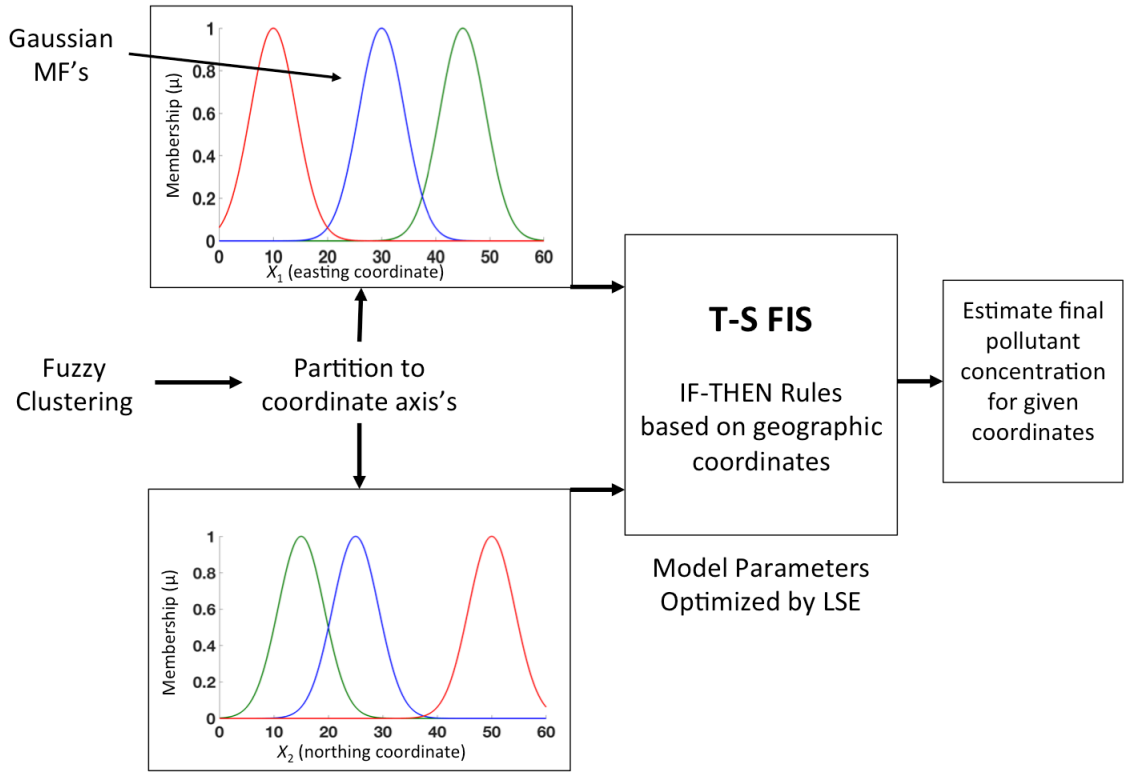


Figure 3-4, Summary of the T-S fuzzy model methodology (modified from Tutmez & Hatipolgu, 2010)

3.3.1 Fuzzy Clustering

GK clustering has shown significant promise for clustering spatially distributed soil geochemical data. However, due to the nearly ubiquitous use of the FCM clustering algorithm in the literature and its proven reliability, it was chosen for this study's fuzzy modeling approach. The FCM clustering algorithm performs an iterative minimization of the objective function J_m to create a fuzzy partition matrix from *N* number of data points into *c* classes (Bezdek, 1981).

$$J_m(U, \boldsymbol{v}) = \sum_{k=1}^{N} \sum_{i=1}^{c} (\alpha_{ik})^m \, \|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 \;\; k = 1, \dots, N \;\; i = 1, \dots, c \qquad (6)$$

where $U = [X, p]^T$ is the dataset to be clustered, X are the spatial UTM coordinates (easting/northing) and *p* is the pollutant concentration (Tutmez & Hatipolu, 2007), α_{ik} is the membership of the *k*th data point in the *i*th cluster, $\boldsymbol{x}_k$ is a vector describing the location of the *k*th data point, specifically its geographic coordinates and pollutant concentration (x,y,z), $\boldsymbol{v}_i$ is a vector describing the location of the *i*th cluster centre, *m* is a weighting exponent where $m \in (1, \infty)$ that defines the fuzziness of the clustering, if $m = 1$ clusters are crisp, that is data points cannot have membership to more than one cluster, as *m* increases cluster boundaries become softer (Kruse et al., 2007). An *m* value between 1.5 and 3 has been shown to produce the best results (the fuzziness coefficient is manually optimized by the modeler to produce the best results (Muhammad & Glass, 2011)). To perform FCM clustering α_{ik} is randomly initialized, and the initial cluster centre locations calculated by:

$$\boldsymbol{v}_i = \frac{\sum_{i=1}^{N} \alpha_{ik}^m \boldsymbol{x}_k}{\sum_{i=1}^{N} \alpha_{ik}^m} \qquad (7)$$

α_{ik} is then calculated by:

$$\alpha_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|}{\|\boldsymbol{x}_k - \boldsymbol{v}_j\|} \right)^{2/(m-1)}} \tag{8}$$

the objective function J_m is then calculated. Equations 7 and 8 are then repeated until J_m improves by less than a certain threshold or a maximum number of iterations is reached. The minimization of J_m is accomplished by a simple Picard iteration (Bezdec, Ehrlich, and Full, 1984). The output membership matrix also satisfies the probabilistic property, that each point can only have membership to different clusters such that each points total membership is equal to 1 (Tutmez et al., 2007; Bezdec et al., 1984).

$$\sum_{i=1}^{c} \mu_{ik} = 1 \, , \forall k = 1, \dots, N \tag{9}$$

The output of the FCM algorithm is two *m* x *n* matrices; the membership matrix and the cluster centre matrix. In the membership matrix *m* = *k* and *n* = *i*, such that a position in the membership matrix describes the k^{th} points belonging to the i^{th} cluster. Due to the probabilistic property (9), the sum of the n^{th} column of the membership matrix is always equal to 1. The cluster centre matrix stores the coordinate location and value of the i^{th} cluster in 3 columns. This methodology for the FCM algorithm is employed in MATLAB r2017a using the fcm.m function using the default parameters of a minimum objective function improvement of $1e^{-5}$ and a maximum number of iterations of 100 (Bezdec et al., 1984). These defaults are based on Bezdec et al. (1984) and were deemed a sufficient level of convergence for this research, as further refinement of the cluster centre locations is so minimal the effect on model performance is thought to be negligible. Determining the optimal number of clusters (cluster validation) can be accomplished many different ways, however there is no general consensus on which type of cluster validation

methods work best for which clustering algorithms, what types of data, and for which types of fuzzy models (Kajornrit & Wong, 2013). Since this is beyond the scope of this research, I adopt a simplistic cluster validation method where the number of clusters that produce the most accurate prediction of the validation data, is the optimal number (**Section 3.3.3)**. This will allow the most accurate predictions from the T-S fuzzy model to be compared to OK and IDW with no uncertainty about the number of clusters being chosen, impacting the models performance.

3.3.2 Clusters and FIS

After clustering, the membership of each data point to each cluster can be partitioned onto the two coordinate axes of the map space and a membership function assigned. The membership functions used are Gaussian shaped and given by the formula:

$$\mu_{A_i}(x_{ki}) = e^{\frac{-(x_{kij}-c_{ij})^2}{2\sigma_i^2}} \quad (10)$$

where $\mu_{A_i}(x_{ki})$ is the membership of the k^{th} point to the i^{th} cluster centre (c_i) on the j^{th} coordinate axis and σ_i defines the width of each membership function. Although the membership functions are Gaussian shaped, in this case σ represents the width of the membership functions independent of the standard deviation. Although the true distribution of each cluster may not be Gaussian, the use of Gaussian shaped membership functions is intended to model the linguistic term "close" in the FIS and so using the same membership function shape for each cluster is thought to be justified (Tutmez & Dag, 2007; Pham, 1997). σ is often held constant or chosen arbitrarily to optimize performance (Dag & Mert, 2008). Independently, Selection of sigma has been accomplished using geostatistics to determine the range of spatial dependence of each cluster (Tutmez et al., 2007). However, the purpose of this research is to determine if fuzzy modeling has the ability to outperform the geostatistical approach (OK) when the spatial

distribution of data becomes sufficiently sparse that variography fails. Therefore, I employ a simplistic approach to determine a constant σ based on model performance (**Section 3.3.3)**. Using a constant σ may over simplify differences between clusters; however, the FCM algorithm produces roughly circular clusters such that a constant σ for each cluster may be adequate (Muhammad & Glass, 2011). The impact of using constant width membership functions will be discussed further to assess its effect on model performance, if any. Once the membership functions are assigned, for any location on the map space, the degree of membership to each individual cluster can be easily determined. Next, a rule is introduced for each cluster, which take the form:

$$R_i = if\ x_1 is\ close\ to\ C_{i1}\ and\ x_2\ is\ close\ to\ \ C_{i2}\ then \qquad (11)$$

$$p_i = a_{i1}x_1 + a_{i1}x_2 + b_i \quad i = 1, \dots, K$$

where x_1 and x_2 are the Cartesian input coordinates, C_{i1} and C_{i2} are the cluster centres of the i^{th} cluster on the respective coordinate axis's. As previously discussed, the term "close" is modeled using the Gaussian shaped membership functions, essentially the closer a point is to a cluster centre the greater its membership (Tutmez & Dhag, 2007; Pham, 1997). p_i is the output of i^{th} rule calculated using the input coordinates and the model parameters a_{i1}, a_{i1}, and b_i for that rule. The model parameters for each rule are solved by least squares regression taking into account the membership each point has to each cluster. This is accomplished by using the form of a weighted least squares regression (WLS) where the weights for each points contribution are the memberships generated by fuzzy clustering. The input of the weights still satisfies the requirement that off diagonal elements of the matrix are null and because the FCM clustering algorithm generates clusters that are distinct as possible, the observed memberships from each cluster are unequal and thus the requirement of heteroscedasticity is maintained. The inputs of

the regression are: X_e =[X;1] which denotes a matrix of the Cartesian input coordinates with a column vector of ones, Γ_i which is a *m* x *n* matrix with the membership of each sample point to the i^{th} cluster on the main diagonal, and a column vector of the known pollutant concentrations (z) at the input coordinate locations. The solution to the least squares problem is (Muhammad & Glass, 2011; Setnes et al., 1998):

$$\left[a_{1,2}^T, b_i\right]^T = \left[X_e^T \Gamma_i X_e\right]^{-1} X_e^T \Gamma_i z \tag{12}$$

Using the resulting model parameters a crisp pollutant value for each rule can be estimated given a certain geographic coordinate. The outputs from all rules that produce consequents are aggregated to the solve the final pollutant estimation (p^*), given by the equation:

$$p^* = \frac{\sum_{i=1}^{K} \beta_i(x) p_i}{\sum_{i=1}^{K} \beta_i(x)} \tag{13}$$

where $\beta_i(x)$ is the degree of activation of the i^{th} rule calculated by the equation:

$$\beta_i(x) = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j), \quad i = 1,2,\dots,K \;\; n = 2 \tag{14}$$

where $\mu_{\tilde{A}_{ij}}(x_j)$ is the membership of the antecedent input of the i^{th} rule to the fuzzy set $\tilde{A}_{ij}$ on the *j* coordinate axis (Setnes et al., 1998). This methodology was implemented in MATLAB r2017a to best represent the most prevalent methods in the literature (Appendix A).

3.3.3 Sensitivity Analysis of Model Parameters

In order to determine the optimal number of clusters and fuzziness parameter (m) for FCM clustering and the membership function width (σ) for the FIS, a pseudo-optimization must be performed. This is accomplished by iteratively increasing the number of clusters and at each

iteration varying *m* through a reasonable range, and at each iteration of *m* varying σ through a reasonable range (Figure 3-5). The model performance results from all combinations will then be assessed and the combination of number of number clusters, *m*, and σ that yields the most accurate prediction of the respective validation sets will be retained as optimal. This method seeks to produce the most accurate result possible for comparison with OK and IDW. The range for number of clusters was varied arbitrarily from 2-n to ensure a clear optimal performance was observed; n was initial set at 200 and the range was varied for each training-validation pair as needed to achieve an optimal prediction. The range for *m* was varied from 1.1 to 3 by 0.1 to ensure a clear optimal model performance was observed, based on optimal ranges stated in the literature. Finally, the σ were varied initially by 10 m increments from 30 m to 220 m to ensure a clear optimal width was obtained; the range was subsequently adjusted as required to achieve a clear optimal performance. The range test for σ is based on the assumption that equal membership coverage in the FIS on the two partitioned map coordinate axes will produce the most accurate result. The code used for the pseudo-optimization is included in Appendix A. Using this methodology will demonstrate which model parameters are required to produce the most accurate spatial predictions for this training and validation data and allow the comparison of the most accurate T-S fuzzy model results to OK and IDW.

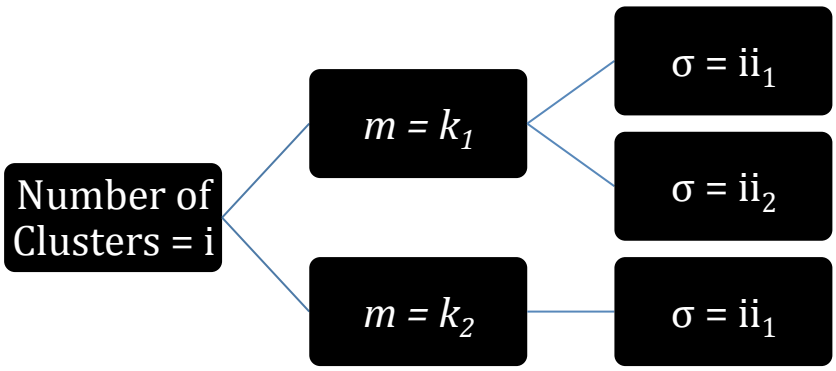


Figure 3-5, Workflow of the pseudo-optimization of clustering and FIS parameters, where i is the number of clusters at the i^{th} iteration, *k* is *m* at the k^{th} fuzziness coefficient, and ii is the σ at the ii^{th} width

3.4 Model Training and Validation

In order to assess the predictive ability of the T-S fuzzy model, kriging, and IDW, the sample data must be split into training and validation sets, where the training data is used to predict the known pollutant concentrations at the locations in the validation set. The standard approach in the literature is a validation data holdback of 20% to 35% (Kord & Moghaddam, 2014; Tutmez & Hatipoglu, 2010; Tutmez & Dag, 2007). The purpose of this research is to test the relative performance of each method under increasingly sparse data conditions. To accomplish this, the amount of training data used to will be incrementally reduced from 75%, to 66%, to 50%, to 33%, to 25%, and finally to 12.5%. At each increment the data not used for training will be utilized for validation. The training-validation subset selection (splits) at each increment will be performed by uniformly sampling from the total data set. Since the data is in the form of three column vectors, selecting different rows for training and validation is a simple indexing operation in MATLAB r2017a (Table 3-1). To determine the sensitivity of the T-S fuzzy model to regional variation in the training and validation sets, the systematic selection will include, three training - validation subsets at each increment. With the exception of the 50% -

50% training-validation subset, since only two combinations of training and validation data are possible when systematically selecting rows for training and validation.

Table 3-1 *Percent of total data used for training-validation increments and subsets and MATLAB indexing for row selection, where n = total number of rows*

Data Reduction Increment	**A**			**B**			**C**			**D**			**E**			**F**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**-**	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Training Set	75%			66.6%			50%			33.3%			25%			12.5%		
Validation Set	25%			33.3%			50%			66.6%			75%			87.5%		
Rows selected for Validation	1:4:n	2:4:n	3:4:n	1:3:n	2:3:n	3:3:n	1:2:n	-	2:2:n	All rows not selected for training are used for validation.								
Rows selected for Training	All rows not selected for validation are used for training.									1:3:n	2:3:n	3:3:n	1:4:n	2:4:n	3:4:n	1:6:n	3:6:n	5:6:n

These 18 training and validation sets will be used to make inferences about the performance of the T-S fuzzy model and its associated parameters. Since OK and IDW are well-established methods and both require significant modeler input through the ArcMap user interface, only the first subset from each reduction increment will be used to assess the relative performance of the models under increasingly sparse data conditions.

For each training and validation set the following procedures were followed for each method.

Kriging: Variography will be performed to determine range of spatial dependence of the training set, and fit a model to calculate the appropriate kriging weights for use in OK. The raster output from OK will then be exported from ArcMap and imported to MATLAB r2017a to perform model validation.

T-S fuzzy modeling: Cluster validation, selection of *m,* and σ will be performed by the pseudo-optimization described in Section 3.3.3. The most accurate prediction of each validation set using the optimal number of clusters, *m*, and σ will be retained as the result for that data reduction increment. The resulting pollutant predictions will then be compared to OK and IDW.

IDW: The training set from each split will simply be used to estimate the concentration at each validation point using optimized parameters in ArcMap. The raster output from IDW will then be exported from ArcMap and imported to MATLAB r2017a to perform model validation.

3.5 Model Performance

The key things that determine the competency of a spatial interpolator are its predictive ability, its ability to handle data of different types and variance, and the smoothness or abruptness of the surface generated (Li & Heap, 2008). The performance metrics used to assess the model performance of T-S fuzzy modeling, OK, and IDW will be the coefficient of determination (R^2), root mean squared error (RMSE), and the mean absolute error (MAE). R^2 is commonly used and provides a metric by which to assess the variance of the model prediction. A higher R^2 value indicates that a higher proportion of the total variation of the prediction is explained by the model (Tutmez & Hatipuglu, 2010). An R^2 value equal to 1 would indicate a "perfect" prediction. R^2 is calculated by the equation:

$$R^2 = \frac{\sum_i^N (z_i - z_i^*)^2}{\sum_i^N (z_i - \bar{z})^2} \tag{18}$$

and RMSE is calculated by the equation:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - z_i^*)^2} \tag{19}$$

Where in both cases z_i is the known value, z_i^* is the model estimate, $\bar{z}$ is the mean of the known values, and N is the number of points in the validation data set. RMSE has units equal to *z* and has been used extensively, however, it is sensitive to outliers (Willmott and Matsura, 2007). Willmott and Matsura (2007) suggest that MAE is the most credible metric for assessing predictions from spatial interpolation. MAE also has the same units as *z* and is calculated by the equation:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|(z_i - z_i^*)| \tag{20}$$

where z_i, z_i^*, and N are again the known value, the model estimate, and the number of validation points respectively. As is intuitive for both RMSE and MAE, a lower error is best. Although both MAE and RMSE are very similar metrics, using them together provides a convenient way to assess what proportion of the prediction errors are large, since RMSE is sensitive to larger errors (Willmott & Matsura, 2007). Such that, the closer RMSE is to MAE, the fewer large prediction errors exist. Additionally, it is important to determine the impact of changing the size of the training and validation sets on the observed performance. This can be accomplished by using the Akaike Information Criterion (AIC). Use of AIC allows the relative comparison of performance results from validation sets of different length or when a different number of model parameters is employed (Khan and Valeo, 2013). AIC is calculated by the equation:

$$\text{AIC} = \ 2[\ln(RMSE)] + 2(k) \tag{21}$$

Where RMSE is calculated from Equation 19, N is the length of the validation set, and k is the length of the training set. Additionally, the ability of the methods to predict extremely high pollutant concentrations is also very important because of the significance of high concentrations of POCs to human health. Therefore, the ability of the methods to predict outlier points will also

be assessed. This will be accomplished by calculating a separate MAE for points that are statistical outliers within the validation sets (oMAE). Although RMSE is more sensitive to outliers, when assessing the error of only points that are statistical outliers, one is essentially measuring the error of a new population. Therefore, since MAE measures the average magnitude of errors it was deemed a suitable metric. For this research, outlier points will be that which are more than three scaled median absolute deviations (MAD) away from the median; calculated using the isoutlier.m function in MATLAB r2017a with default parameters. oMAE will be calculated for the first data subset at each reduction increment with the predictions from IDW and Kriging and for the T-S fuzzy model using the most accurate predictions resulting from the optimal parameters determined in the pseudo-optimization. Using MAE, R^2, RMSE, AIC, and oMAE the performance of the three spatial prediction methods will be assessed at each iteration of the data reduction. In order to give MAE, R^2, and RMSE equal weighting, during both the pseudo-optimization and the data reduction analysis, a simple scoring system was introduced. Each model result for MAE, R^2, and RMSE will receive a score between 1 and 10, the mean score then determines the overall performance of that model result with 10 being best and 1 being worst. The ranges for the scoring system will be chosen arbitrarily based on the performance ranges observed from each metric. This approach seeks to produce enough separation between scores where some kind of conclusion can be reached about the differences in performance. Since oMAE is not an established performance metric, it will be examined separately at each increment of the data reduction and will not be included in the scoring system. Plotting the model performance results of each method under the increasingly sparse data conditions will allow conclusions to be reached regarding the efficacy of T-S fuzzy modeling compared to OK and IDW.

4. Analysis and Results

4.1 Creation of Training and Validation Sets and Scoring System

For this research a T-S fuzzy model using FCM clustering with constant width Gaussian shaped membership functions was used to make spatial predictions of lead concentrations in soil. The amount of data used to make the predictions was iteratively reduced to determine if the T-S fuzzy modelling approach has the ability to outperform OK and IDW under sparse data conditions. Prior to the spatial predictions being made and the data reduction analysis performed, the data was systematically separated into training and validation sets. The data in each training set was used to train the T-S fuzzy model, which was then used to solve the known lead concentrations at the locations in the respective validation sets. The predicted values were them compared to the known values and the performance assessed. Systematic selection of training and validation sets has been shown an adequate method for testing spatial interpolators (Liao et al., 2018). The purpose of this research was to test the relative performance of the T-S fuzzy model, OK, and IDW; the selection of the training and validation sets did not take into account over-fitting during training and validation set selection. Since, the purpose of systematically using less data at each increment was to simulate real world conditions where fewer samples were collected. Table 4-1 contains the lengths of data sets obtained during this selection and Table 4-2 summarizes the relevant statistics about each sub set.

Table 4-1 *Summary of lengths of training and validation sets*

Data Reduction Increment	**Training Set Length**			**Validation Set Length**		
Split #	1	2	3	1	2	3
A	1151	1151	1151	384	384	384
B	1023	1023	1024	512	512	511
C	767	768	-	768	767	-
D	512	512	511	1023	1023	1024
E	384	384	384	1151	1151	1151
F	256	256	256	1279	1279	1279

Due to the systematic technique used to select the individual sub sets of each data reduction increment, there are minor differences in the lengths of some training – validation subsets (splits) within the same increment, dictated by the odd or even nature of the row selection. This length difference of one between certain training – validation subsets at the same reduction increment is assumed to have a negligible impact on the comparison between the results from these sets.

Training Sets	Skewness			Kurtosis		
Split #	1	2	3	1	2	3
A	1.35	1.58	1.64	7.79	9.35	9.91
B	1.66	1.69	1.40	10.28	9.75	8.22
C	1.30	1.78	-	7.58	10.5	-
D	1.48	1.33	1.88	8.01	8.50	11.2
E	1.97	1.62	1.45	11.5	9.91	8.13
F	1.60	2.28	1.14	8.58	13.1	6.94
Validation Sets						
A	1.97	1.62	1.45	11.5	9.91	8.13
B	1.48	1.33	1.89	8.01	8.50	11.2
C	1.78	1.30	-	10.5	7.58	-
D	1.66	1.69	1.40	10.3	9.75	8.22
E	1.35	1.58	1.64	7.79	9.35	9.91
F	1.59	1.33	1.68	9.70	7.71	9.97

Training Sets	Range (ppm)			Mean (ppm)			Standard Deviation		
Split #	1	2	3	1	2	3	1	2	3
A	73.4	84.4	83.4	21.0	21.3	21.1	8.60	9.07	9.05
B	84.4	84.4	72.1	21.5	21.21	21.5	8.94	9.25	8.76
C	72.1	84.4	-	20.7	21.8	-	8.49	9.43	-
D	72.0	72.1	84.4	20.1	21.3	21.7	9.07	8.44	9.41
E	83.4	72.1	73.4	21.9	21.0	21.6	10.0	8.75	8.79
F	70.4	84.4	66	20.9	22.2	22.2	9.37	10.2	8.58
Validation Sets									
A	83.40	72.10	73.4	21.9	21.1	21.6	10.0	8.75	8.79
B	72.0	72.1	84.4	20.8	21.3	21.7	9.07	8.44	9.41
C	84.4	72.1	-	21.8	20.7	-	9.43	8.49	-
D	84.4	84.4	72.1	21.5	21.2	21.0	8.94	9.25	8.76
E	73.4	84.4	83.4	21.0	21.3	21.1	8.60	9.07	9.05
F	84.4	72.1	84.4	21.3	21.1	21.1	8.91	8.71	1.14

Table 4-2 *Summary statistics of all training and validation sets*

The results from the systematic selection of the training and validation sets reflect the statistics of the over-all lead data set, however, some discrepancies do exist between the training and validation sets. Specifically, phenomena that may affect model performance are where training

sets have a lower standard deviation, range, mean, and kurtosis than the validation sets. In these cases the validation sets will be more difficult to predict due to the increased variance and greater number of outliers. This is due to the fact that spatial interpolation methods generally predict towards the mean, such that if a validation set is more complex than its training set, model performance will suffer (Li & Heap, 2011). Conversely, if the training set contains a greater range of values than its respective validation set, it will be able to more easily predict the narrower range of values because the training set can more easily account for variation in the validation set. The results from the each model will be subsequently compared to the statistics for the training and validation sets to determine how great of an effect these discrepancies had, if any. The systematic selection of the training and validation sets produced relatively well spatially distributed validation points throughout the data at each increment. Figure 4-1 displays the map expression of data reduction increment A, subset 1 training and validation points. Minor groupings of validation points are visible, however, they are uniform throughout the entire map space since the selection of the training validation sets was systematic.

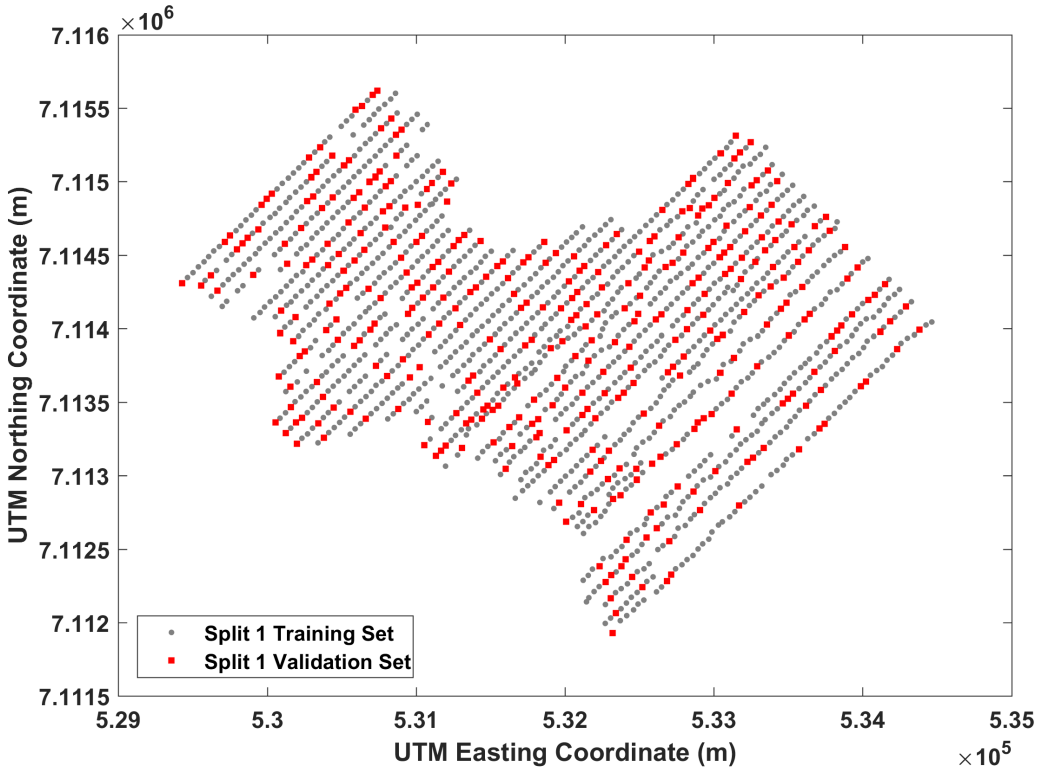


Figure 4-1, Map expression of data reduction increment A, subset 1 (split 1) training and validation sets

For each subset at the training reduction increments the locations of the validation points were varied to determine the T-S fuzzy models sensitivity to regional variation within the training and validation data. Figure 4-2 displays the map expression of data reduction increment A, subset 2 where the spatial differences in the training and validation sets are visible.

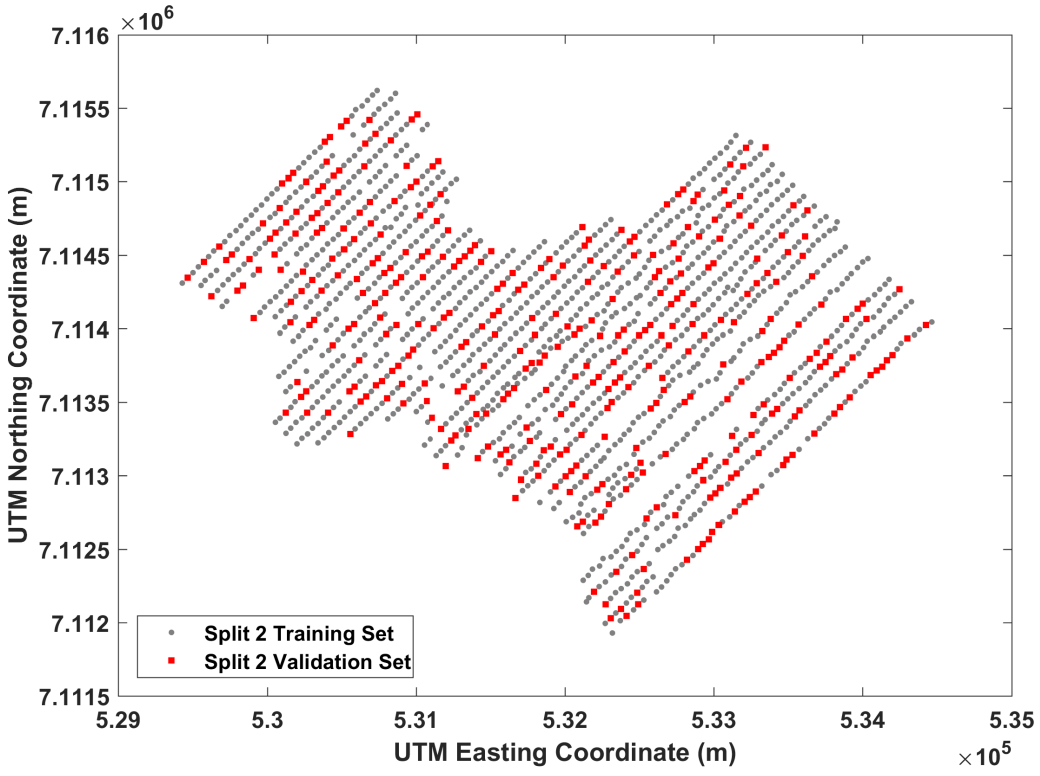


Figure 4-2, Map expression of data reduction increment A, subset 2 training and validation sets

As the amount of data used for training was reduced the number of validation points increased, as is visible Figure 4-3, which displays the map expression of data reduction increment E, subset 1. As with the initial grouping of validation points, when very little data is used for training, minor grouping of training points also occurs. Again, these points are evenly distributed throughout the map space and were considered adequate for simulating data sparsity in this research. The map expressions of the remaining data reduction increments and their subsets are displayed in Appendix B.

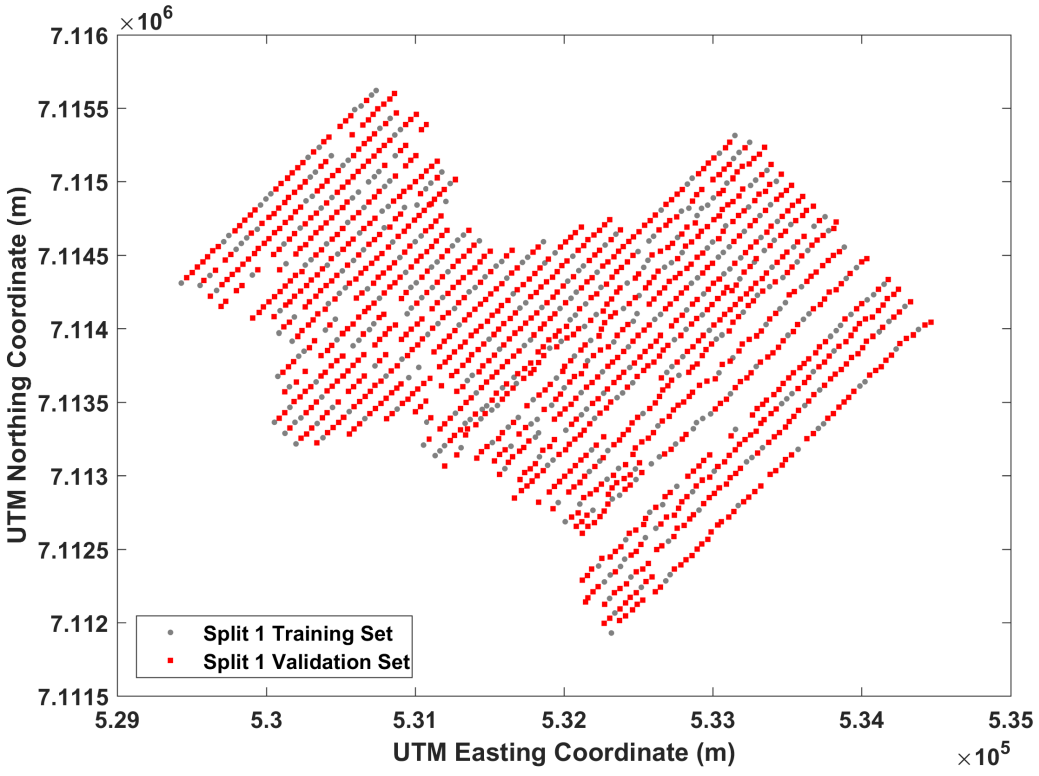


Figure 4-3, Map expression of data reduction increment E, subset 1 training and validation sets

After selection of the training and validation sets, prediction results were obtained for the T-S fuzzy model for all of the training – validation subsets at each data reduction increment using the pseudo-optimization and IDW and OK using the first training – validation subset from each increment. Prior to the selection of optimal model parameters and further analysis, the scoring system to rate the models was developed. For the three main performance metrics: MAE, R^2, and RMSE, the observed ranges of performance from each metric were analyzed to ensure the scoring ranges chosen for each metric were able to produce meaningful scoring separation between the models during the analysis (Table 4-3). For completeness, during the pseudo-optimization of the T-S fuzzy model, combinations of parameters that produced extremely poor model results were tested, therefore, the total ranges observed from the T-S fuzzy method were

deemed to large to include. To produce a meaningful scoring system for comparisons, the top 50% of the range of results from OK and IDW were chosen to define the scoring system. These ranges were divided into 10 equal increments to produce the bins used in the scoring system, where a score of 10 is the best and 1 is the worst (Table 4-4). The average of the three performance metrics scores were used to determine the optimal model parameters based on performance for the T-S fuzzy model and to analyze the relative performance of OK and IDW.

Table 4-3 *Summary of Max and Min model performance from OK, IDW, and the T-S fuzzy model*

Performance Metric	**T-S Fuzzy Model**		**OK**		**IDW**		**Total OK and IDW Range**	
	Max	Min	Max	Min	Max	Min	Max	Min
MAE	43.300	4.268	4.758	4.153	4.483	4.092	4.758	4.092
R^2	0.492	0.022	0.516	0.339	0.529	0.395	0.529	0.339
RMSE	104.900	6.480	7.844	6.334	7.515	6.245	7.844	6.245

Table 4-4 *Bin ranges for model scoring*

	Performance Metrics		
Score	**MAE**	**R^2**	**RMSE**
10	≤ 4.0920	≥ 0.5290	≤ 6.2450
9	≤ 4.1586	≥ 0.5100	≤ 6.4049
8	≤ 4.2252	≥ 0.4910	≤ 6.5648
7	≤ 4.2918	≥ 0.4720	≤ 6.7247
6	≤ 4.3584	≥ 0.4530	≤ 6.8846
5	≤ 4.4250	≥ 0.4340	≤ 7.0445
4	≤ 4.4916	≥ 0.4150	≤ 7.2044
3	≤ 4.5582	≥ 0.3960	≤ 7.3643
2	≤ 4.6248	≥ 0.3770	≤ 7.5242
1	≤ 4.6914	≥ 0.3580	≤ 7.6841

There are several issues with this scoring system; mainly it is overly sensitive, such that very minor increases in performance are rewarded with a much larger score. Previous research has

shown that scoring systems, which are overly complex and represent more variation, than is present in the data's population are not representative of actual performance (Khan, 2015). Initial exploration of bin-widths indicated that because the performance of the three methods is very poor for the spatial prediction of the data used in this study and the difference between model performance at each training reduction increment is very small, use of a more representative scoring system with wider bins would only lead to the conclusion that model performance is essentially unaffected by using less data. However, others have shown this is not the case (Li & Heap, 2011). Therefore, for the sake of making some kind of inference about the T-S fuzzy models performance under sparse data conditions the scoring system in Table 4-4 was employed. Because of the known pitfalls with using such a sensitive scoring system, the MAE and R^2 values will also be extensively analyzed to ensure false conclusions are not reached.

4.2 T-S Fuzzy Model Performance

Once the scoring system was in place the results of the pseudo-optimization were analyzed. When determining the optimal number of clusters, it was observed that the performance of the of T-S fuzzy model would reach an initial peak or plateau, where subsequent model runs using a higher number of clusters would not produce a higher score (Figure 4-4). In the interest of maintaining minimal complexity within the model bases on the principle of Occam's razor and to reduce the risk of over clustering, the lowest number of clusters that yielded the highest model performance was retained as optimal for each data reduction increment and subset. In cases where more than one T-S fuzzy model iteration with the optimal number of clusters, but different parameters, produced the same mean score, the result with the lowest MAE was deemed optimal. MAE was selected because of its prevalence in assessing the accuracy of spatial predictions (Willmot & Matsuura, 2006). Figures 4-4 displays the results from the

selection of optimal number of clusters for data reduction increments A-F for subset 1. In each case the highest mean score from each number of clusters tested is displayed. Therefore, for each cluster the results being displayed are the optimal performance with optimal *m* and σ for that number of clusters.

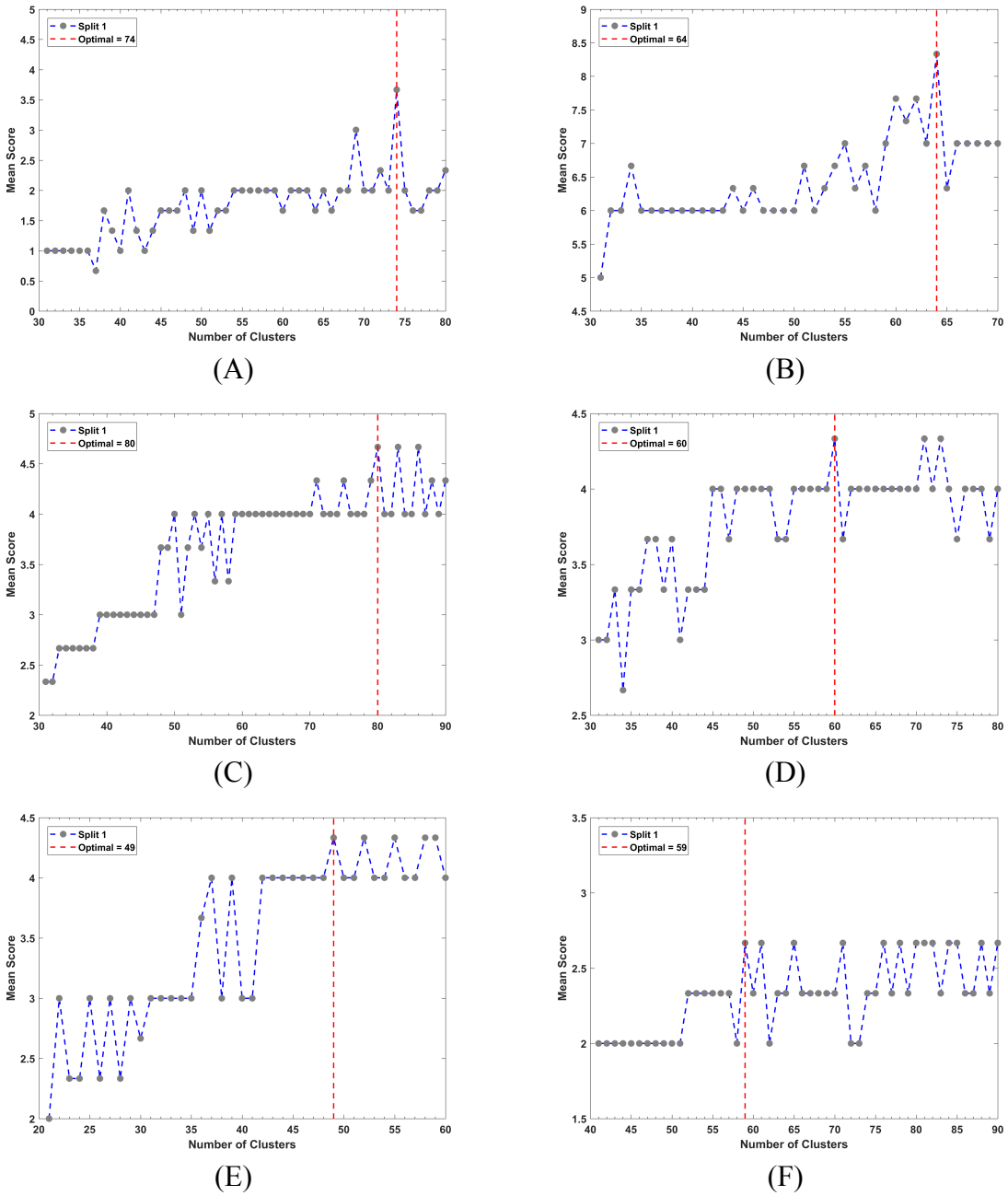


Figure 4-4, Results from the determination of the optimal number of clusters for data reduction increments A-F subset 1

In all cases a clear initial peak was reached and dictated the optimal number of clusters for that particular training set. The optimal number of clusters for each set is displayed in Table 4-5 and the remainder of the results for determining the optimal number of cluster for the remaining training and validation subsets are displayed in Appendix B. After identifying the number of clusters that yielded the highest model performance, the combination of fuzziness (*m*) and membership function width (σ) that contributed to the highest mean score was identified. Figures 4-5 displays the results from determining the optimal *m* value for data reduction increments A-F for subset 1. In each case the highest mean score for each *m* value tested with the optimal number of clusters is displayed.

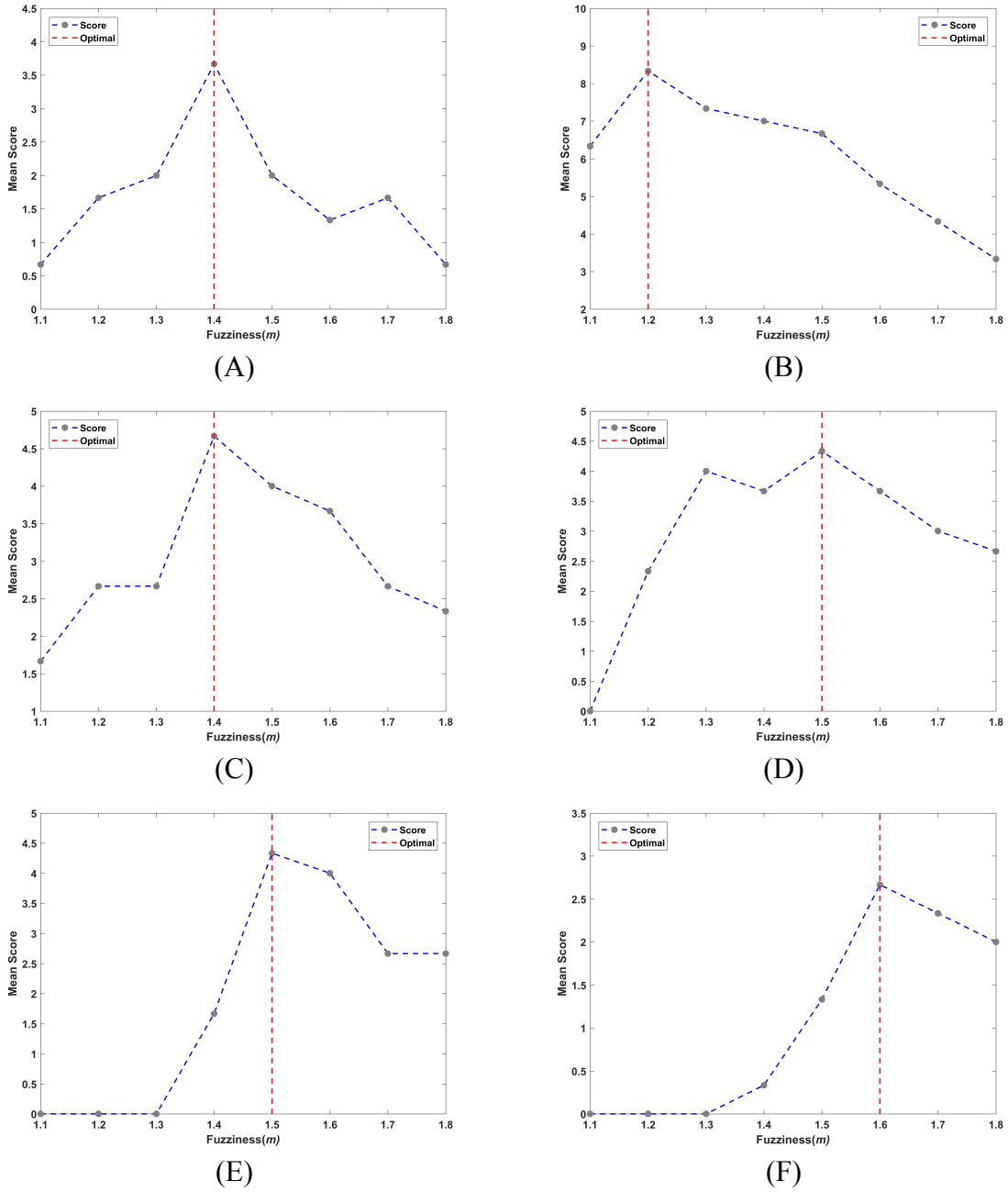


Figure 4-5, Results from the determination of *m* for the optimal number of clusters for data reduction increments A-F subset 1

For subset 1, increments A-F a single *m* value produced the highest model performance for each increment of the data reduction and the *m* value had a visibly large effect on the performance of the T-S fuzzy model. To further investigate the effect of the number of clusters on *m*, Figure 4-6

displays the MAE and Figure 4-7 displays the R^2 for each number of clusters for the relevant *m* values tested in the pseudo-optimization where σ is at its optimal. From figures 4-5, 4-6, and 4-7 An *m* value between 1.4-1.6 was the most common for the training data used, it is surmised that the *m* value for different spatial data would be reliant on the spatial dependence of the data its self. From Figure 4-6 its clear that as training data becomes sparser, the selection of *m* becomes more critical. Such that, when the spatial density of the data is high, most values of *m* yield an adequate prediction but as the amount of training data used becomes more sparse, *m* values that do not sufficiently increase fuzzy overlap perform very poorly. Additionally, Figure 4-6 shows that as the amount of data used for training is reduced, the selection of the number of clusters has very little impact on MAE and a greater impact on R^2. Using the scoring system still yielded a clear absolute model performance, but the performance was only slightly better that if fewer clusters were used. This indicates a major challenge when using T-S fuzzy modeling; how to determine the optimal number of clusters when results are similar. Since the purpose of this study was to compare the relative performance the T-S fuzzy model to OK and IDW, comparing the best possible score was advantageous, but because of the insensitivity of the model results to the number of clusters, if the T-S fuzzy model was being used as a spatial interpolator, determining the exact number of clusters would be less critical. Although multiple values of *m* displayed the ability to produce an accurate prediction, when data density became sparser, a selection of *m* that was to low, had a severe impact on model performance. Therefore, selecting an adequate *m* value is important in achieving an accurate prediction, specifically when data density is sparse. The remainder of the results for determining the optimal *m* for the remaining training and validation subsets are displayed in Appendix B.

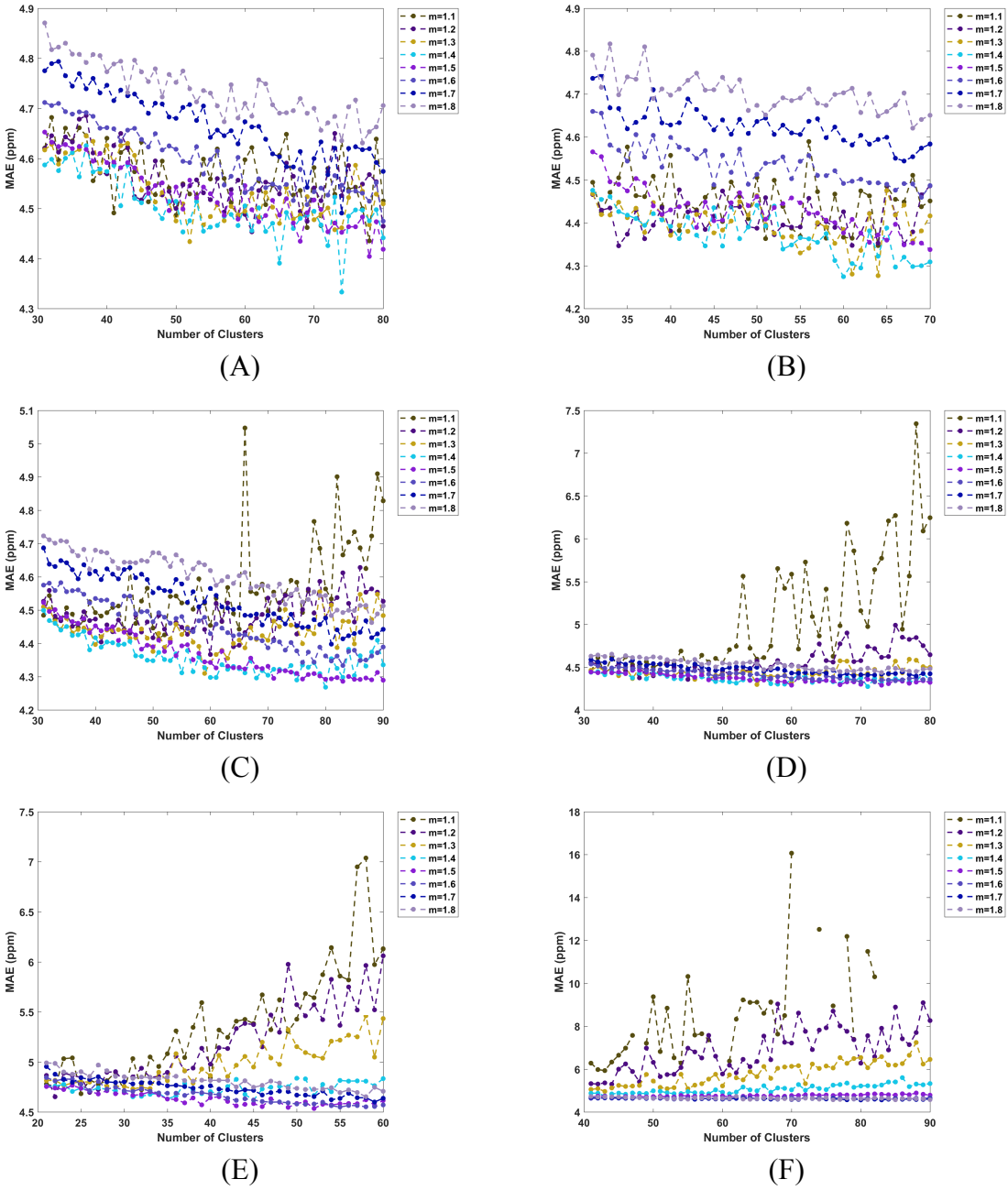


Figure 4-6, MAE plotted against number of clusters for *m*=1.1 to 1.8 for Training sets A-F, subset 1

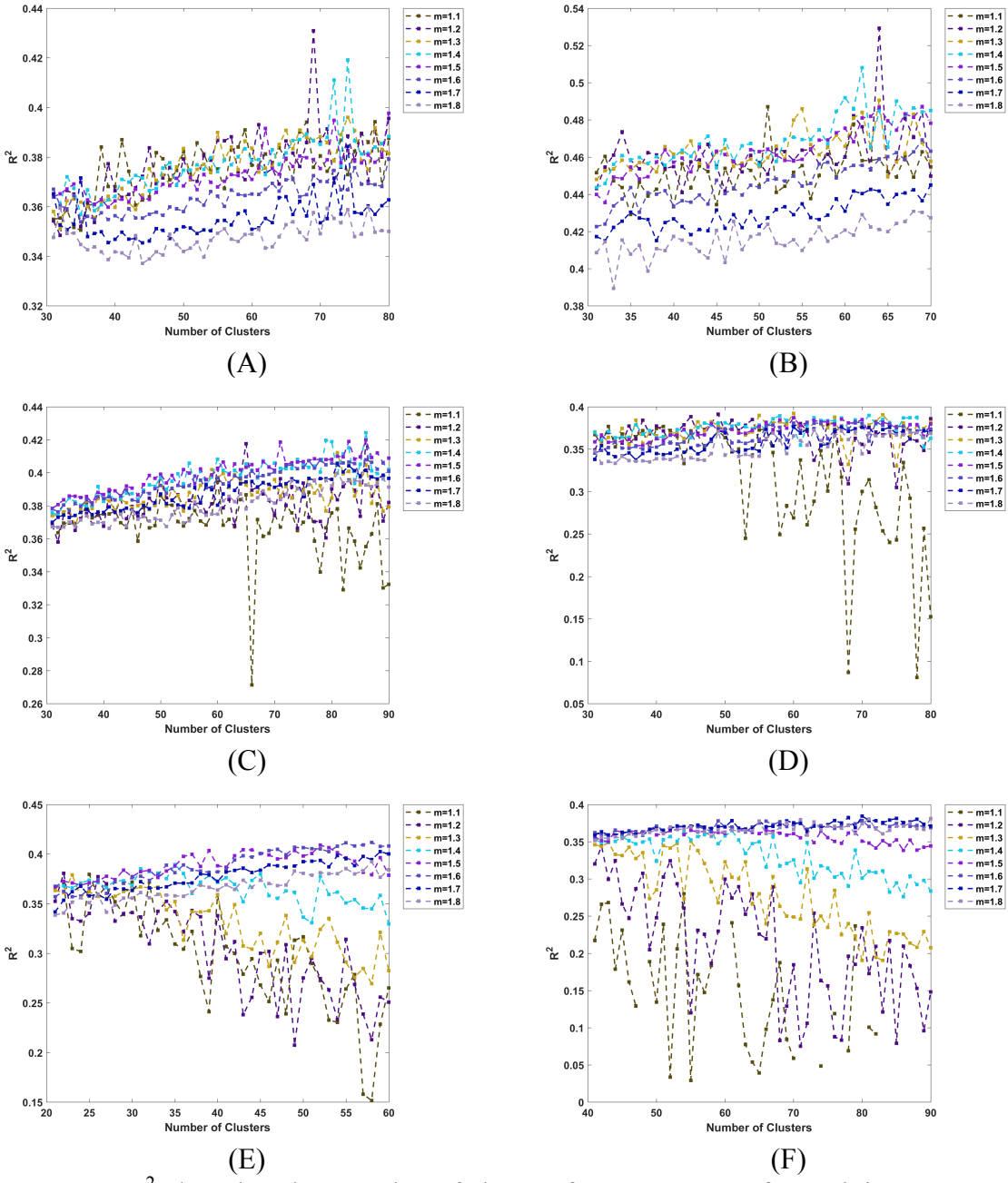


Figure 4-7, R^2 plotted against number of clusters for *m*=1.1 to 1.8 for Training sets A-F, subset 1

Finally, the optimal σ, which produced the highest model performance, was identified. Figure 4-8 displays the model performance of for σ, while the optimal number of clusters and *m* are constant for data reduction increments A-F for subset 1. σ appears to be a much less sensitive parameter than number than *m*. In most cases as σ was varied, an optimal model score was

obtained using many different widths of membership function for all subset 1 increments except increment B. In all other cases for subset 1 the MAE was used to determine the optimal width. Although the changes in MAE were very subtle a clear minimum was always observed and allowed the selection of an optimal width. Figure 4-9 and 4-10 display the effect of number of clusters on σ. As is clear in Figure 4-8 many different widths are capable of producing an accurate model result. The observed range of σ that generally produced an adequate result are 80 m to 150 m, which was not greatly effected by the amount of data used to solve the validation points. Although not conclusive, the minimal effect σ has on over-all performance may confirm using constant width membership functions with FCM clustering for this specific data had a negligible impact on performance.

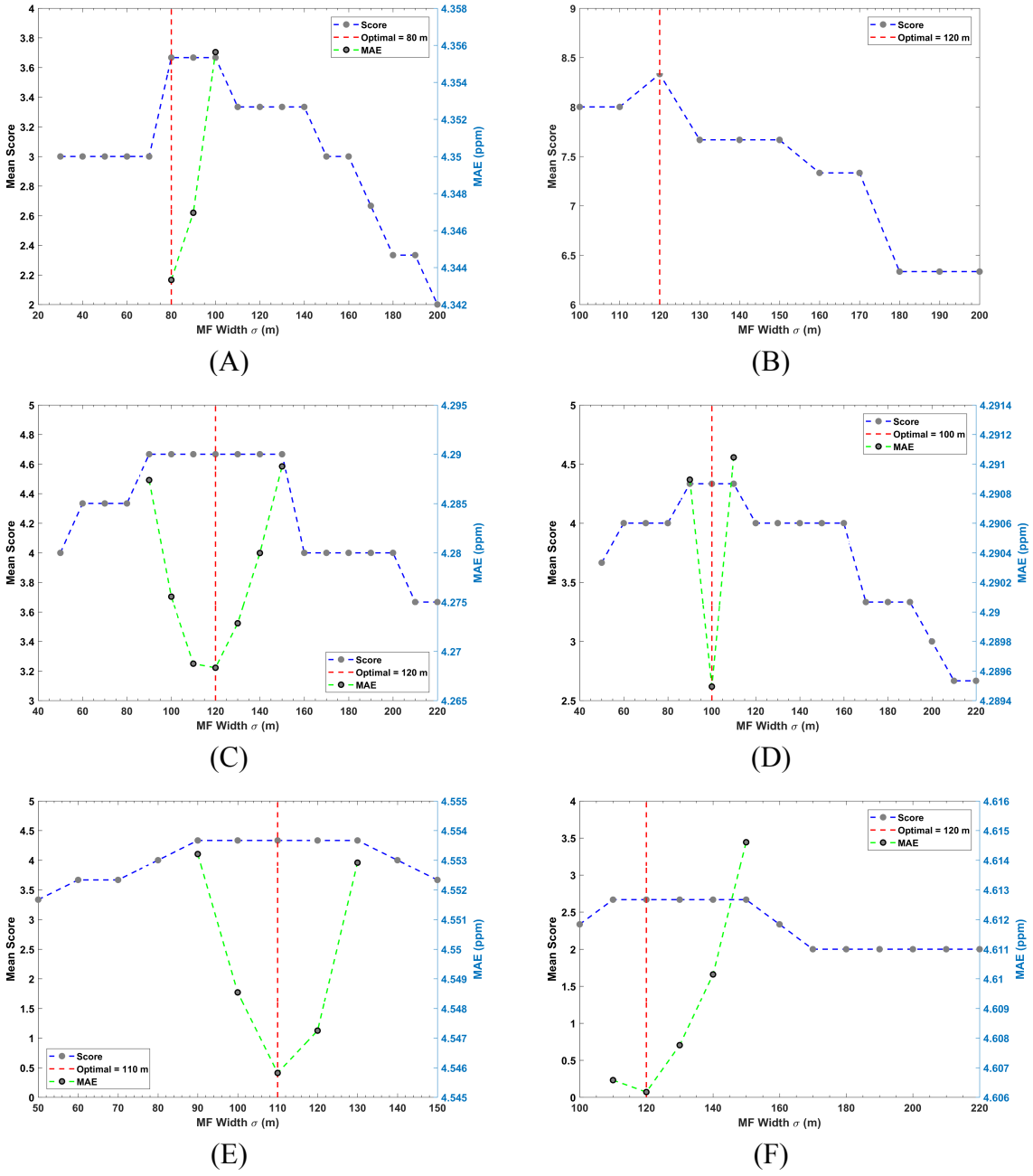


Figure 4-8, Results from the determination the optimal σ for data reduction increment A-F, subset 1

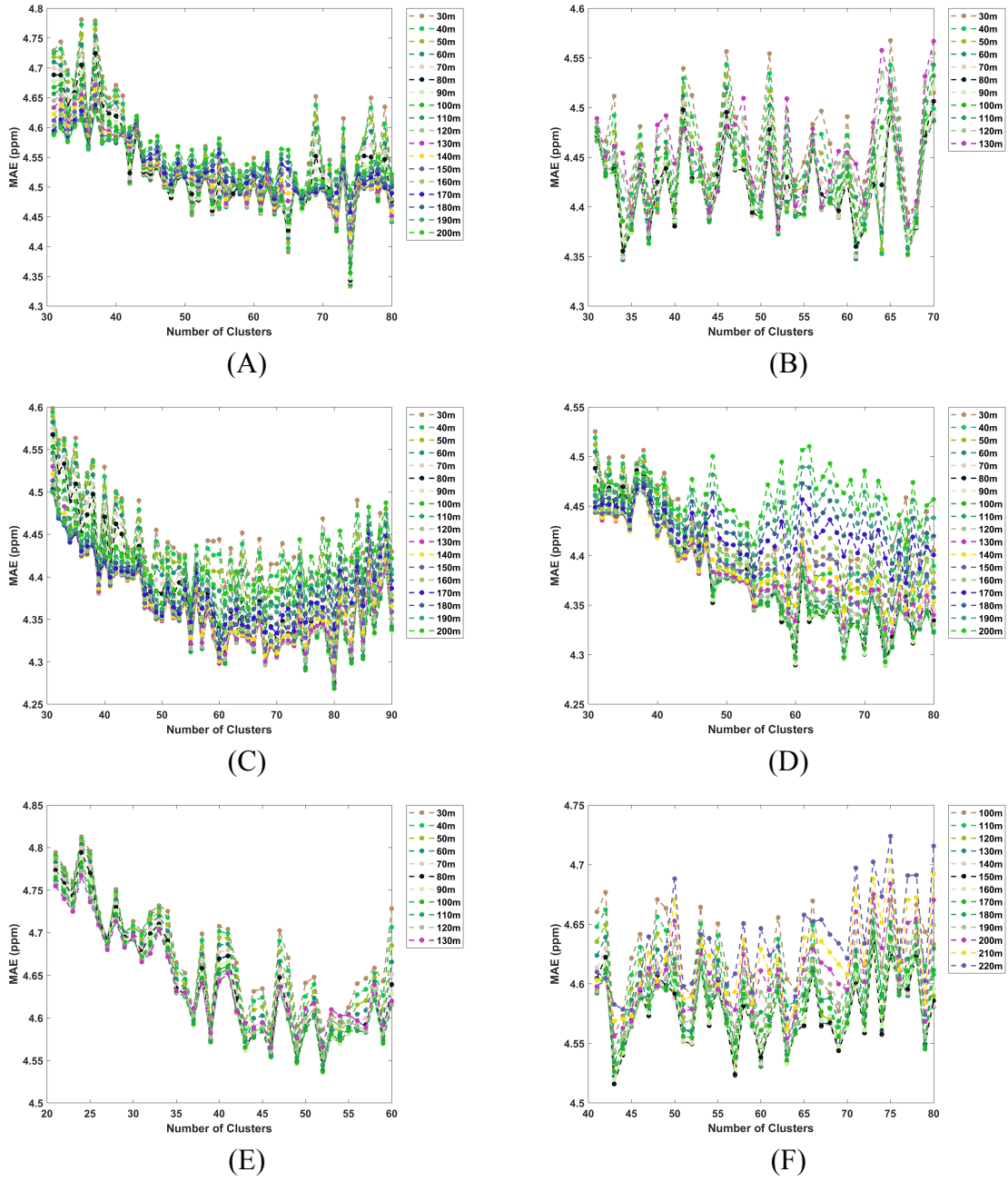


Figure 4-9, MAE plotted against number of clusters for ranges of σ tested during the pseudo-optimization for Training sets A-F, subset 1

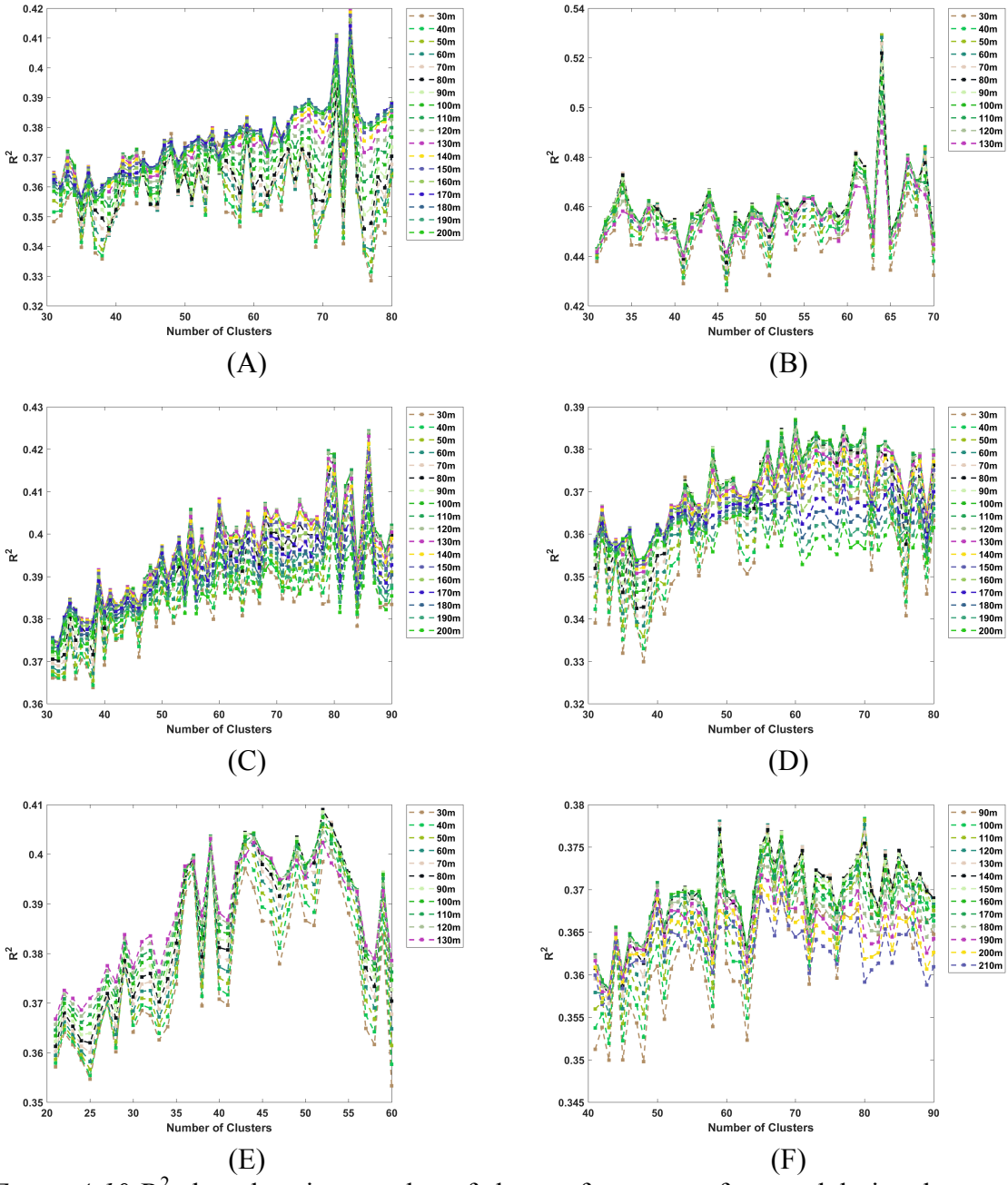


Figure 4-10, R^2 plotted against number of clusters for ranges of σ tested during the pseudo-optimization for Training sets A-F, subset 1

The results of the pseudo-optimization for each parameter are summarized in Table 4-5. By plotting the mean of the 3 subsets from each reduction increment, several conclusions can be reached (Figure 4-11).

Table 4-5 *Results for pseudo-optimization of model parameters for the T-S fuzzy model*

Data Reduction Increment	**A**			**B**			**C**		
Split #	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
# of Clusters	74	78	70	64	66	60	80	68	-
m	1.4	1.5	1.1	1.2	1.8	1.2	1.4	1.5	-
σ (m)	80	50	90	120	70	140	120	60	-
Data Reduction Increment	**D**			**E**			**F**		
Split #	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
# of Clusters	60	49	75	49	51	29	59	53	64
m	1.5	1.4	1.5	1.5	1.5	1.5	1.6	1.7	1.5
σ (m)	100	140	100	110	140	180	120	120	160

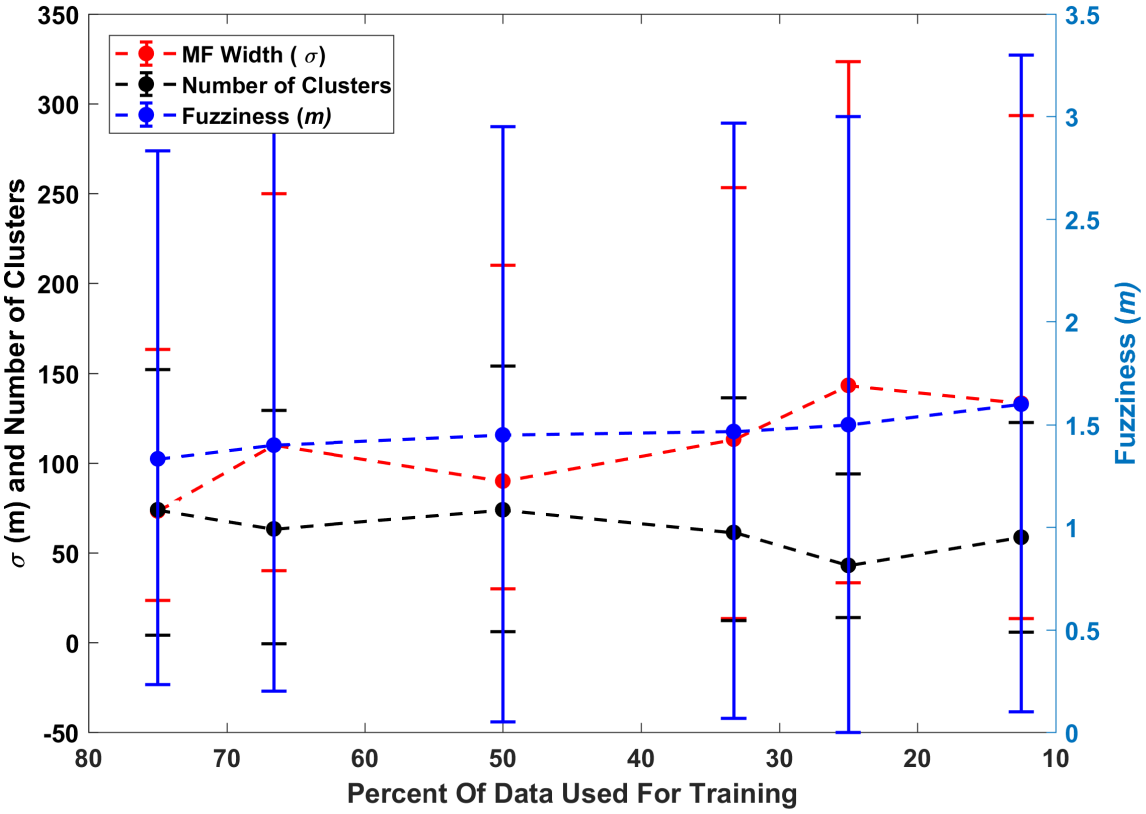


Figure 4-11, Mean values with error bars of T-S fuzzy model parameters for subsets 1-3 plotted through the data reduction

A minor trend exists for *m* and σ that as the spatial distribution of the data becomes sparser, the optimal *m* and σ increase and in general the number of clusters decreases. This is an intuitive result, as the amount of data available for clustering becomes smaller as does the optimal number of clusters. To produce the most accurate result possible from the FIS, the membership function coverage throughout the map space must be maintained. The increased *m* values, increase the fuzzy overlap between clusters, this in conjunction with the increased membership function widths accomplishes the increased coverage. The optimal number of clusters shows a weak correlation with the standard deviation and kurtosis of the training data used for clustering (Figure 4-12). An increase in the standard deviation and kurtosis indicates a greater complexity in the data, which inherently has more distinct regions within it, leading to more clusters.

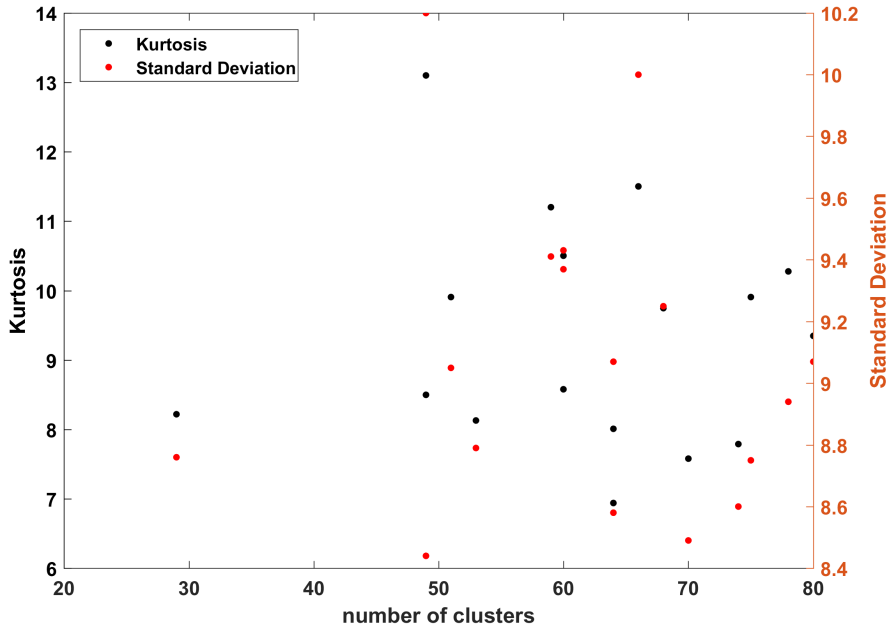


Figure 4-12, Trend analysis of kurtosis and standard deviation as a function of number of clusters

Using the optimal model parameters determined during the pseudo-optimization the performance of the T-S fuzzy model can be assessed. The initial step of the T-S fuzzy model is FCM clustering to break the initial training data down into smaller more easily modeled fuzzy regions. The results of the FCM clustering were relatively consistent for all data reduction increments and subsets, with clusters evenly distributed through out the map space. Figure 4-13 and Figure 4-14 display the spatial extent of the clusters with the training data used for clustering for data reduction increments A and F for subset 1. The location of the training points within the contours indicates the membership the respective points have that the cluster centres. The contours in Figure in 4-13 and 4-14 represent the gradient of membership between the different cluster centres and allow the visualization of how the training points contribute to each cluster centre. However, the contour plots are only capable of displaying the membership a point has to a single cluster centre. When in reality each point will have membership to many different cluster centres. All contour plots were generated using a function developed by Balasko, Abonyi, and Feil (2005).

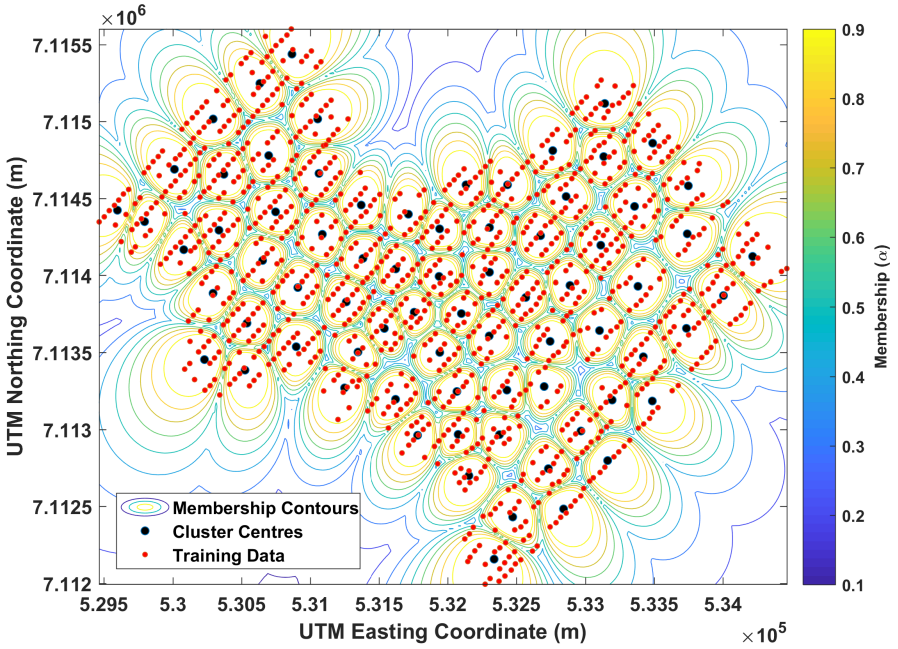


Figure 4-13, Visual representation of clustered training data for data reduction increment A, subset 1

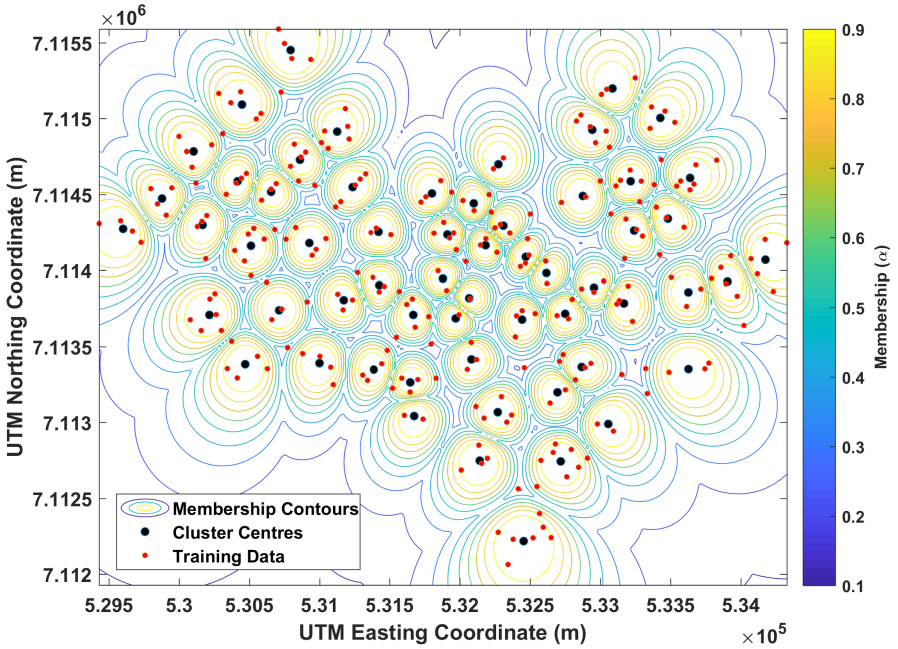


Figure 4-14, Visual representation of clustered training data for data reduction increment F, subset 1

Cluster centres often occur concurrently with grouped training data, this is not however always the case. Figure 4-15 and 4-16 provide a zoomed in view of the central map region for data reduction increments A and F, subset 1 respectively.

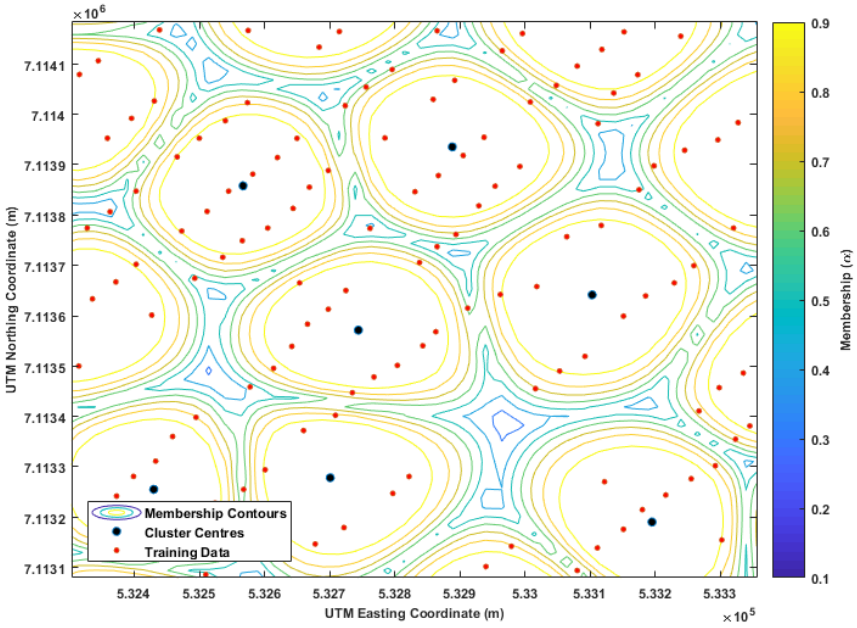


Figure 4-15, Visual representation of clustered training data for data reduction increment A, subset 1, zoomed view of the central map area

Very clear in Figures 4-15 and 4-16 are the differences in the density of the training data. In Figure 4-15 only small regions exist (~100 m^2) that do not have at least a membership of 0.3 to a cluster centre. As where in Figure 4-16 a large region (~400 m^2) exists where there the membership to the surrounding cluster centres is very low. Because membership to the surrounding clusters within this area is so low, it would be difficult for the T-S fuzzy model to produce an accurate prediction from within this regions or regions similar to this because the any point within these regions would fail to produce a meaningful weight in the FIS. Due to these regions low degree of membership and subsequent low weighting in the FIS, they would be more

likely to under predict pollutant concentrations, which may explain the reduced performance of the T-S fuzzy model once the data density become sufficiently sparse. Contour plots from all reduction increments and subsets are displayed in Appendix B.

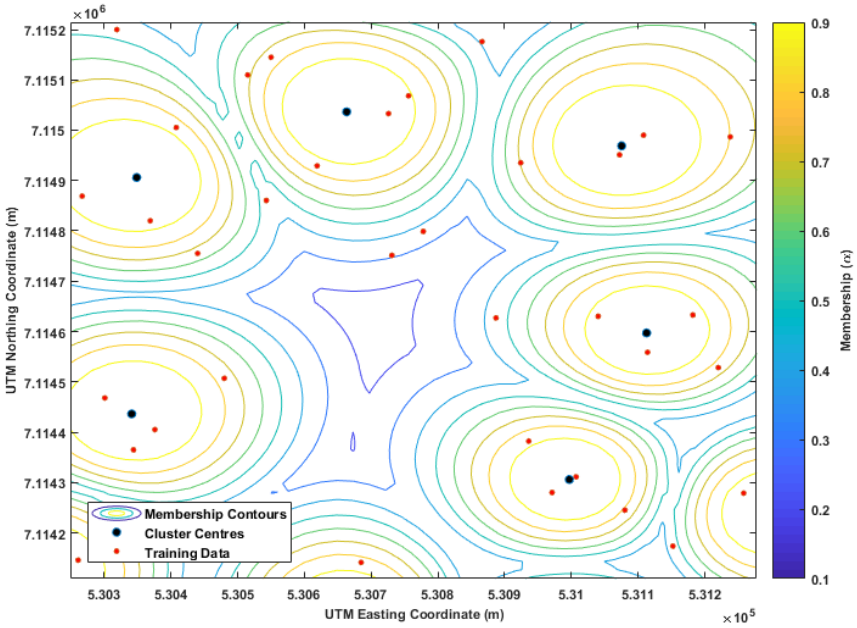


Figure 4-16, Visual representation of clustered training data for data reduction increment F, subset 1, zoomed view of central map area

After FCM clustering was performed the membership matrices and cluster centre matrices were subjected to the FIS, which for the given validation set, made predictions of lead concentrations at the locations specified in the validation set. Using the optimal width for the membership functions, the FIS calculated the membership to clusters within the map space based on their geographic position. Figure 4-17 displays the membership functions for the easting and northing coordinate axis for data reduction increment E, subset 3. This set was chosen because its results from the pseudo-optimization had the fewest number of clusters, making for easier visual analysis.

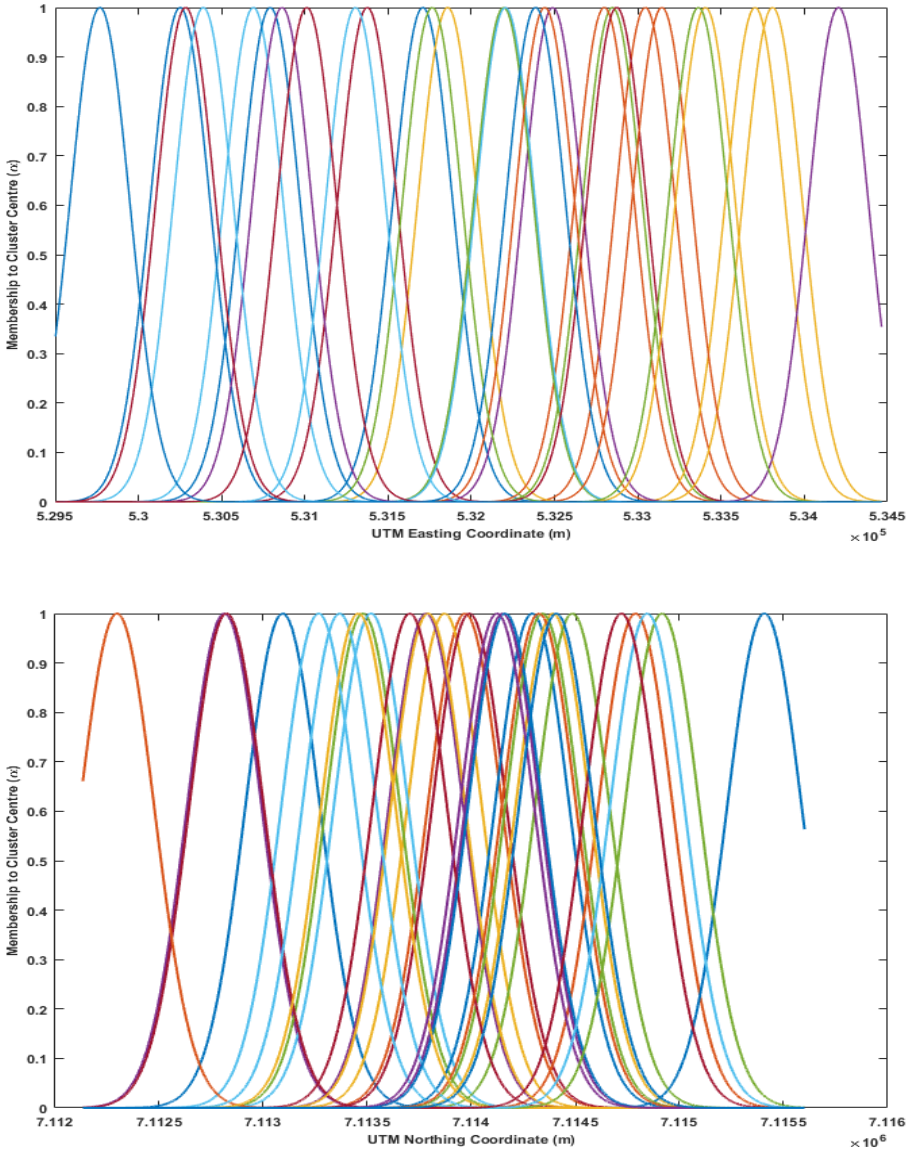


Figure 4-17, Membership functions for the easting and northing coordinate axis of the FIS for data reduction increment E, subset 3

Due to the size of the area in the analysis, when looking from either coordinate axis many membership functions overlap one another. However, it is important to remember that in the FIS when solving a concentration at a validation point location, membership to both coordinate axes is necessary to produce a meaningful output from that rule. To better visualize the spatial overlap

of the membership functions Figure 4-18 displays a zoomed view of the membership functions on the eastern margin of the easting coordinate axis of data reduction increment F, subset 1.

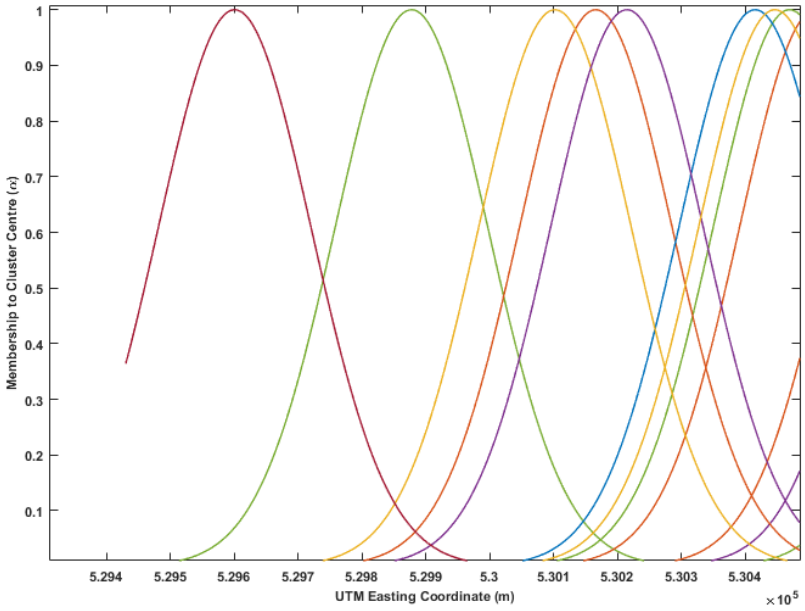


Figure 4-18, Membership functions for the easting coordinate axis from the FIS for data reduction increment F, subset 1, view is zoomed to east end of the axis to better view membership function overlap

Each tick on easting coordinate axis in Figure 4-18 is equal to 100 m. Figure 4-18 also illustrates the low density of membership functions that occur at the margins of the map space. This may lead to break down of spatial predictions at the margins of the data (Kajornit et al., 2016). Using the optimal constant width membership function the FIS was able to make predictions of lead concentrations at the validation point locations for each respective set. Using the optimal parameters for each data set it was possible to calculate an MAE for the outlier points within each validation set for comparison during the data reduction and between the different modeling approaches. Table 4-6 contains the complete results for each data reduction increment and all subsets for all of the aforementioned performance metrics and Table 4-7 includes the individual

performance metric scores and mean scores resulting from the pseudo-optimization for the optimal parameters.

Table 4-6 *Complete Results for the T-S fuzzy model for the validation data from all data reduction increments and subsets*

Data Reduction Increment	**A**			**B**			**C**		
Split #	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
MAE (ppm)	4.34	4.15	4.00	4.36	4.03	4.51	4.27	4.34	-
RMSE (ppm)	7.68	6.92	6.13	6.25	6.05	7.60	7.23	6.51	-
R^2	0.416	0.377	0.515	0.529	0.492	0.360	0.419	0.422	-
AIC	2306.1	2305.9	2305.6	2049.7	2049.6	2052.1	1538.0	1539.7	
oMAE (ppm)	41.88	41.30	23.17	17.69	28.49	27.63	30.65	25.36	
Data Reduction Increment	**D**			**E**			**F**		
Split #	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
MAE (ppm)	4.29	4.60	4.35	4.55	4.56	4.56	4.61	4.73	4.56
RMSE (ppm)	7.06	7.20	6.56	6.71	7.12	7.09	7.16	6.82	6.96
R^2	0.3867	0.397	0.444	0.4032	0.385	0.388	0.378	0.399	0.415
AIC	1027.9	1025.8	771.81	771.93	771.92	771.92	515.94	515.84	515.88
oMAE (ppm)	29.95	27.55	24.80	22.99	28.57	30.16	26.00	25.05	25.53

Table 4-7 *Summary of individual performance metric scores and mean scores for the T-S fuzzy model using optimal model parameters*

Data Reduction Increment	**A**			**B**			**C**		
Split #	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
MAE Score	6	5	10	6	10	3	7	6	-
RMSE Score	1	9	10	9	10	1	3	8	-
R^2 Score	4	2	9	10	8	1	4	4	-
Mean Score	3.67	5.33	9.67	8.33	9.33	1.67	4.67	6.00	-
Data Reduction Increment	**D**			**E**			**F**		
Split #	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
MAE Score	7	2	6	3	3	3	2	0	3
RMSE Score	4	4	8	7	4	4	4	6	5
R^2 Score	2	3	5	3	2	2	2	3	4
Mean Score	4.33	3.00	6.33	4.33	3.00	3.00	2.67	3.00	4.00

In general the performance of T-S fuzzy model decreases as sparsity of the data increases. Figure 4-19 displays the measured lead concentrations plotted against the predicted lead concentrations from the T-S fuzzy model from all data reduction increments and subsets.

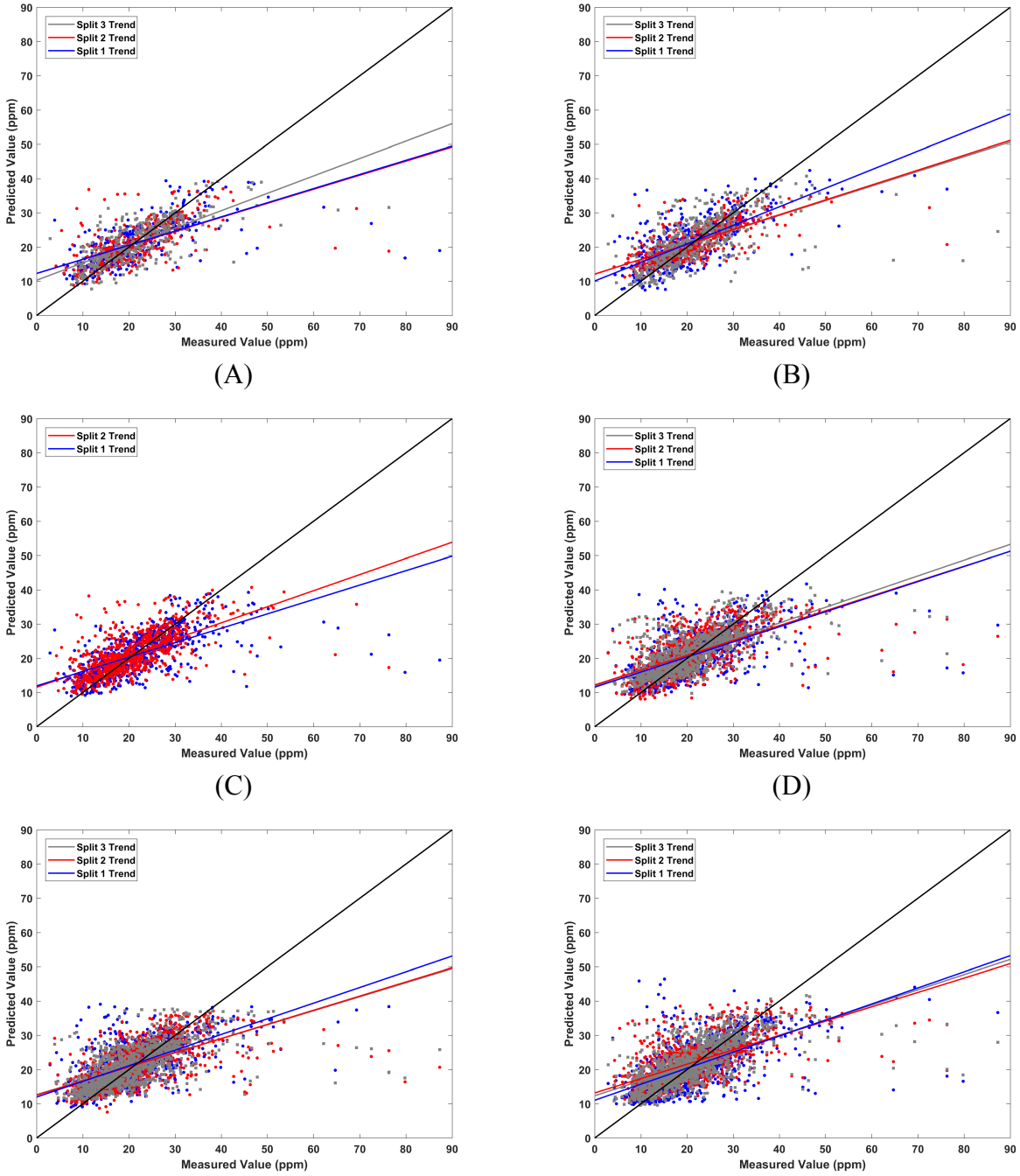


Figure 4-19, T-S fuzzy model performance results for Data increments A-F and all subsets, trend line colour matches the subset split in each plot

As fewer training points are used to make the spatial predictions there is more scatter of points from the 1:1 reference line in each plot and a deflection of the trend lines to a lower slope. This is quantified by the lower R^2 value as the data conditions become sparser. As the amount of data is reduced the results from the three subsets at each increment also become more similar. This is most likely attributable to the larger size of the validation sets at these increments being more representative overall. There are clear outliers in the validation sets in each reduction increment that the model fails to predict. It is possible that these outlier points are occurring at the margins of the data, where the density of membership functions is low. Furthermore, the predictive ability of the model at the margins may be skewing the overall performance results. Figure 4-20 is an east-west cross-section plot of the measured lead concentration (P measured) and the predicted lead concentration (P predicted) for data reduction increment A, subset 1. The second order trend lines on this plot indicate the T-S fuzzy model is performing most adequately in the centre of the map space and confirms that performance near the margins is poor. This is may be the result of low membership function density, however, outlier points at margins of the map space may also contribute to a poorer model performance in this area.

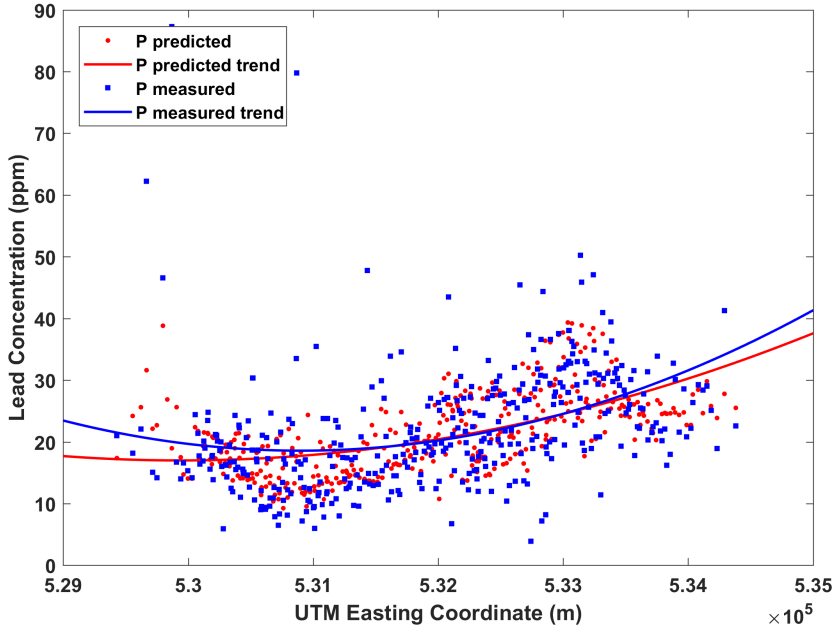


Figure 4-20, East-west (left to right) cross section plot of P measured and P predicted from the T-S fuzzy model for data reduction increment A, subset 1

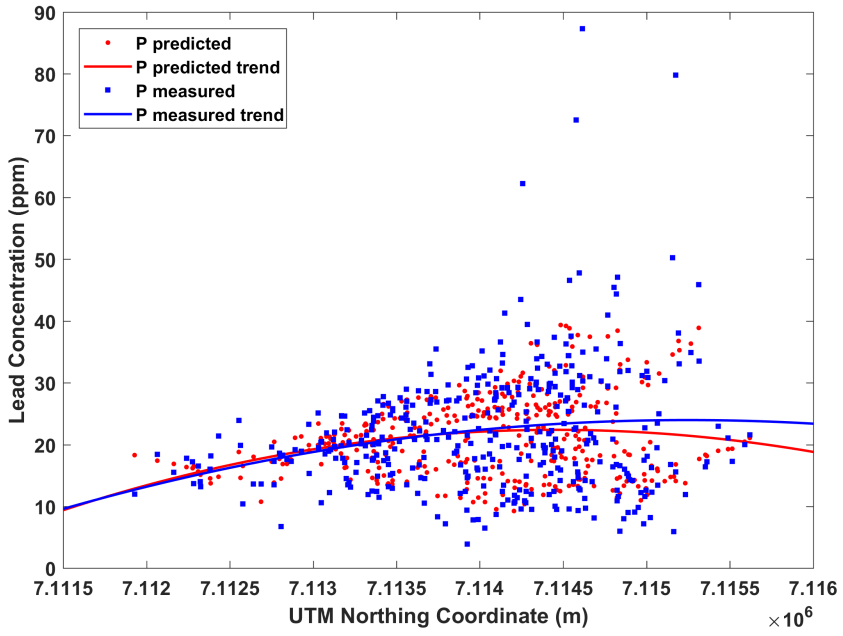


Figure 4-21, South to north (left to right) cross section plot of P measured and P predicted from the T-S fuzzy model for data reduction increment A, subset 1

The north-south cross-section plot for data reduction increment A, subset 1 displays less model break down in the southern region of the map where the range of lead concentrations is much lower (Figure 4-21). However the area in the north where many outlier validation points occur is severely affecting the model result. Therefore, the break down of predictive ability near the margins may be the result of both low membership density and the specific location of outlier points. The model is very clearly predicting values towards the mean in all cases. Additional cross-section plots from the remaining reduction increments are displayed in Appendix B.

The performance of the T-S fuzzy model for the prediction of lead in soil appears to be very poor. However, the kurtosis of the data used for the analysis is very high. Previous studies have shown that spatial interpolators perform poorly on data with more outliers and higher variance (Li & Heap, 2008). To quickly assess if the quality of the data is leading to poor results in this analysis, the T-S fuzzy model was used to predict values from a synthetic data set. The synthetic data used were normally distributed and were extracted from a simple surface created using a MATLAB R2017a (MathWorks, 2017) (Figure 4-22).

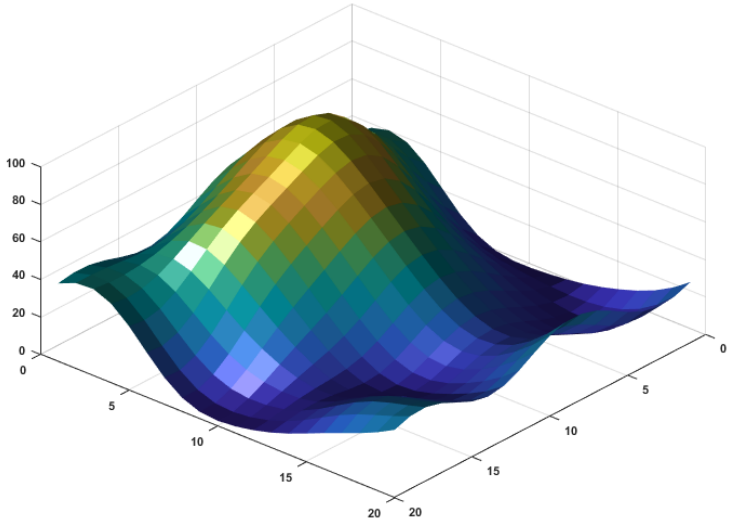


Figure 4-22, Synthetic data surface with arbitrary axis units

The same methodology for selection of a training and validation set was employed and a split of 75% for training and 25% for validation was selected for this example (Figure 4-23). Table 4-8 displays the summary statistics of the synthetic data set, which exhibits a significantly lower kurtosis and skewness than the lead in soil data utilized in this analysis. This data is unitless but was given a range similar to that exhibited by the lead data.

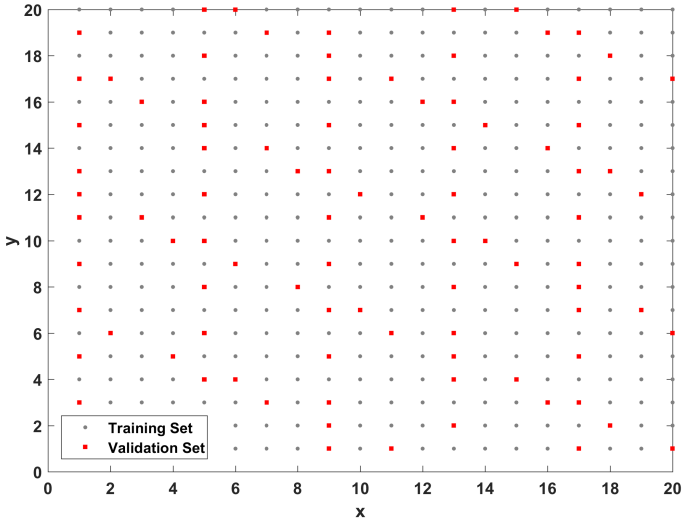


Figure 4-23, Spatial expression of the training and validation sets extracted from the synthetic surface

Table 4-8 *Summary statistics for synthetic data example of the T-S fuzzy method*

Synthetic Data	Range (ppm)	Mean (ppm)	Standard Deviation	Skewness	Kurtosis
Training set (75%)	84.40	34.09	21.60	0.60	2.43
Validation Set (25%)	78.22	34.24	21.77	0.64	2.35

Figure 4-24 displays the measured synthetic data values plotted against model predicted values and Table 4-9 summarize the performance metrics for the synthetic data example.

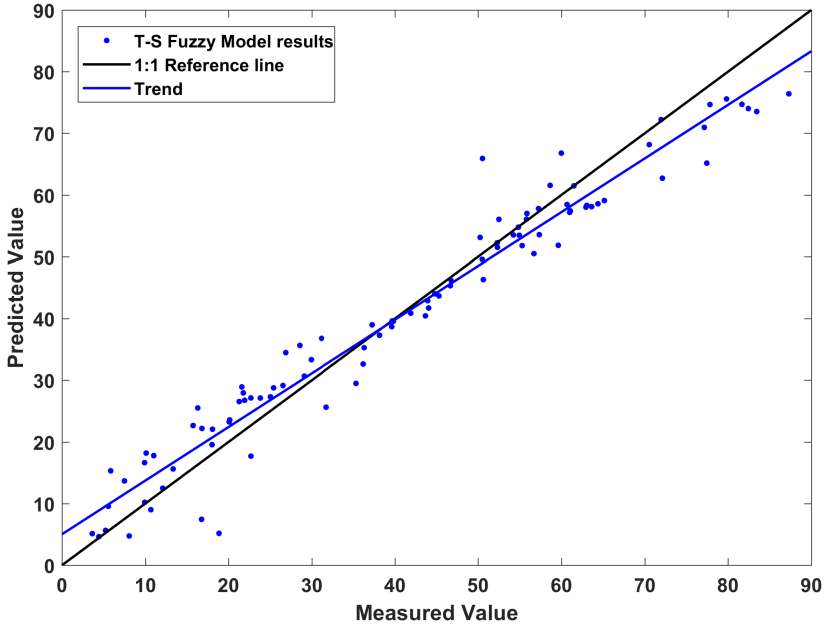


Figure 4-24, Plot of measured values vs predicted values from T-S fuzzy model using synthetic data.

Table 4-9 *Results for the T-S fuzzy model from the unitless synthetic data set*

Performance Metric	**MAE**	**R^2**	**RMSE**
Results	4.09	0.953	5.23

As is clear from the Figure 4-16 and Table 9 the T-S model performs very well using the simplistic synthetic data set. Therefore, the poor performance observed using the lead in soil data is the result of the quality of the data. Since this data represents real world lead concentrations in soil, although difficult to model, is still a good source of data for comparing the T-S fuzzy model, OK, and IDW. Further discussion of performance of the T-S fuzzy model in relation to training and validation sets is reviewed in Section 4.4 during the comparison of the three methods.

4.3 Kriging and IDW Results

The training sets for data reduction increments A-F, subset 1 were exported from MATLAB R2017a and imported into ArcMAP (ESRI, 2011). The Geostatistical Analyst tool was used to perform spatial interpolation using IDW and OK with optimized parameters for each data set. The output from each method and data reduction increment was a complete raster grid of the entire map space, using the default raster pixel size of 14 m^2. This data was then exported as 3 column vectors containing the spatial location and value of prediction from each raster pixel centre in the map space. This data was then imported into MATLAB R2017a; to analyze the performance of the methods, the validation set for each respective training increment was used to identify which points from the ArcGIS raster outputs were validation points, so the predicted concentration at those locations could be compared. This was accomplished using the knnsearch.m function in MATLAB R2017a, this function selected the rows of the ArcGIS output based on there spatial relation to the validation point locations for the respective training – validation sets. Practically, this means that the value of a validation point is compared to the value of the raster pixel it occurs within. Once the validation points from each ArcMap output were determined for each respective increment, the performance metrics for the predictions were calculated and the results subjected to the model scoring system to calculate a mean performance score for OK and IDW for each data reduction increment. Table 4-10 and Table 4-11 display the complete results of the IDW and OK methods respectively.

Table 4-10 *Model results for the IDW spatial predictions*

Data Reduction Increment	MAE (ppm)	R^2	RMSE (ppm)	AIC	oMAE (ppm)	Mean Score
A	4.17	0.442	7.51	2306.0	42.33	5.00
B	4.09	0.529	6.25	2049.7	21.78	9.33
C	4.22	0.432	7.14	1537.9	30.52	5.33
D	4.22	0.4031	6.95	1027.8	29.88	5.33
E	4.48	0.395	6.82	771.84	21.88	4.00
F	4.44	0.4086	6.88	515.86	26.98	4.33

Table 4-11 *Model results for the OK spatial predictions*

Data Reduction Increment	MAE	R^2	RMSE	AIC	oMAE (ppm)	Mean Score
A	4.49	0.392	7.84	2306.1	42.92	2.00
B	4.15	0.516	6.33	2049.7	22.95	9.00
C	4.46	0.384	7.45	1538.0	33.14	2.67
D	4.46	0.362	7.18	1027.9	31.95	3.00
E	4.76	0.339	7.25	771.96	22.74	1.00
F	4.65	0.353	7.34	515.99	25.89	1.33

The effects the data reduction had on both the OK and IDW results are similar to the T-S fuzzy model. That is as the data conditions become sparser the predictive ability of the methods are poorer. The data reduction also had specific implications for the model parameters of OK and IDW (Table 4-12).

Table 4-12 *Results from variography and IDW parameter optimization*

Data Reduction Increment	OK Variography Results				IDW Parameters
	Range (m)	Sill	Nugget	Model Type	Optimized *p*
A	2788.71	51.18	43.56	Stable	1.866
B	397.24	57.59	0	Stable	1.959
C	2709.77	47.31	43.48	Stable	1.691
D	4327.50	83.65	44.69	Stable	1.328
E	131.42	74.88	0.075	Stable	1.977
F	111.54	66.91	0.067	Stable	1.205

For IDW, as the density of data in the map space was reduced the power factor *p* also reduced (Figure 4-25). This is necessary, because as contributing points become more spread out, a lower *p* gives a greater weighting to more distance points. The result of the increased data sparsity on the variography for OK, is that once a certain threshold is reached, insufficient data is present to adequately determine the spatial dependence of the data. At this point the range of spatial dependence drops to be being relatively low (Figure 4-21). This occurs at increment E in the analysis, where 25% of the data is being used for training. For increment E and F the model performance is also considerably lower, indicating that the data density is sufficiently low that variography may be failing.

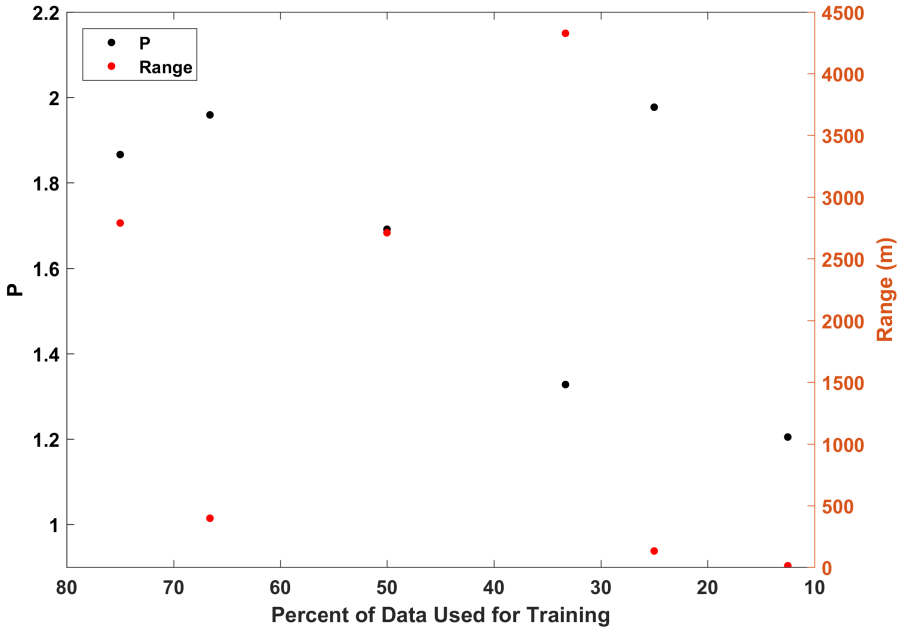


Figure 4-25, Optimal *p* values and ranges for all data reduction increments

From the variography for the all data reduction increments, the optimal model fit to each experimental variogram by the Spatial Analyst Tool was a "stable" model type. The binned variograms from each increment are quite similar, however the binned semivariance values become more scattered as the data become sparser. This is indicative of the changes in range observed. Figure 4-26 displays a comparison the binned experimental semi-variogram from data reduction increment B and F. Most notable is that for increment B, which had the highest model performance, the experimental variogram has very few outlier points, which leads to a more accurate prediction of the range and fitted model. The experimental variogram for increment F displays a much greater scatter, which contributed to the poor model performance at this increment. Binned experimental variograms for the remaining training reduction increments are displayed in Appendix B.

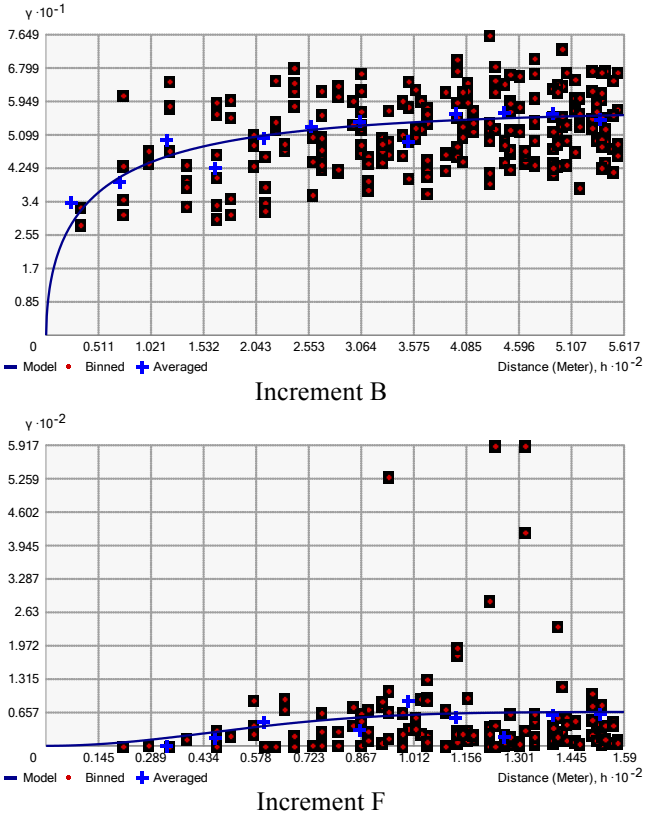


Figure 4-26, Binned experimental variograms from data reduction increments B and F, subset 1

Further discussion of the performance of OK and IDW is reviewed in Section 4.4 during the comparison with the T-S fuzzy model.

4.4 Comparison of the T-S Fuzzy Model, OK, and IDW Under Increasingly Sparse Conditions

Initially it is useful to compare the three methods measured vs. predicted concentration plots for each data reduction increment to make initial comparisons of the performance of the methods. Figure 4-27 displays the results from the increments A-F for the three methods. In all cases there is very minimal difference between the three methods. All three methods under predict the outliers in the validation sets and predict towards the mean of the training sets. For all three methods, data reduction increment B produced the highest performance scores; this is due

to the fact that at this increment subset 1's training set has a much higher range, variance, and kurtosis than the respective validation set. To better assess the overall predictive ability of each method, the mean scores from each data reduction increment for subset 1 from each method were plotted against the amount of data used for training at each increment (Figure 4-28).

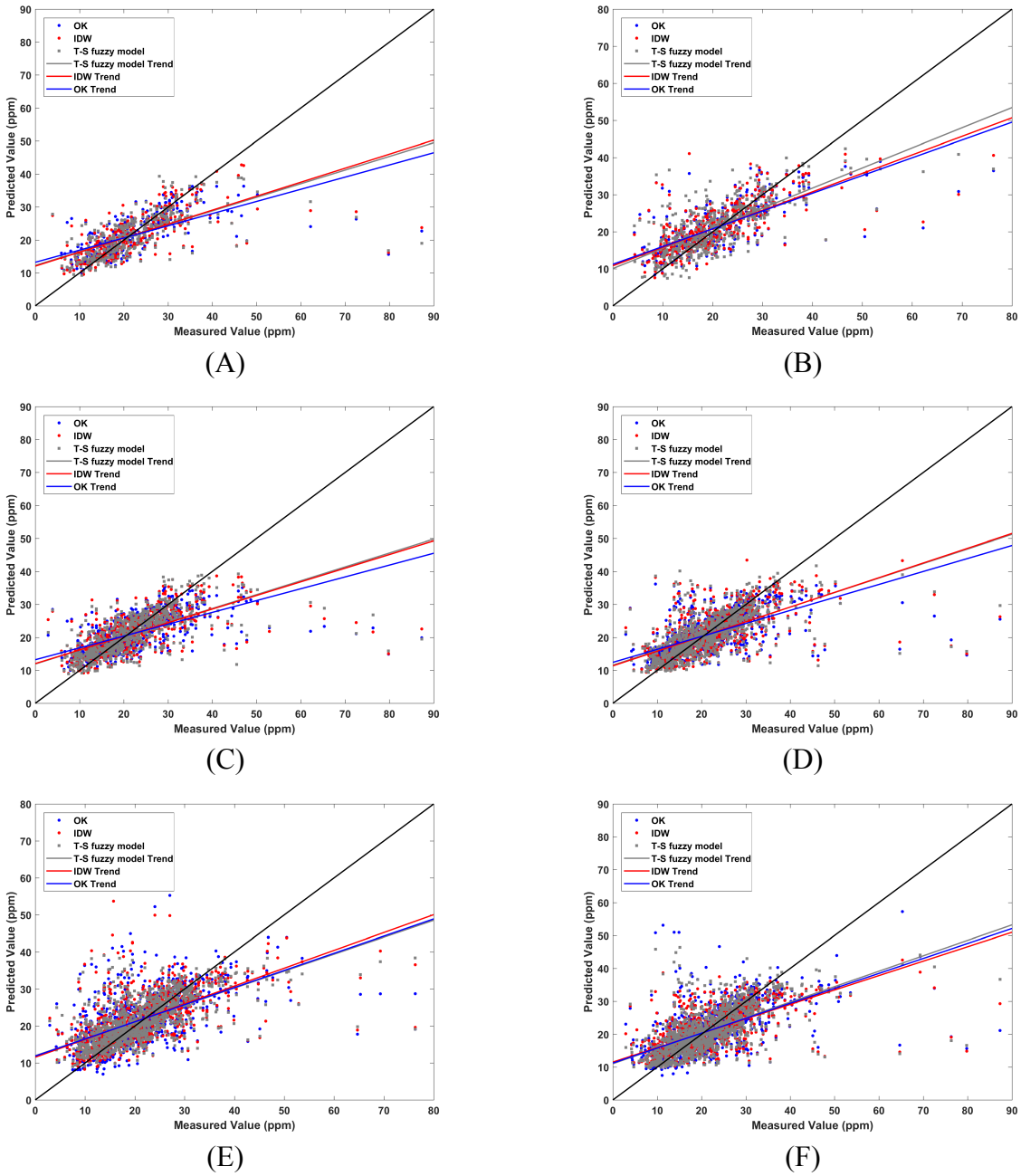


Figure 4-27, Measured vs. predicted lead concentrations for OK, IDW, and the T-S fuzzy model for all data reduction increments, subset 1

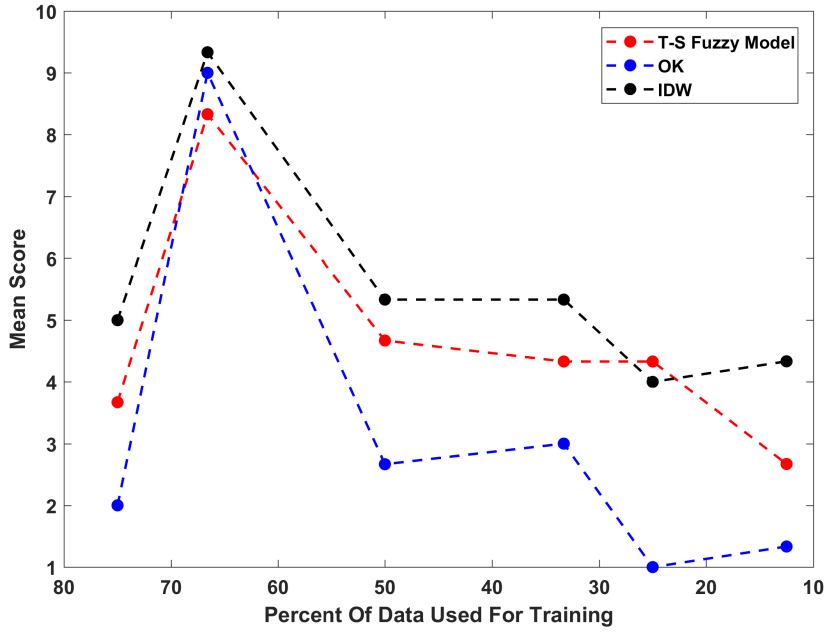


Figure 4-28, Mean scores for OK, IDW, and the T-S fuzzy model from each increment of the data reduction, subset 1

Based on an initial review of the T-S fuzzy models results for the 3 subsets at each data reduction increment, it appears to be sensitive to the variation in the training and validation sets. Therefore, it is useful to additionally plot the highest and lowest observed mean scores for the T-S fuzzy model from the subsets from each increment (Figure 4-29). Since Increment C (50% training, 50% validation) only has two subsets its mid value is displayed as the mean of the score from subsets 1 and 2.

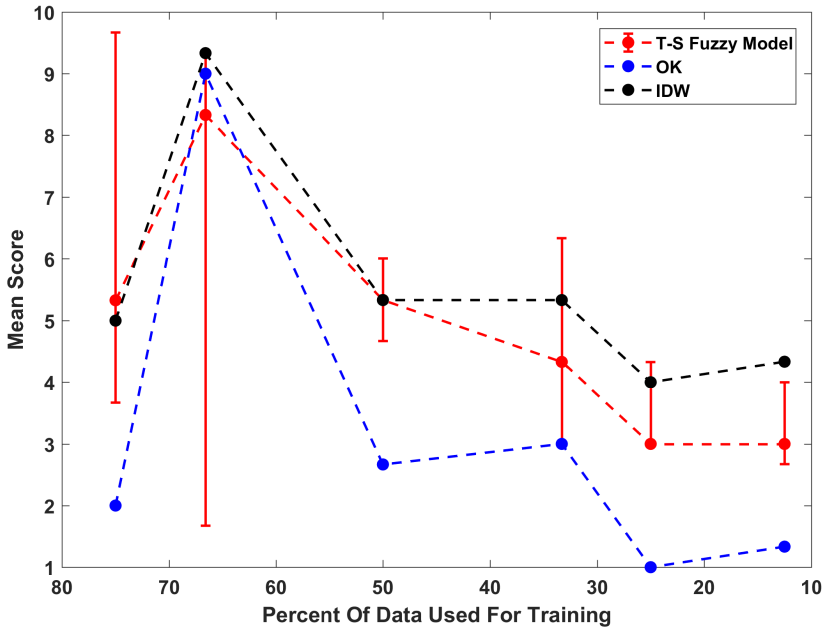


Figure 4-29, Mean scores for OK and IDW from each increment of the data reduction, subset 1 and the mean scores from T-S fuzzy model with error bars indicating the max and min performance from subsets 1-3 for all increments

Overall the three models perform very similarly. For practicality, the analysis was not performed using subsets 2 and 3 for IDW and OK. However, it is hypnotized that if randomly selected training and validation sets were tested at each increment over many iterations the average results from the three models would be close to equal. The discrepancies identified previously between the training and validation sets appear to have large effect on the performance of models. Specifically, between increment A and B, where for increment A the kurtosis and range are higher in the validation set and for increment B the kurtosis and range are higher in the training set. However, subset 3 of increment B, where a very poor performance is exhibited by the T-S fuzzy model the validation set again, has a much higher kurtosis and range. As the size of the validation sets increase the range of predictions decreases. This is thought to be the result of the greater number of points ability to reduce drastic variation between the validation set.

The effect of differences between the training and validation sets does appear to be significant. However, in a real world situation if fewer samples were collected and analyzed the results could have a lower range and kurtosis, so for the purpose simulating spatial data sparsity the training and validation sets used are useful because they illuminate weakness's in the models tested. Specifically, that if the samples fail to capture the overall variance of a system, their spatial predictions will be poor. Overall through out the data reduction increments, on average all the methods model performances decline, which is the desired result of the analysis. For Increments A-F, subset 1 where direct comparisons are available, IDW consistently outperformed the other methods and the T-S fuzzy model minorly outperformed OK. To further quantify these observations the performance of the individual metrics from the optimal T-S fuzzy model results and for OK and IDW from increments A-F for subset 1 can also be plotted against the amount of training data used at each increment (Figures 4-30 to 4-32).

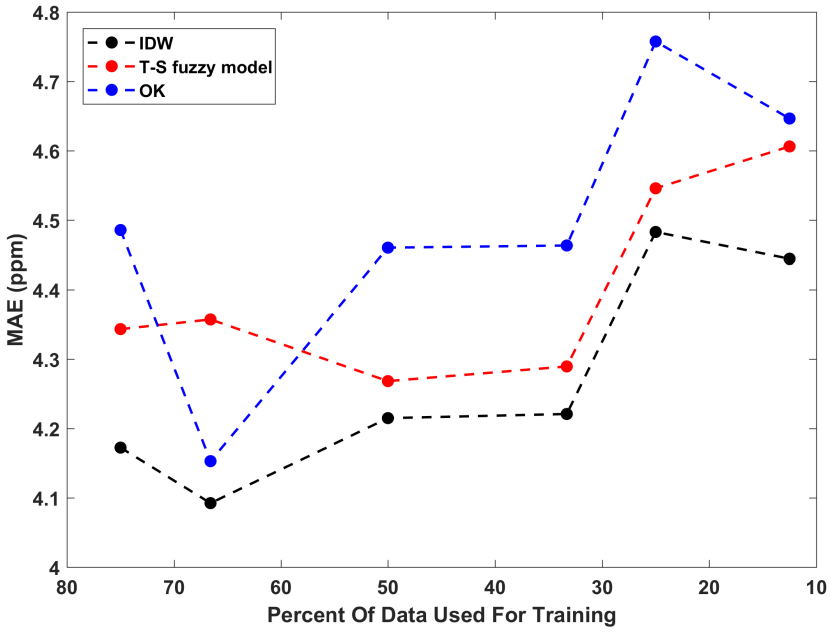


Figure 4-30, MAE calculated for OK, IDW, and the T-S fuzzy model at all data reduction increments for subset 1

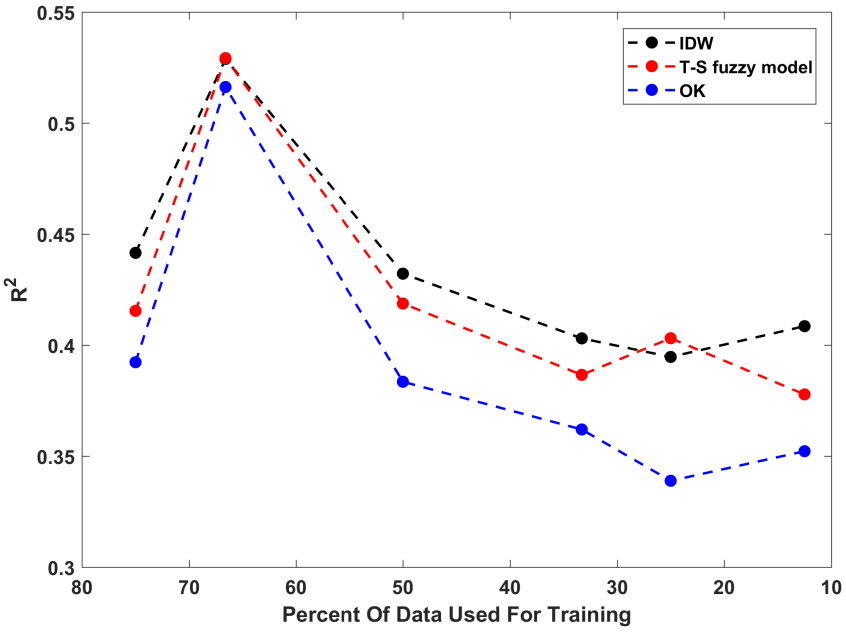


Figure 4-31, R^2 calculated for OK, IDW, and the T-S fuzzy model at all data reduction increments for subset 1

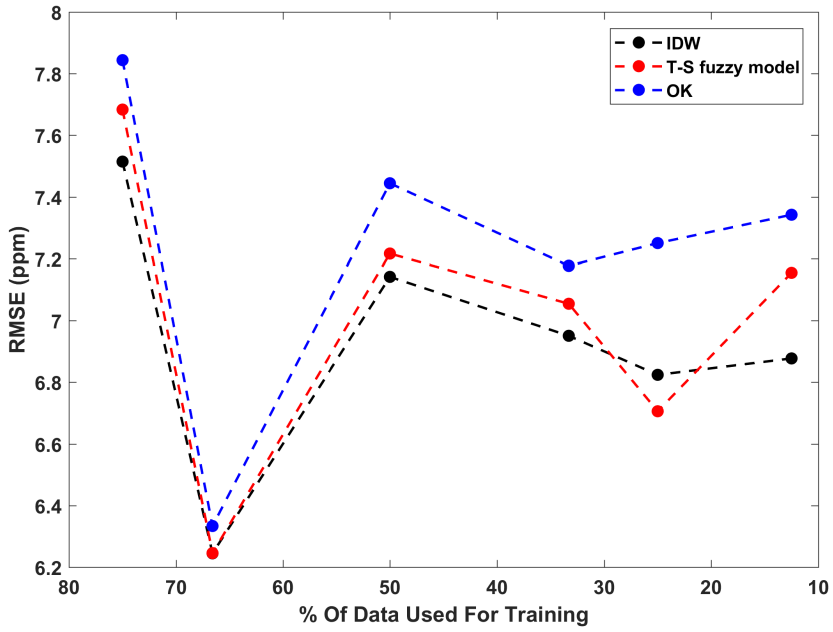


Figure 4-32, RMSE calculated for OK, IDW, and the T-S fuzzy model at all data reduction increments for subset 1

The differences between the three methods are so subtle it is difficult to draw any significant conclusions from the results. As is consistent with the mean scores, based on these performance metrics IDW appears to have produce the best results, with the T-S fuzzy model minorly outperforming OK. This comparison of OK and fuzzy modeling is consistent with other findings in the literature. To further quantify the similar performances of the three methods and account for the differences in validation set length the AIC was plotted for each model, for each increment of the data reduction for training-validation, subset 1 (Figure 4-33), the AIC values are also summarized in Tables 4-6, 4-10, and 4-11.

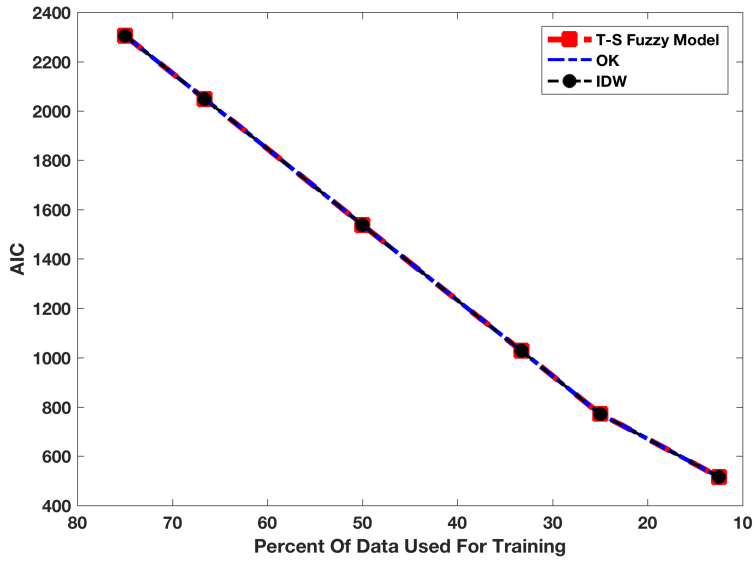


Figure 4-33, AIC values for the T-S fuzzy model, OK, and IDW plotted against the amount of training data used to make spatial predictions

From analyzing the AIC values at each reduction increment for the T-S fuzzy model, OK, and IDW were within 1 point of each other, indicating that there is no significant difference between the model results at each increment (Burnham and Anderson, 2004). The AIC values decrease substantially as the amount of data used for training is reduced and the amount of validation data increases, indicating that changes in the data's spatial density do not greatly affect individual

model performance. Further supporting the results observed from the other performance metrics, that there is no advantage to using this T-S fuzzy model for this data. A common criticism of OK is its tendency to have a significant smoothing affect of the predicted values (Zarco-Perello & Simões, 2017; Goovaerts, 1999). For the lead predictions made in this analysis OK, IDW, and the T-S fuzzy model all appear to have a significant smoothing effect at all data reduction increments (Figure 4-34). However, the quality of the data used may be indicative of this result. Nevertheless, all three methods perform very poorly for the spatial prediction of this lead in soil data.

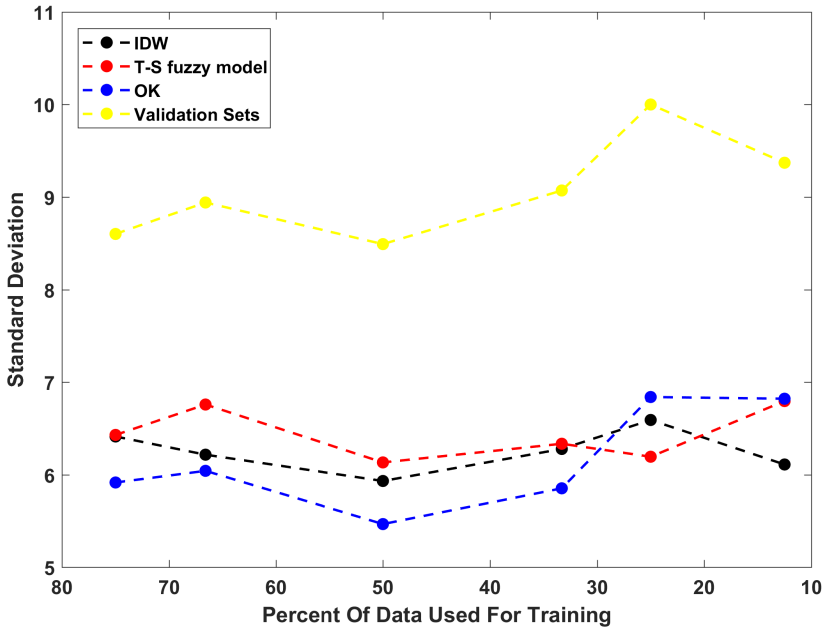


Figure 4-34, Standard deviation of the predictions from OK, IDW, and the T-S fuzzy model and the standard deviation of the measured values from the validation sets at all data reduction increments for subset 1

The ability of spatial interpolators to predict extremely high values within the data is very important (Li & Heap, 2008). To assess the ability of OK, IDW, and the T-S fuzzy model to predict outliers within the validation sets, a separate MAE was calculated for each data reduction

increment for all subsets. Figure 4-35 diplays the oMAE results for OK, IDW, and the T-S fuzzy model for all data reduction increments for subset 1.

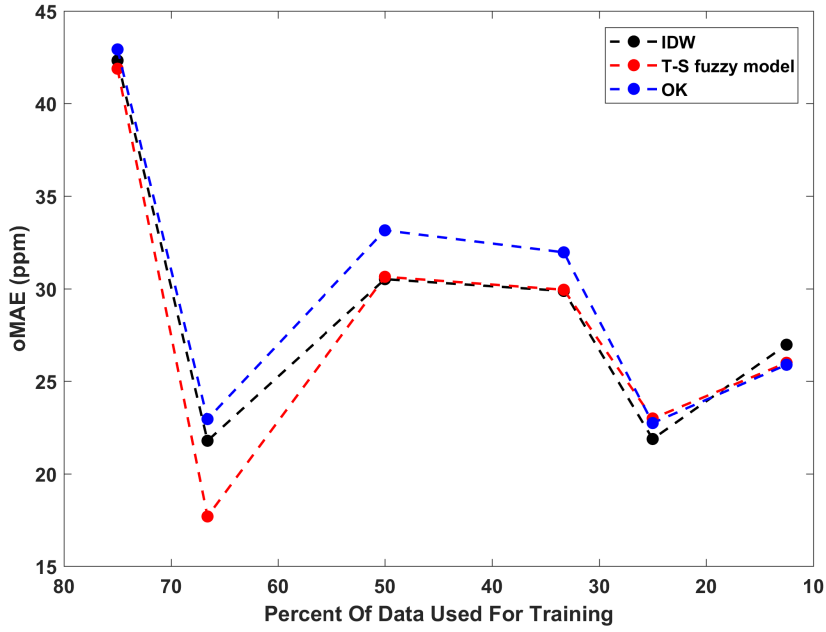


Figure 4-35, oMAE from OK, IDW, and the T-S fuzzy model for all increments of the data reduction for subset 1

The ability of the three methods to predict outlier points within the respective validation sets appears to be very similar. As with all other performance metrics, it appears the performance of the three methods is difficult to separate. The similarity in model performance may be related to the models themselves. Both IDW and OK generate weights for surrounding points based on a similar model shapes. The T-S fuzzy model utilized Gaussian shaped membership functions to determine the weighting for each of the clusters predictions. Although the three models are not using identical profiles in their predictions of unknown points, its possible that the similarity in the results is simply a function of the similarities within the models themselves. However, further research would be required to quantify this. What is clear is that T-S fuzzy model does not

appear to outperform OK and IDW for the spatial predictions of lead in soil under increasingly sparse data conditions. Since fuzzy clustering is at the core of the T-S fuzzy modeling approach, it can be surmised that the performance of the method may be limited by the quality of the fuzzy clusters. Criticisms have been made of the FCM clustering algorithms ability to only extract spherical clusters from with in the data (Muhammad & Glass, 2011). If the spatial trends of lead concentrations within this data set are consistently non-spherical it is possible that FCM was not well suited to clustering this data. The T-S fuzzy model did perform comparably well to OK and IDW under sparse data conditions and through an example with synthetic data appeared to have a high predictive ability. In reality, pollutant concentrations in marine sediment generally occur with much less abrupt transitions with fewer outliers (Kazemi & Hosseini, 2011). Therefore, for the practical application of determining the T-S fuzzy models efficacy for modeling dioxins in marine sediment, the data set used in this analysis may not have been adequate. However, testing the methods using "difficult" to model data does provide some insight into the models weaknesses. Additionally, the data used in this research had a negatively skewed distribution. A factor that was not considered was normalizing the distribution of the data prior to the analysis, which may have impacted the performance of the models. However, for testing the relative performance of T-S fuzzy model against IDW and OK for spatial prediction using incrementally less spatially distributed data, the results are still of merit. Previous research comparing IDW and OK have found that in general the two methods perform very similarly and that IDW often produces a lower error (Qiao et al., 2018; Zarco-Perello & Simões, 2017; Mirzaei & Sakizadeh, 2016). This research supports that conclusion. Furthermore, this research did prove that the T-S fuzzy modeling approach may be a viable spatial prediction tool. If refined into a packaged spatial prediction tool, where optimal model parameters are automatically selected, spatial

predictions could be made with very little modeler input. The data driven nature of the approach would make it advantageous over Kriging, which requires expert modeler knowledge to correctly navigate variography. Therefore, development of a fuzzy clustering based, T-S structured FIS specifically for spatial interpolation may be a worthwhile endeavor. Finally, Li and Heap (2011) reviewed 18 studies that compared different spatial interpolators and concluded that the predictive ability of a spatial interpolator has more to do with the variance of the training data that the spatial density. If the variance of the population is not accounted for in the samples collected, then spatial prediction, regardless of the method used will be poor. Therefore, in practical applications ensuring the variance of the system of interest is captured by the samples is of the upmost importance. However, without a priori knowledge of the system this may be very difficult.

4.5 Future Research

Although the T-S model used in this analysis performed very similarly to the most prevalent methods used for spatial interpolation in the literature, several assumption were made that may have affected the models performance. Future research with a focus on the effect of membership function shape and varied membership function widths may improve the spatial predictive ability of the T-S fuzzy model. The predictive ability of the FIS is heavily governed by the partitioning of the data space into fuzzy regions by the FCM clustering algorithm. Further research should also investigate the effect that different clustering algorithms have on predictive ability. To analyze the quality of the fuzzy clustering methods used, a possible solution may be the introduction of a fuzzy performance metric, which could be used to identify the amount of the validation data captured by the fuzzy clusters above a certain membership as a proxy for how well those clusters are able to model the data. Additionally, future research should consider the

impact of normalization or transformation of data prior to the analysis. Within the FIS the equation that solves the contribution of each cluster is a simple linear equation (Equation 12). It may be possible to achieve a more accurate model prediction using a non-linear approach; since the linear solution for each cluster may underrepresent the inherent complexity of each cluster and lead to under-fitting in the model predictions; future research should investigate this possibility. In practice the T-S fuzzy model would use all the available sample data in a map space to make predictions at every unknown location. To address the issue of over-fitting, a possible hybridized approach may be warranted, where when predicting unknown values at locations proximal to known data points a method be employed that would govern those prediction; something such as IDW may be sufficient to ensure over-fitting does not occur around known sample points. Determining the optimal combination of fuzzy clustering algorithm, membership function shapes and widths, and equations for the FIS may present significant computational challenges, since many variables may be dependent. Therefore, an approach such an ANFIS where each variable is assigned to the node of a neural network, may lead to the most optimal performance. However, care should be taken to ensure transparency, such that the optimum from each variable can be easily confirmed as logical. Based on the AIC values calculated for this research, the reduction of the spatial density of the training data had a negligible effect on model performance. Therefore, future research should attempt a different methodology for simulating data sparsity.

5. Conclusion

Dioxins are a persistent environmental pollutant that occur in marine harbour sediments around the world and are the result of anthropogenic activities. Dioxins have the ability to enter food webs, bio-accumulate in animal tissues, and pose a significant risk to human health. In order to most adequately remediate contaminated sediments, the boundaries of the contamination must first be delineated. In marine environments collecting point samples and performing spatial interpolation accomplishes this. However, the cost of collecting and analyzing enough samples to produce an accurate pollutant distribution map may be cost inhibitive for certain remediation projects. The ability to collect and analyze fewer samples, yet still produce an adequate pollutant distribution map, would reduce the initial cost of new remediation projects and may lead to more projects being started.

Fuzzy Set Theory has been shown as a way to reduce uncertainty do to data sparsity and provides a convenient way to mathematically quantify gradational changes using membership functions. This is advantageous for modeling spatial phenomena like pollutant concentrations that do not have crisp boundaries, but gradationally change between areas with different pollutant concentrations. Fuzzy modelling takes advantage of Fuzzy Set Theories ability to quantify gradational changes, and uses fuzzy clustering to give each data point membership to each of the theoretical clusters in the data. This additional information may help fuzzy modeling produce more accurate spatial predictions. Fuzzy modeling has been used to make many types of spatial predictions in the geosciences, however limited research has been conducted in determining its efficacy for spatial modeling under sparse spatial data conditions. To determine if fuzzy modeling is a suitable method for spatial prediction under sparse data conditions, a data set of spatially distributed lead in soil concentrations was utilized. The data was collected in Yukon

Territory for the purpose of mineral exploration, it is useful because it presents the concentrations of several environmentally sensitive pollutants in point samples spatially distributed over a large area. Lead was selected due to its high rank as a pollutant of concern in urban settings. By assessing the predictive ability of fuzzy modeling under sparse spatial data conditions using lead concentrations, inferences can be made about its possible applicability for modeling dioxins in marine sediment. Specifically, this research used a T-S fuzzy modelling approach to make spatial predictions and the spatial density of the data used to make the predictions was incrementally reduced to simulate increasingly sparse spatial data conditions. This was intended to help in determining if the T-S fuzzy modeling approach has the ability to produce an accurate pollutant distribution map using fewer data points. For context, the performance of the T-S fuzzy model was compared to the traditional spatial interpolation methods IDW and OK during the analysis. The T-S fuzzy model used the FCM clustering algorithm to partition the data into less complex fuzzy regions. The fuzzy regions were represented by Gaussian shaped membership functions in an FIS, which used geographic map coordinates as inputs to produce pollutant concentrations as outputs. Prior to the spatial predictions being made, the data was systematically separated into training and validation sets for six separate data reduction increments. The data reduction increments utilized 75%, 66.6%, 50%, 33.3%, 25%, and 12.5% of the overall data respectively. At each increment the data not used for prediction was used for validation. To ensure the observations about the T-S fuzzy model were robust, at each data reduction increment three subsets were systematically selected to determine the T-S fuzzy models sensitivity to data variation at the same increment. To assess the predictive ability of the three methods MAE, R^2, RMSE, and AIC were chosen as performance metrics. To give each metric equal weighting in the assessment, a binned scoring system was

developed to best assess the relative performace of the spatial predictions. Each metric was rated with a score between 1 and 10, based on its performance; the average score from MAE, R^2, and RMSE represented each models final score. AIC was not included in the scoring system as the AIC scores from each of the methods were nearly identical. The scoring system employed was overly sensitive and only provided relative comparisons between the models. Initially, the parameters for the T-S fuzzy model were determined based on the ability of the model to predict the validation sets from the respective training sets. Specifically, the parameters that were determined included: number of clusters, fuzzy overlap between clusters, and membership function width. The fuzziness was the most sensitive parameter and had the greatest effect on the models performance, especially as data conditions become sparser. The number of clusters appeared to be less critical, but in each case a number of clusters was observed that produced an optimal score for predicting the validation sets. The membership function widths for the FIS had a much smaller impact on performance, a range between 80 m and 150 m appeared best, however all the membership function widths tested produced very similar results. Selecting the correct model parameters was important for determining the best possible T-S fuzzy model performance for the data used in this study. However, it was determined that certain ranges and combinations of parameters produced adequate predictions, indicating that model parameter selection is not the most critical part of fuzzy modeling. Initially, the optimal number of clusters, fuzziness, and membership function widths were determined for each of the six data reduction increments and their subsets. Once the combinations that yielded the most accurate predictions of the validation sets were identified, the performance of the T-S fuzzy model was assessed. It was observed that the T-S fuzzy model is very sensitive to data quality. Differences in the spatial density, variance, and kurtosis have a significant impact on the models performance. This was

quantified by comparing the results of three subsets at each data reduction increment. To determine the efficacy of the T-S fuzzy model under the increasingly sparse spatial data conditions, its results were compared to those from OK and IDW for the same data sets. Because OK and IDW are well-established methods and require significant modeler input, only results from the first subset of each data reduction increment was used for comparison. The three methods performed very similarly at all increments of the data reduction. The three models performances for the individual performance metrics were also very similar, including their ability to predict outliers in the validation data sets. All three methods exhibited a considerable smoothing effect in their predictions, consistently predicting lead concentrations towards the mean. It was determined that the data used in this study was difficult to model based on the high number of outliers in the data set. This was quantified by testing the predictive ability of the T-S fuzzy model on a smooth synthetic data surface. When modelling this data, which had a much lower kurtosis, the T-S fuzzy model performed well. Additionally, the AIC for each method was compared for each increment of the data reduction. The AIC values at each increment for the T-S fuzzy model, OK, and IDW were within 1 point of each other at indicating that there is no significant difference between the model results (Burnham and Anderson, 2004). The AIC values decrease substantially as the amount of data used for training is reduced and the amount of validation data increases, indicating that changes in the data's spatial density do not greatly effect model performance. Based on the statistical distribution of the data set used in this analysis and the similar results of the three methods, its difficult to draw conclusions about the efficacy of the T-S fuzzy model for modeling dioxins in marine sediment. Although, the results from this research, based on a suite of performance metrics tested at varying levels of data sparsity found that the performance of IDW was slightly better the T-S fuzzy model, which was slightly better

than OK. The performance results for OK, IDW, and the T-S fuzzy model were however, incredibly similar, which may have been a product of the similarities in the models themselves. The impact that variation in the training sets had on the predictive ability of the T-S fuzzy model does indicate that if used in situations where less data is available, the results would be less reliable, as by using less data it becomes difficult to capture to total variation of the system being modeled. This indicates that for the area where the data was sampled in Yukon Territory, collecting a sufficient number of samples as to adequately capture the spatial variance of metal concentrations will be critical to the success of spatial modeling for data collected there.

The T-S fuzzy modelling approach is a flexible and transparent data-driven modelling technique. Spatial predictions made using the T-S fuzzy model in this research were comparable to the most prevalent spatial interpolation methods used in the literature. The slight outperformance of OK by fuzzy modeling follows results from other reported research in the literature. However, the fuzzy modeling approach did not clearly outperform the traditional spatial interpolation methods under increasingly sparse spatial data conditions. For the data used in this study the predictive abilities of the models tested was very poor. Further, the AIC values calculated at each data reduction increment indicated that changing the spatial density of the data did not have an appreciable effect on model performance. Therefore, it is not possible to speculate on possible changes to sampling density based on using a fuzzy modeling spatial interpolation approach or its possible efficacy for modeling dioxins in marine sediment. Although, the data used in this analysis was difficult to model spatially, it still provided a good source for the relative comparison of OK, IDW, and T-S fuzzy modeling. For the pure simplicity, interpretability, and minimal computational expense, IDW may be the most adequate spatial interpolation tool for the initial stages of a remediation project. Although it may be

possible to produce a marginally more accurate result using a different spatial interpolation technique or a further optimized T-S fuzzy model, for the modeler input required, IDW may simply be best.

Bibliography

Amini, M., Afyuni, M., Fathianpour, N., Khademi, H., & Flühler, H. (2005). Continuous soil pollution mapping using fuzzy logic and spatial interpolation. *Geoderma*, *124*(3–4), 223–233. https://doi.org/10.1016/j.geoderma.2004.05.009

Bárdossy, G., & Fodor, J. (2001). Traditional and New Ways to Handle Uncertainty in Geology. *Natural Resources Research*, *10*(3), 179–187. https://doi.org/10.1023/A:1012513107364

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Burrough, P. A. (1989). Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*, *40*(3), 477–492. https://doi.org/10.1111/j.1365-2389.1989.tb01290.x

Burrough, P. A., & McDonnell, R. (1998). *Principles of Geographical Information Systems*. Oxford University Press. Retrieved from https://books.google.ca/books?id=shHznQEACAAJ

Collazo-Cuevas, J. I., Aceves-Fernandez, M. A., Gorrostieta-Hurtado, E., Pedraza-Ortega, J. C., Sotomayor-Olmedo, A., & Delgado-Rosas, M. (2010). Comparison between Fuzzy C-means clustering and Fuzzy Clustering Subtractive in urban air pollution. In *2010 20th International Conference on Electronics Communications and Computers (CONIELECOMP)* (pp. 174–179). https://doi.org/10.1109/CONIELECOMP.2010.5440772

Dag, A., & Mert, B. A. (2008). Evaluating Thickness of Bauxite Deposit Using Indicator Geostatistics and Fuzzy Estimation. *Resource Geology*, *58*(2), 188–195. https://doi.org/10.1111/j.1751-3928.2008.00055.x

Gedeon, T. D., Wong, K. W., Wong, P. M., & Huang, Y. (2003). Spatial Interpolation Using Fuzzy Reasoning. *Transactions in GIS*, *7*(1), 55–66. https://doi.org/10.1111/1467-9671.00129

Goovaerts, P. (1999). Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, *89*(1–2), 1–45. https://doi.org/10.1016/S0016-7061(98)00078-0

Hites, R. A. (2011). Dioxins: An Overview and History †. *Environmental Science & Technology*, *45*(1), 16–20. https://doi.org/10.1021/es1013664

Hwang, S., & Thill, J.-C. (2005). Modeling Localities with Fuzzy Sets and GIS. In *Fuzzy Modeling with Spatial Information for Geographic Problems* (pp. 71–104). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-26886-3_4

Jang, J. S. R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(3), 665–685. https://doi.org/10.1109/21.256541

Kajornrit, J., & Wong, K. W. (2013). Cluster validation methods for localization of spatial rainfall data in the northeast region of Thailand. In *2013 International Conference on Machine Learning and Cybernetics* (Vol. 4, pp. 1637–1642). https://doi.org/10.1109/ICMLC.2013.6890861

Kajornrit, J., Wong, K. W., & Fung, C. C. (2016). An interpretable fuzzy monthly rainfall spatial interpolation system for the construction of aerial rainfall maps. *Soft Computing*, *20*(12), 4631–4643. https://doi.org/10.1007/s00500-014-1456-9

Kazemi, S. M., & Hosseini, S. M. (2011). Comparison of spatial interpolation methods for estimating heavy metals in sediments of Caspian Sea. *Expert Systems with Applications*, *38*(3), 1632–1649. https://doi.org/10.1016/j.eswa.2010.07.085

Khan, U. T. (2015). *Environmental prediction and risk analysis using fuzzy numbers and data-driven models* (Thesis). Retrieved from https://dspace.library.uvic.ca//handle/1828/6937

Khan, U. T., & Valeo, C. (2015). A new fuzzy linear regression approach for dissolved oxygen prediction. *Hydrological Sciences Journal*, *60*(6), 1096–1119. https://doi.org/10.1080/02626667.2014.900558

Kord, M., & Asghari Moghaddam, A. (2014). Spatial analysis of Ardabil plain aquifer potable groundwater using fuzzy logic. *Journal of King Saud University - Science*, *26*(2), 129–140. https://doi.org/10.1016/j.jksus.2013.09.004

Kruse, R., Döring, C., & Lesot, M.-J. (2007). Fundamentals of Fuzzy Clustering. In J. V. de Oliveira & W. Pedrycz (Eds.), *Advances in Fuzzy Clustering and its Applications* (pp. 1–30). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470061190.ch1

Kulkarni, P. S., Crespo, J. G., & Afonso, C. A. M. (2008). Dioxins sources and current remediation technologies — A review. *Environment International*, *34*(1), 139–153. https://doi.org/10.1016/j.envint.2007.07.009

Li, J., & Heap, A. D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, *6*(3–4), 228–241. https://doi.org/10.1016/j.ecoinf.2010.12.003

Liao, Y., Li, D., & Zhang, N. (2018). Comparison of interpolation models for estimating heavy metals in soils under various spatial characteristics and sampling methods. *Transactions in GIS*, *22*(2), 409–434. https://doi.org/10.1111/tgis.12319

MAMDANI, E. H., & ASSILIAN, S. (1999). An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. *International Journal of Human-Computer Studies*, *51*(2), 135–147. https://doi.org/10.1006/ijhc.1973.0303

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

McBratney, A. B., & Odeh, I. O. A. (1997). Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, *77*(2), 85–113. https://doi.org/10.1016/S0016-7061(97)00017-7

Mirzaei, R., & Sakizadeh, M. (2016). Comparison of interpolation methods for the estimation of groundwater contamination in Andimeshk-Shush Plain, Southwest of Iran. *Environmental Science and Pollution Research*, *23*(3), 2758–2769. https://doi.org/10.1007/s11356-015-5507-2

Mitrou, P. I., Dimitriadis, G., & Raptis, S. A. (2001). Toxic effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin and related compounds. *European Journal of Internal Medicine*, *12*(5), 406–411. https://doi.org/10.1016/S0953-6205(01)00146-7

Mousavi, S. R., Sarmadian, F., Dehghani, S., Sadikhani, M. R., & Taati, A. (2017). Evaluating inverse distance weighting and kriging methods in estimation of some physical and chemical properties of soil in Qazvin Plain. *Eurasian Journal of Soil Science*, *6*(4), 327–336.

Muhammad, K., & Glass, H. J. (2011). Modelling Short-Scale Variability and Uncertainty During Mineral Resource Estimation Using a Novel Fuzzy Estimation Technique. *Geostandards and Geoanalytical Research*, *35*(3), 369–385. https://doi.org/10.1111/j.1751-908X.2010.00051.x

Nayak, P. C., & Sudheer, K. P. (2008). Fuzzy model identification based on cluster estimation for reservoir inflow forecasting. *Hydrological Processes*, *22*(6), 827–841. https://doi.org/10.1002/hyp.6644

Nourzadeh, M., Hashemy, S. M., Rodriguez Martin, J. A., Bahrami, H. A., & Moshashaei, S. (2013). Using fuzzy clustering algorithms to describe the distribution of trace elements in arable calcareous soils in northwest Iran. *Archives of Agronomy and Soil Science*, *59*(3), 435–448. https://doi.org/10.1080/03650340.2011.636356

Pedrycz, W., & Izakian, H. (2014). Cluster-Centric Fuzzy Modeling. *IEEE Transactions on Fuzzy Systems*, *22*(6), 1585–1597. https://doi.org/10.1109/TFUZZ.2014.2300134

Pham, T. D. (1997). Grade estimation using fuzzy- set algorithms. *Mathematical Geology*, *29*(2), 291–305. https://doi.org/10.1007/BF02769634

Qiao, P., Lei, M., Yang, S., Yang, J., Guo, G., & Zhou, X. (2018). Comparing ordinary kriging and inverse distance weighting for soil as pollution in Beijing. *Environmental Science and Pollution Research*, 1–12. https://doi.org/10.1007/s11356-018-1552-y

Setnes, M., Babuska, R., & Verbruggen, H. B. (1998). Rule-based modeling: precision and transparency. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *28*(1), 165–169. https://doi.org/10.1109/5326.661100

Shahbeik, S., Afzal, P., Moarefvand, P., & Qumarsy, M. (2014). Comparison between ordinary kriging (OK) and inverse distance weighted (IDW) based on estimation error. Case study: Dardevey iron ore deposit, NE Iran. *Arabian Journal of Geosciences*, *7*(9), 3693–3704. https://doi.org/10.1007/s12517-013-0978-2

Shepard, D. (1968). A Two-dimensional Interpolation Function for Irregularly-spaced Data. In *Proceedings of the 1968 23rd ACM National Conference* (pp. 517–524). New York, NY, USA: ACM. https://doi.org/10.1145/800186.810616

Sonmez, H., Gokceoglu, C., & Ulusay, R. (2004). a mamdani fuzzy inference system for the geological strength index (gsi) and its use in slope stability assessments. *International Journal of Rock Mechanics and Mining Sciences*, *41*(3), 513–514. https://doi.org/10.1016/j.ijrmms.2003.12.092

Stahl, K., Moore, R. D., Floyer, J. A., Asplin, M. G., & McKendry, I. G. (2006). Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agricultural and Forest Meteorology*, *139*(3–4), 224–236. https://doi.org/10.1016/j.agrformet.2006.07.004

Sugeno, M., & Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, *1*(1), 7-. https://doi.org/10.1109/TFUZZ.1993.390281

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(1), 116–132. https://doi.org/10.1109/TSMC.1985.6313399

Travis, C. C., & Hattemer-Frey, H. A. (1991). Human exposure to dioxin. *Science of The Total Environment*, *104*(1), 97–127. https://doi.org/10.1016/0048-9697(91)90010-C

Tutmez, B., & Dag, A. (2007). Use of Fuzzy Logic in Lignite Inventory Estimation. *Energy Sources, Part B: Economics, Planning, and Policy*, *2*(1), 93–103. https://doi.org/10.1080/15567240600629302

Tutmez, B., & Hatipoglu, Z. (2007). Spatial estimation model of porosity. *Computers & Geosciences*, *33*(4), 465–475. https://doi.org/10.1016/j.cageo.2006.07.008

Tutmez, B., & Hatipoglu, Z. (2010). Comparing two data driven interpolation methods for modeling nitrate distribution in aquifer. *Ecological Informatics*, *5*(4), 311–315. https://doi.org/10.1016/j.ecoinf.2009.08.001

Tutmez, B., & Tercan, A. E. (2007). Spatial estimation of some mechanical properties of rocks by fuzzy modelling. *Computers and Geotechnics*, *34*(1), 10–18. https://doi.org/10.1016/j.compgeo.2006.09.005

Tutmez, B., Tercan, A. E., & Kaymak, U. (2007). Fuzzy Modeling for Reserve Estimation Based on Spatial Variability. *Mathematical Geology*, *39*(1), 87. https://doi.org/10.1007/s11004-006-9066-4

Willmott, C. J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science*, *20*(1), 89–102. https://doi.org/10.1080/13658810500286976

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zarco-Perello, S., & Simões, N. (2017). Ordinary kriging vs inverse distance weighting: spatial interpolation of the sessile community of Madagascar reef, Gulf of Mexico. *PeerJ*, *5*. https://doi.org/10.7717/peerj.4078

Zhou, S.-M., & Gan, J. Q. (2008). Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets and Systems*, *159*(23), 3091–3131. https://doi.org/10.1016/j.fss.2008.05.016

Appendix A

Matlab code used for execution of the T-S Fuzzy Model

```
function [C_n,U_d,R2,RMSE,MAE,oMAE,P_out] = TSFUZZ_FCM(XYZ_data_training,XYZ_data_val,num_clusts,m,MFwidth)
%C_n = Cluster centre matrix
%U_d = membership matrix
%P_out = predicted value vector same length with corresponding rows to
%XYZ_data val
%XYZ_data_training = training data XYZ columns
%XYZ_data_val = validation data XYZ columns
%num_clusts = number of clusters
%m = fuzziness
%MFwidth = membership function width
options = [m NaN 1e-5 0];
%Run fcm clustering
[C_n,U_d] = fcm(XYZ_data_training,num_clusts,options);
% Determine linear solution for each cluster via least squares regression
ws=warning('off','all'); %error warning matrix close to singular, this is what we want, suppress warning for speed
X_e = ones(length(XYZ_data_training),3);
X_e(:,1:2) = XYZ_data_training(:,1:2);
for i=1:length(C_n)
gamma(:,:,i) = diag(U_d(i,:));
end
z = XYZ_data_training(:,3);
% solve least squares regression parameters for each cluster
for i=1:length(C_n)
coefs(:,:,i) = (inv(X_e'*gamma(:,:,i)*X_e))*X_e'*gamma(:,:,i)*z;
end
%for plotting membership functions
%figure
%x_plot = min(XYZ_data_training(:,1)):1:max(XYZ_data_training(:,1));
%for k=1:length(C_n)
% y(:,:,k) = gaussmf(x_plot,[MFwidth C_n(k,1)]); %clust_std_x(k)
% plot(x_plot,y(:,:,k),'LineWidth',1)
% xlabel('UTM Easting Coordinate (m)')
% ylabel('Membership to Cluster Centre (\alpha)')
% set(gca,'fontsize',12,'fontweight','bold')
% set(gcf,'PaperType','a5','PaperUnits','centimeters','PaperPosition',[0 0 21 14.8],'PaperOrientation','portrait')
% hold on
%end
%hold off
%y_plot = sort(XYZ_data_training(:,2));
%y_plot = min(XYZ_data_training(:,2)):1:max(XYZ_data_training(:,2));
%figure
%for i=1:length(C_n)
% yy(:,:,i) = gaussmf(y_plot,[MFwidth C_n(i,2)]);
% plot(y_plot,yy(:,:,i),'LineWidth',1.5)
% xlabel('UTM Northing Coordinate (m)')
% ylabel('Membership to Cluster Centre (\alpha)')
%set(gca,'fontsize',12,'fontweight','bold')
%set(gcf,'PaperType','a5','PaperUnits','centimeters','PaperPosition',[0 0 21 14.8],'PaperOrientation','portrait')
%hold on
%end
warning(ws) %turn warnings back on
%%
for validation points - solve there membership to each cluster based on
%their map coordinate location
for k = 1:length(XYZ_data_val) %for each validation point
for i=1:length(C_n) % there is an output from each cluster
Beta_X(i,:,k) = gaussmf(XYZ_data_val(k,1),[MFwidth C_n(i,1)]);
Beta_Y(i,:,k) = gaussmf(XYZ_data_val(k,2),[MFwidth C_n(i,2)]);
end
```

```
end
% solve the rule output for each cluster
for k=1:length(XYZ_data_val) %for each point in the validation set
for i=1:length(C_n(:,1)) % solve output from each cluster
rule_out(i,:,k) = (coefs(1,:,i)*XYZ_data_val(k,1))+(coefs(2,:,i)*XYZ_data_val(k,2))+coefs(3,:,i);
end
end
Beta = [Beta_X.*Beta_Y]; %Beta = degree of activation of rule
% perform rule agregation using all rules
for k = 1:length(XYZ_data_val)
for i = 1:length(C_n)
P_out_top(i,:,k) = (Beta(i,:,k)*rule_out(i,:,k));
P_out_bottom(k) = sum(Beta(:,:,k));
end
end
%solve the final P_out vector, which can be compared to the validation data
%vector for comparison.
for k = 1:length(XYZ_data_val)
P_out(k) = sum(P_out_top(:,:,k))/P_out_bottom(k);
end
P_out = P_out';
%determine which point in validation set are outliers
KK = isoutlier(XYZ_data_val(:,3));
OL_P = P_out(KK);
OL_V = XYZ_data_val(KK,3);
%calculate performance metrics
oMAE = sum(abs(OL_V-OL_P))/length(OL_P)
R2 = (corr(XYZ_data_val(:,3),P_out))^2;
MAE = (1/length(P_out))*sum(abs(XYZ_data_val(:,3)-P_out));
RMSE = sqrt((1/length(P_out))*sum((XYZ_data_val(:,3)-P_out).^2));
End
```

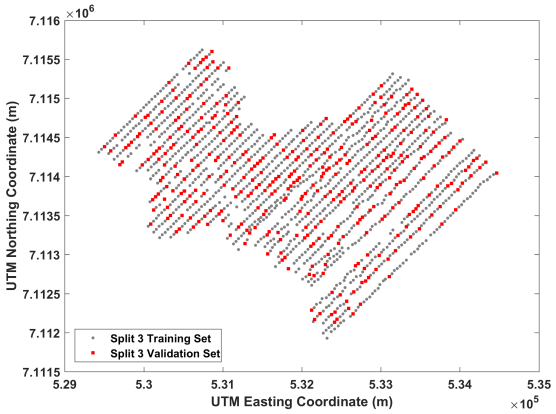
Code for Determination of optimal model parameters

```
XYZ_data_training = XYZ; %Trainging set
XYZ_data_val = XYZval; % Validation set

fuzz = [1.1:0.1:2]; %range of m
width = [50:10:200]; %range of widths for sigma


for i=1:30 %number of clusters
for kk = 1:length(fuzz)
   m = fuzz(kk);

for j = 1:length(width)

num_clusts = i+4;
MFwidth = width(j);
rng(0) %reset the random number generator each loop for consistency
% calls TS_GK, the function that is the fuzzy model
[C_n,MAE_FCM,R2_FCM,RMSE_FCM,var_pred,P_out] =
TSFUZZ_FCM(XYZ_data_training,XYZ_data_val,num_clusts,m,MFwidth);
%C_n is cluster centre locations
%P_out is predicted pollunant vector
%Record the output of each performance metric

FuzzMP(i).out(j,1,kk) = R2_FCM;
FuzzMP(i).out(j,2,kk) = MAE_FCM;
FuzzMP(i).out(j,3,kk) = RMSE_FCM;
FuzzMP(i).out(j,5,kk) = var_pred; % variance of prediction
FuzzMP(i).out(j,7,kk) = fuzz(kk);
```

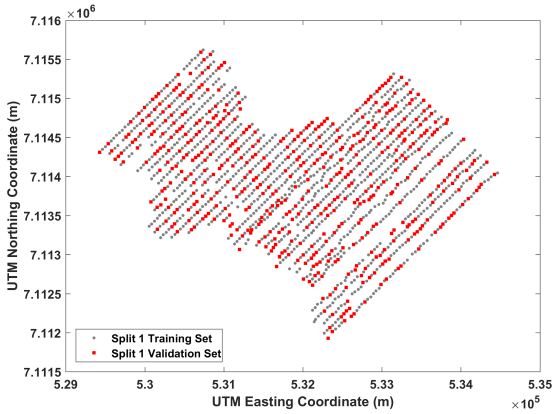
```
FuzzMP(i).out(j,8,kk) = num_clusts;
FuzzMP(i).out(j,9,kk) = width(j);

end
end
end
%save('FuzzMP','FuzzMP') %save outputs to determine which combination of
%parameters produces the most accurate prediction.
```

Appendix B

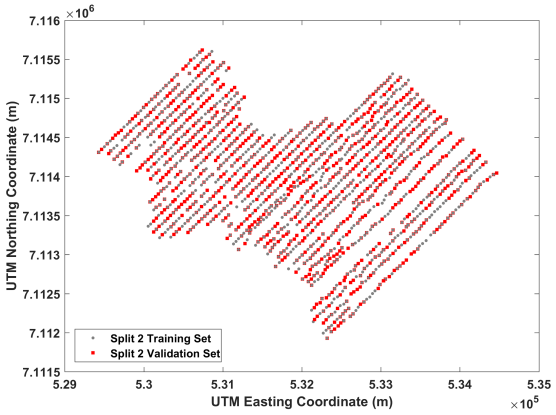
Additional figures from Chapter 4



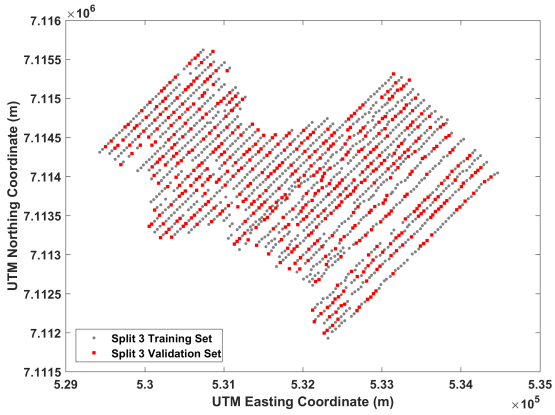
(A3)



(B1)

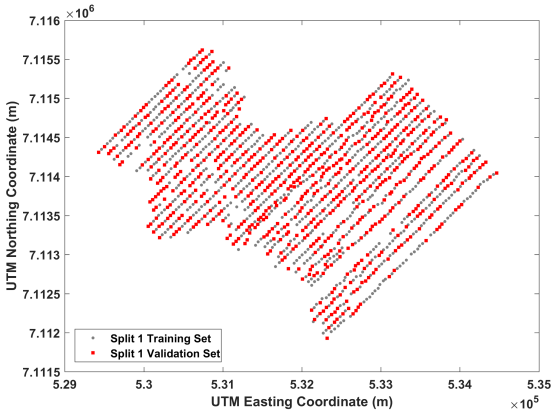


(B2)

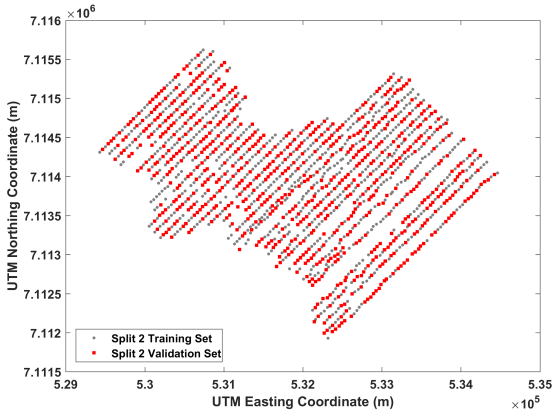


(B3)

Figure B-1, Spatial expression of training – validation subsets A3, B1, B2, and B3

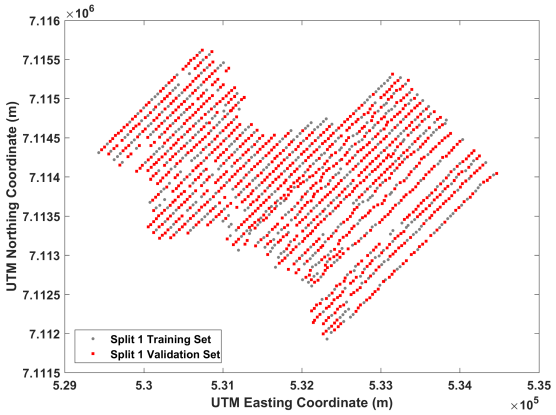


(C1)

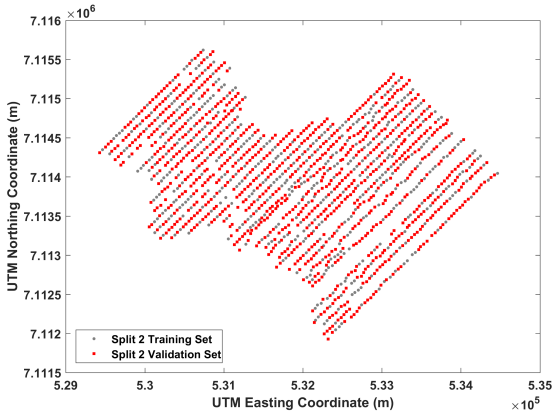


(C2)

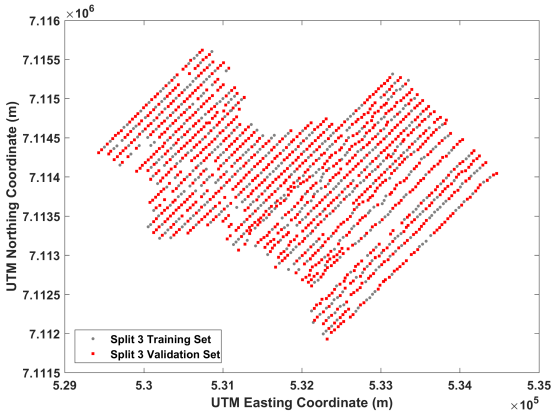
Figure B-2 Spatial expression of training – validation subsets for data reduction increment C, subsets 1 and 2



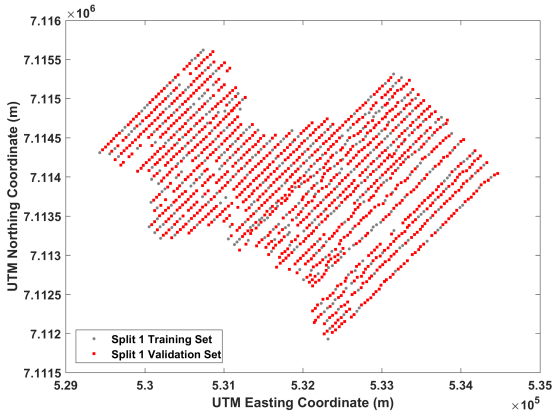
(D1)



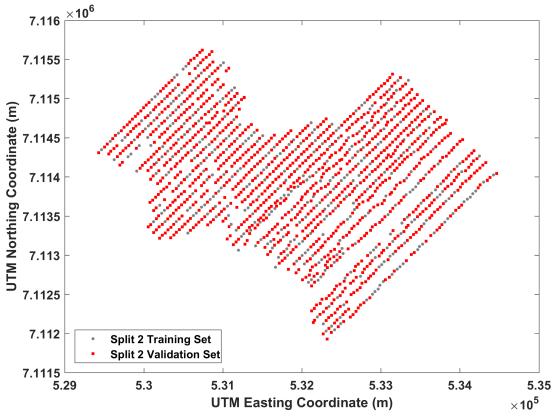
(D2)



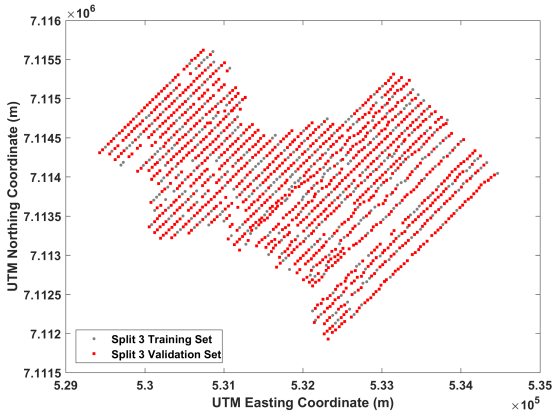
(D3)



(E1)



(E2)



(E3)

Figure B-3, Spatial expression of training – validation subsets for data reduction increments D and E, subsets 1-3

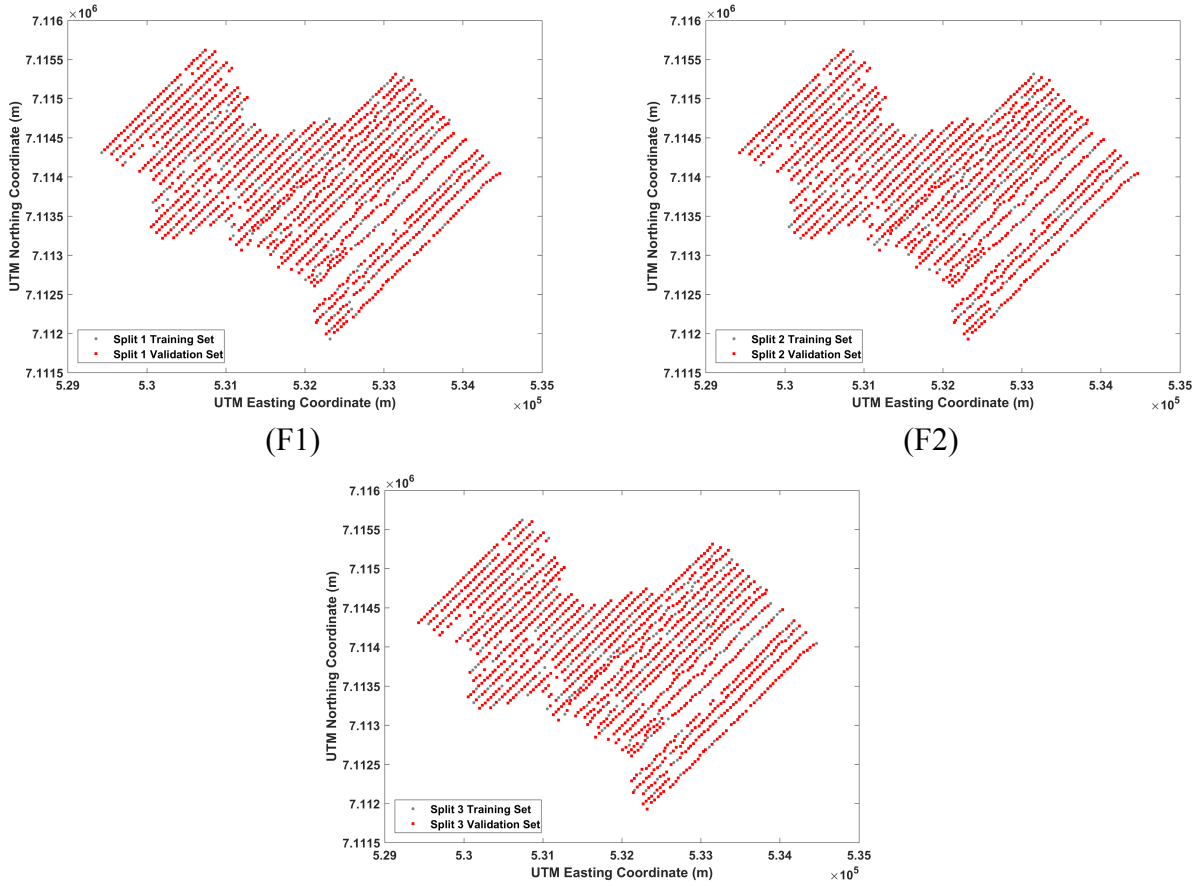
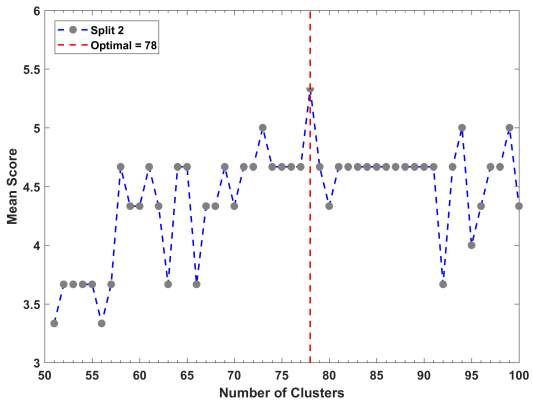
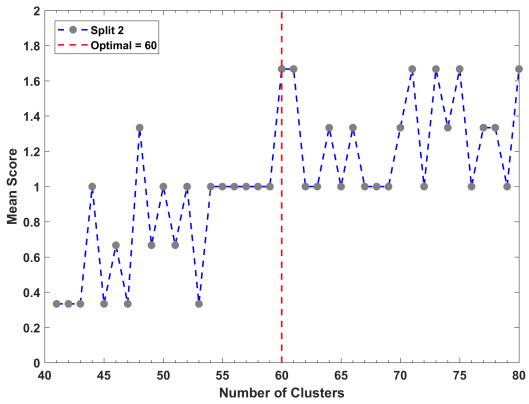


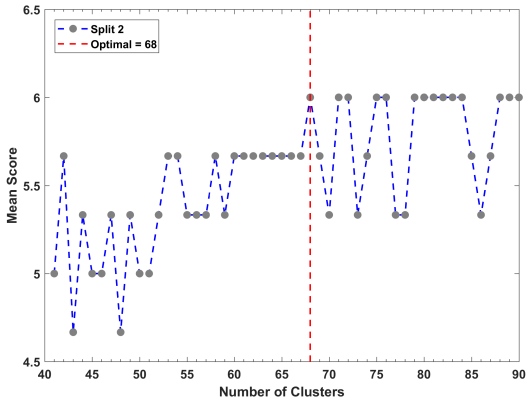
Figure B-4, Spatial expression of training – validation subsets for data reduction increment F, subsets 1-3



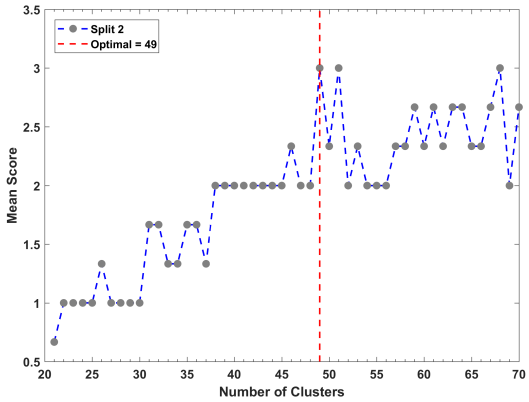
(A)



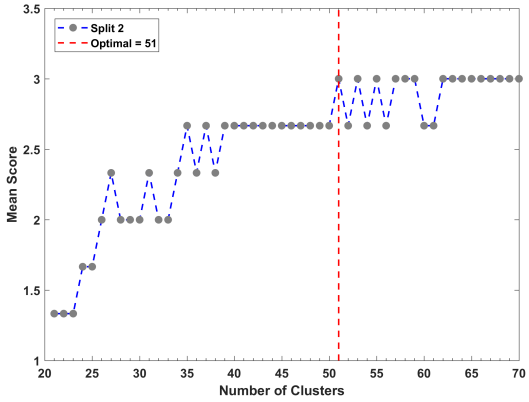
(B)



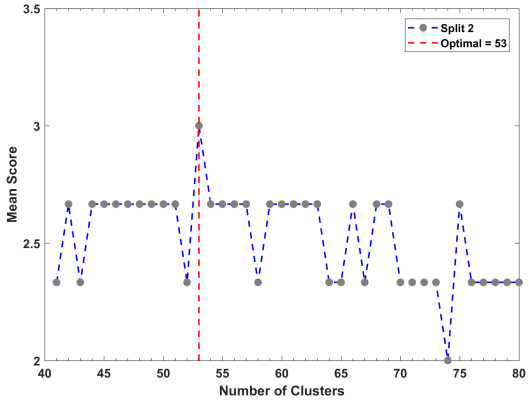
(C)



(D)

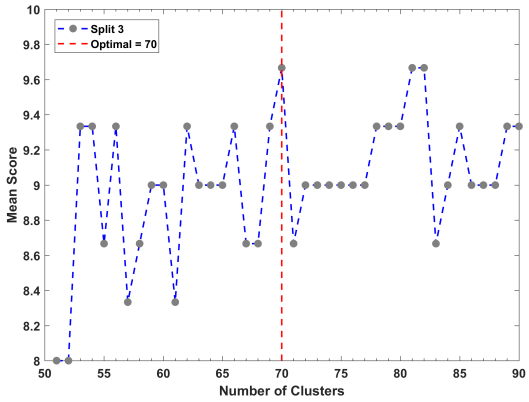


(E)

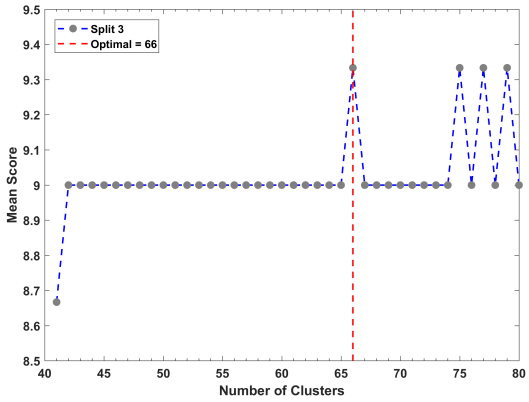


(F)

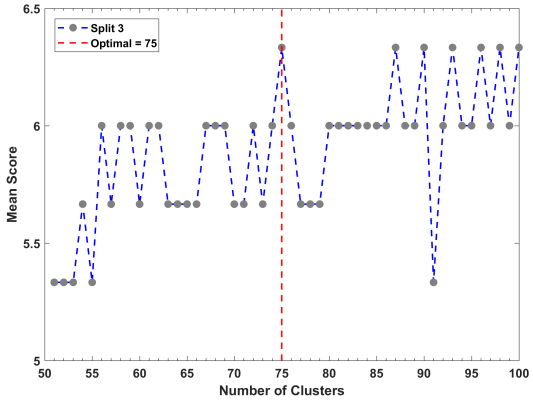
Figure B-5, Cluster validation results for data reduction increments A-F subset 2



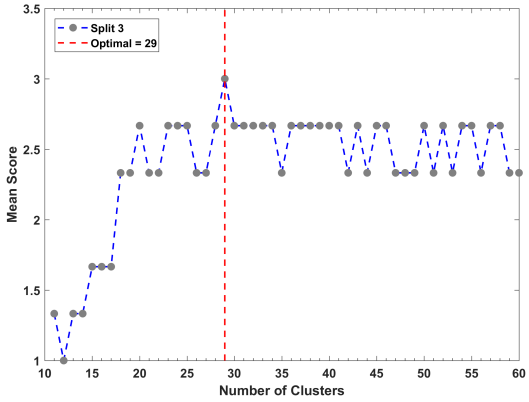
(A)



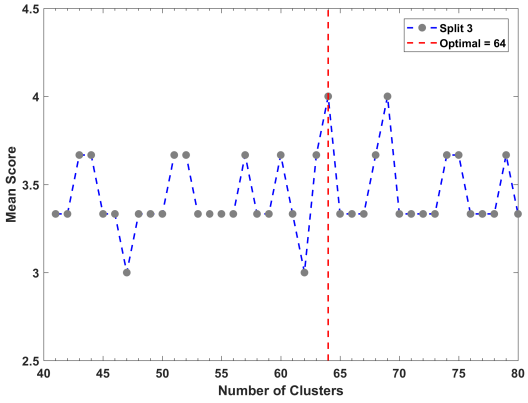
(B)



(D)

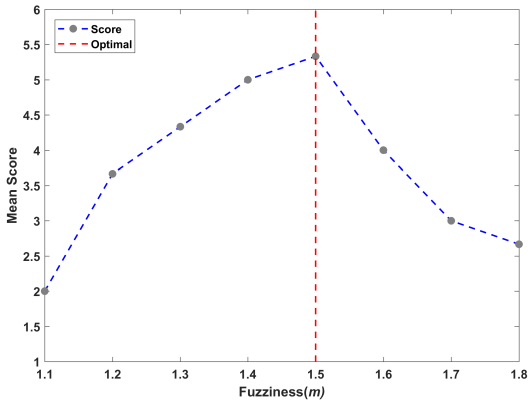


(E)

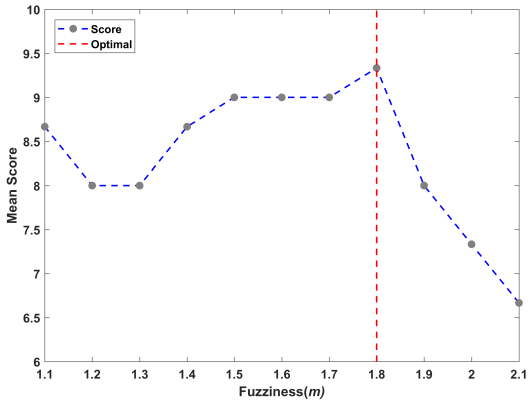


(F)

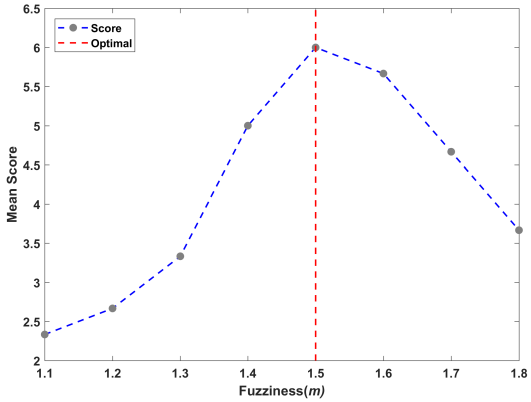
Figure B-6, Cluster validation results for data reduction increments A-F subset 3



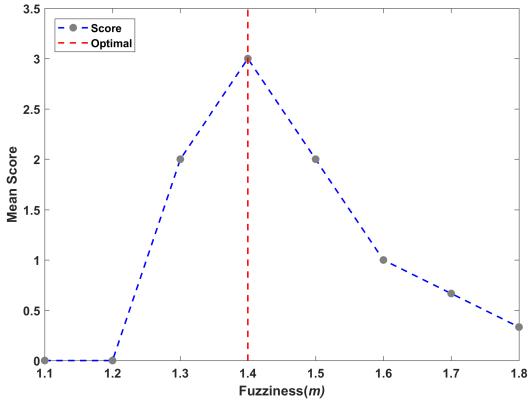
(A2)



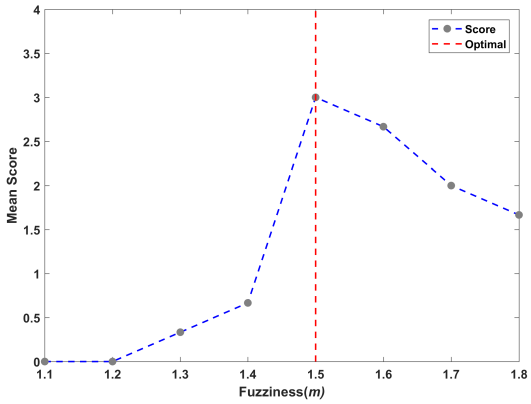
(B2)



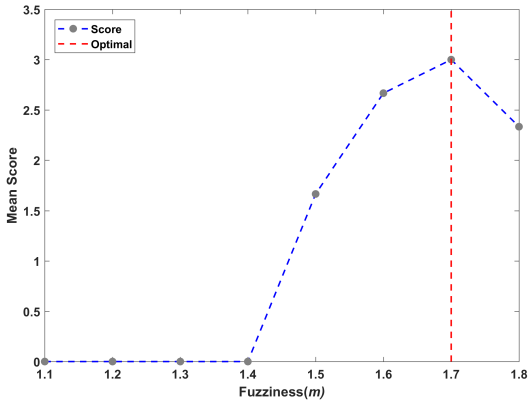
(C2)



(D2)

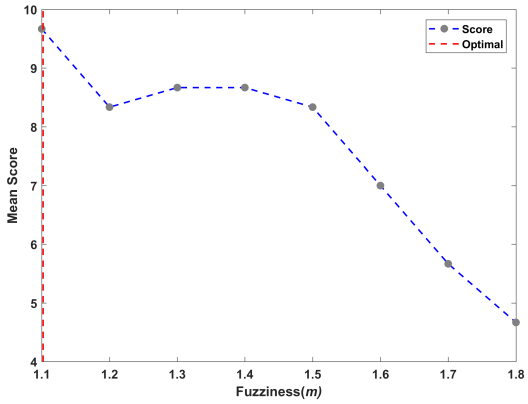


(E2)

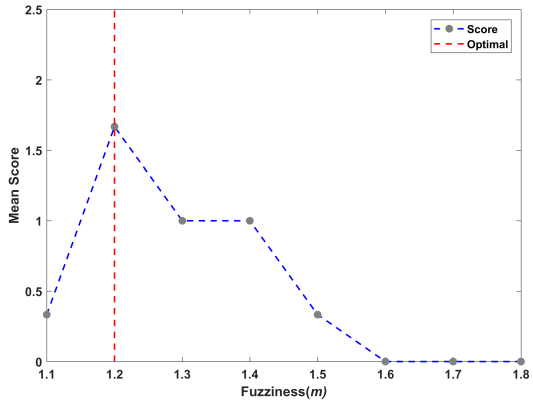


(F2)

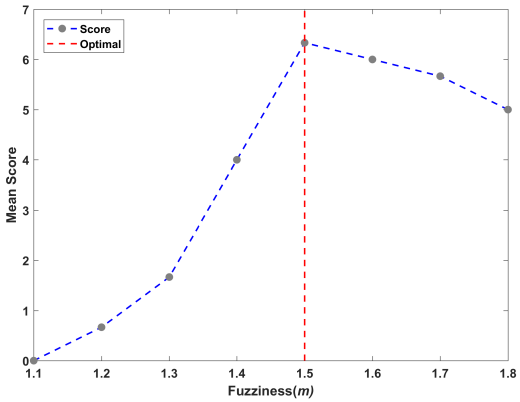
Figure B-7, Fuzziness parameter (*m*) validation results for data reduction increments A-F subset 2



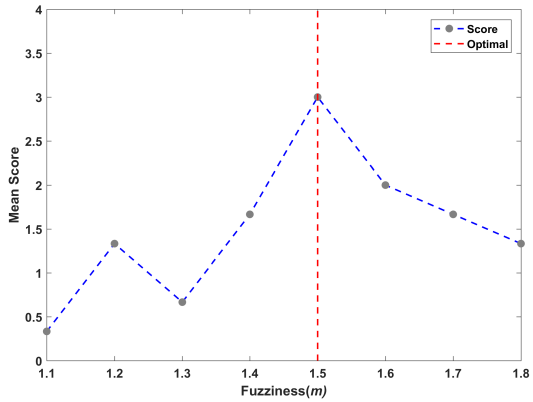
(A3)



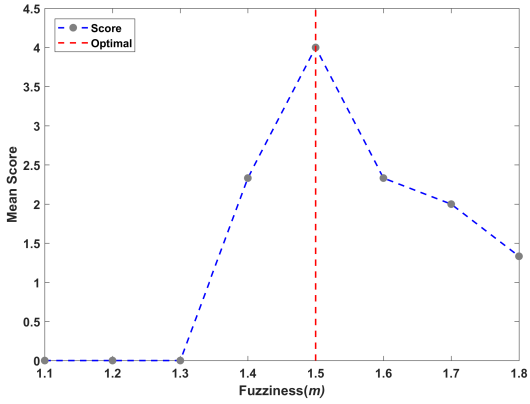
(B3)



(D3)

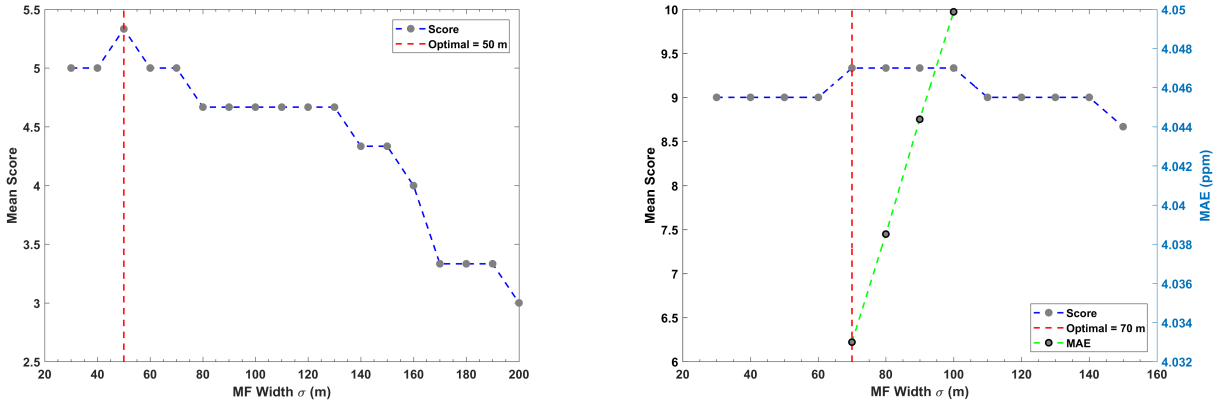


(E3)



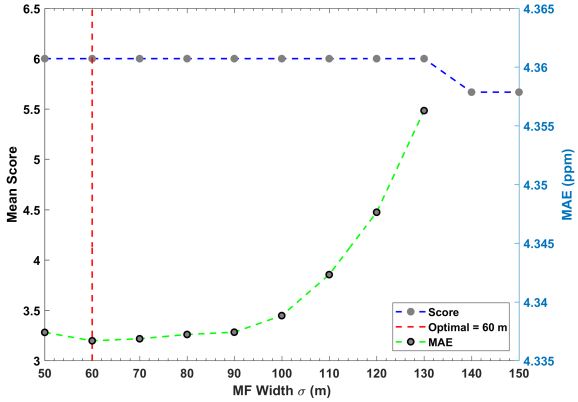
(F3)

Figure B-8, Fuzziness parameter (*m*) validation results for data reduction increments A-B and D-F subset 3

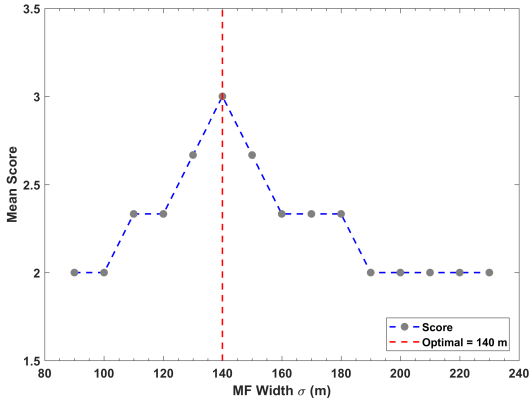


(A2)

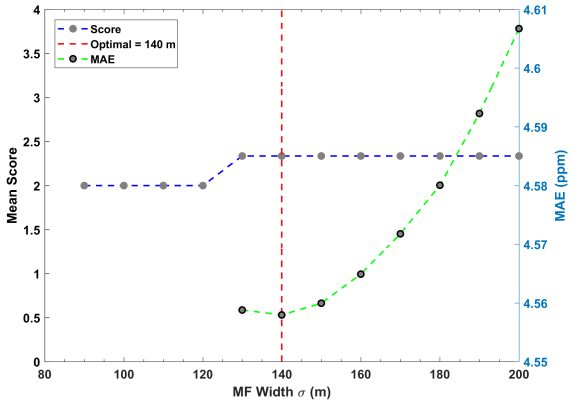
(B2)



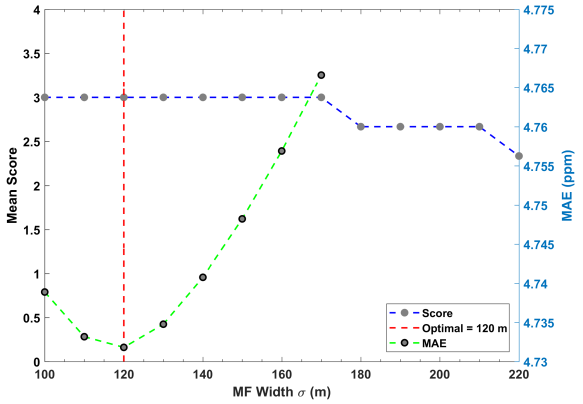
(C2)



(D2)

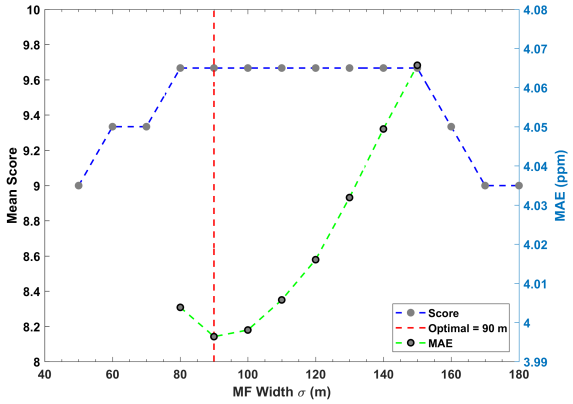


(E2)

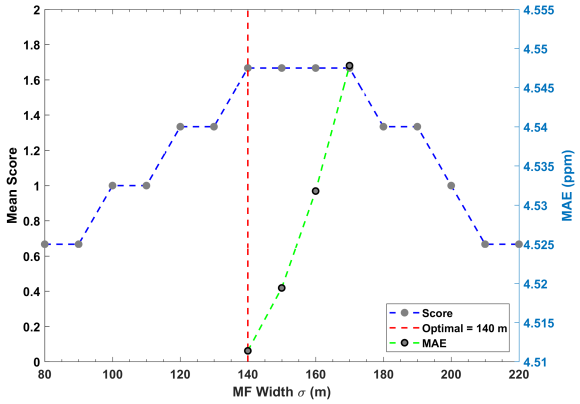


(F2)

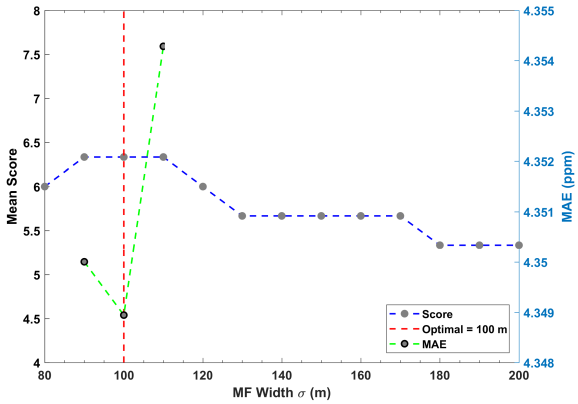
Figure B-9, Membership function width (σ) validation results for data reduction increments A-F subset 2



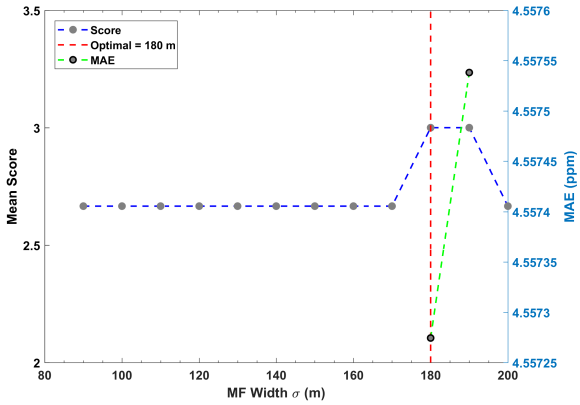
(A3)



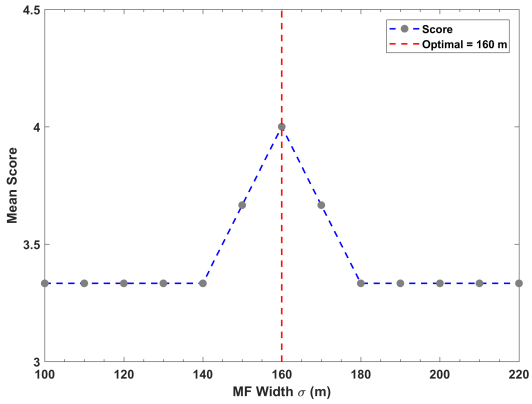
(B3)



(D3)



(E3)



(F3)

Figure B-10, Membership function width (σ) validation results for data reduction increments A-B and D-F subset 3

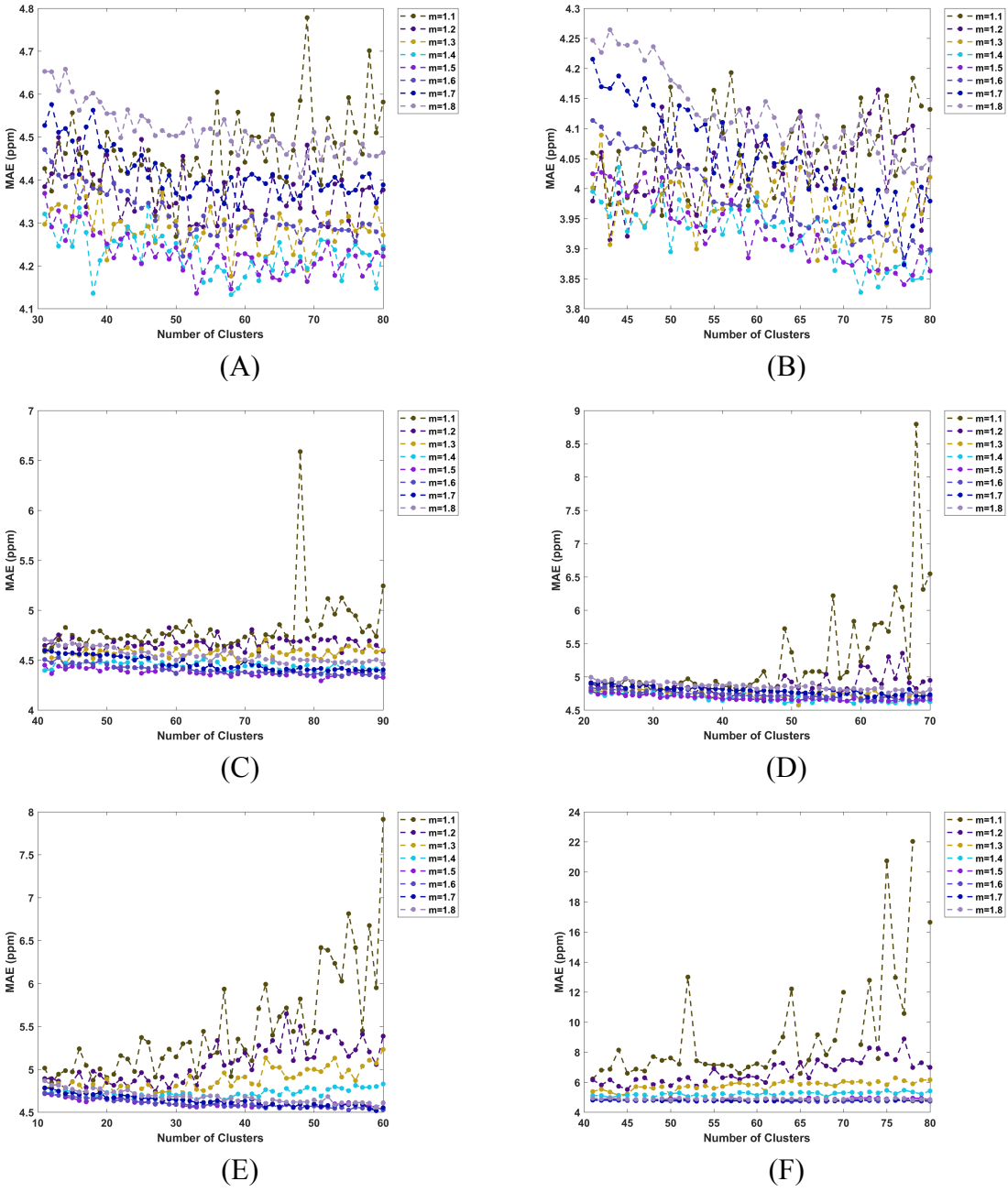


Figure B-11, MAE plotted against number of clusters for *m*=1.1 to 1.8 for Training sets A-F, subset 2

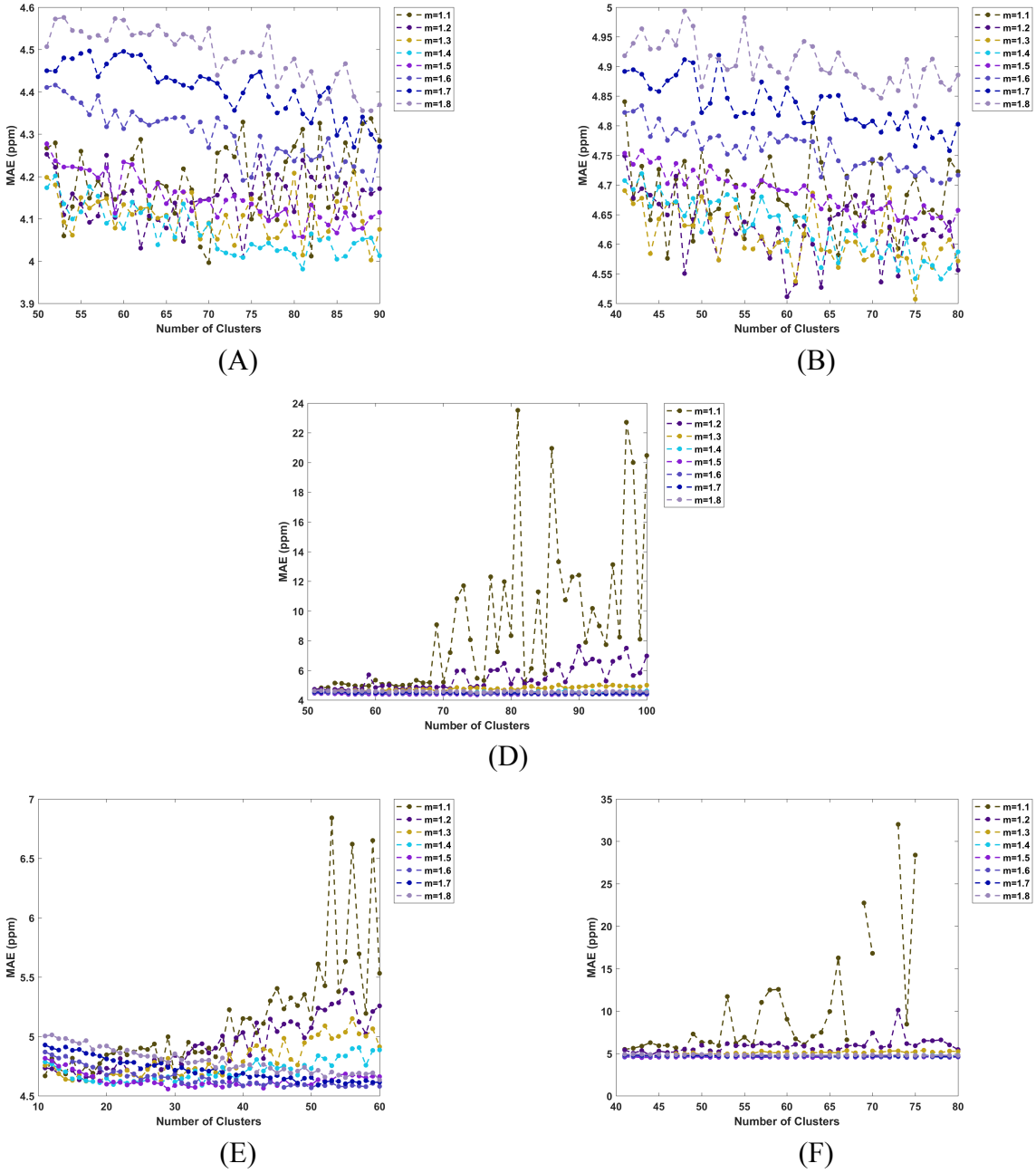
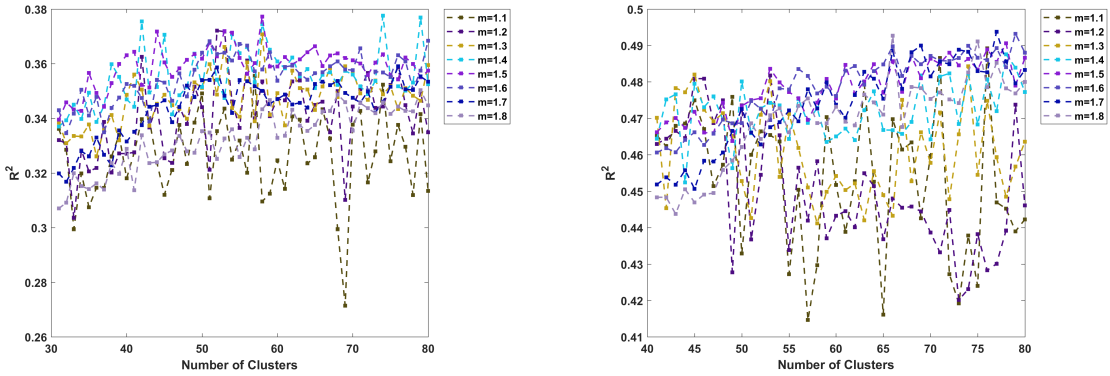
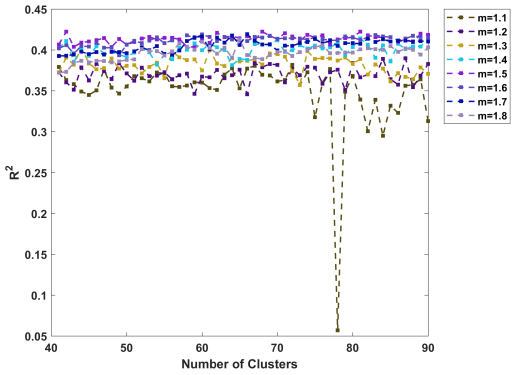


Figure B-12, MAE plotted against number of clusters for *m*=1.1 to 1.8 for Training sets A-F, subset 3

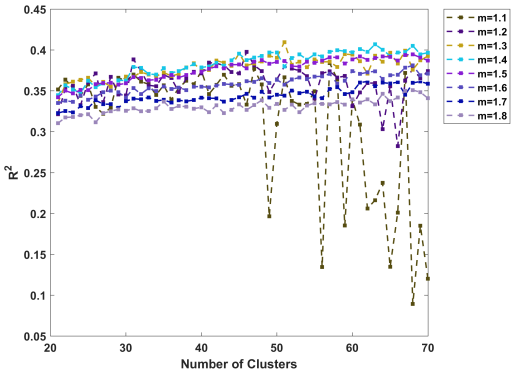


(A)

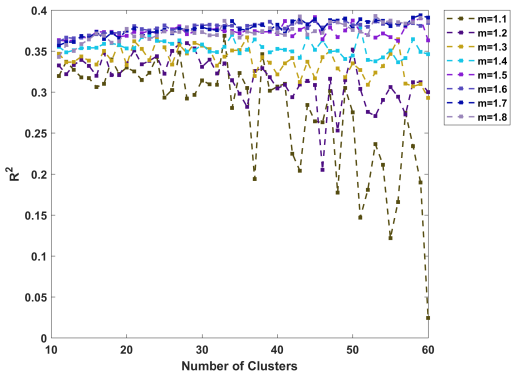
(B)



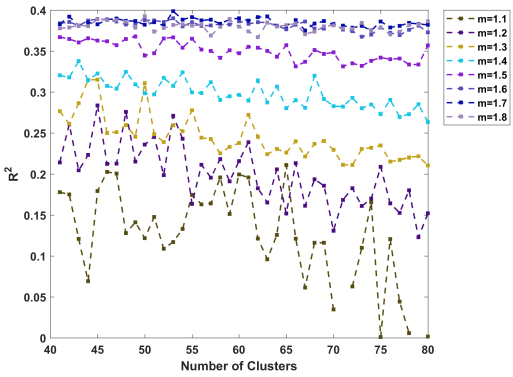
(C)



(D)

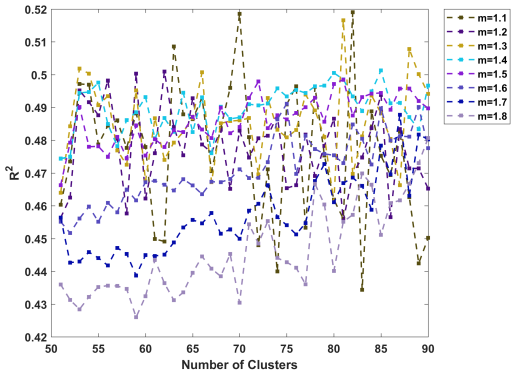


(E)

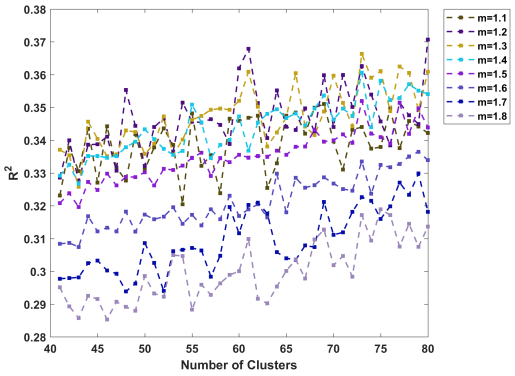


(F)

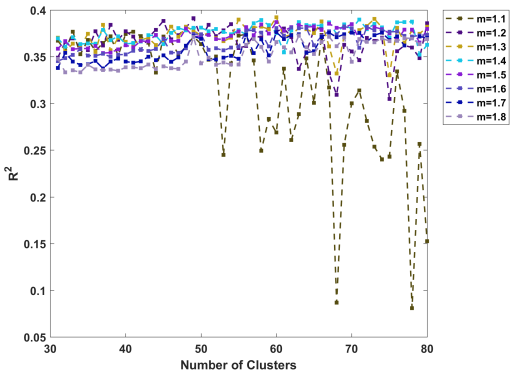
Figure B-13, R^2 plotted against number of clusters for *m*=1.1 to 1.8 for Training sets A-F, subset 2



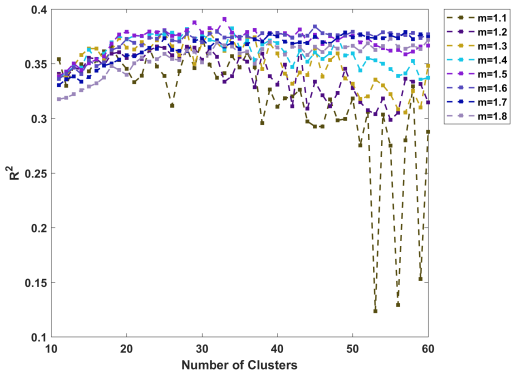
(A)



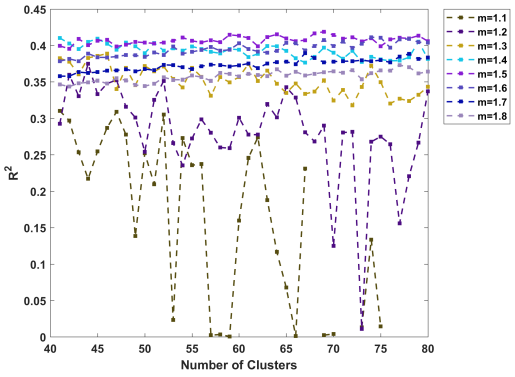
(B)



(D)



(E)



(F)

Figure B-14, R^2 plotted against number of clusters for *m*=1.1 to 1.8 for Training sets A-F, subset 2

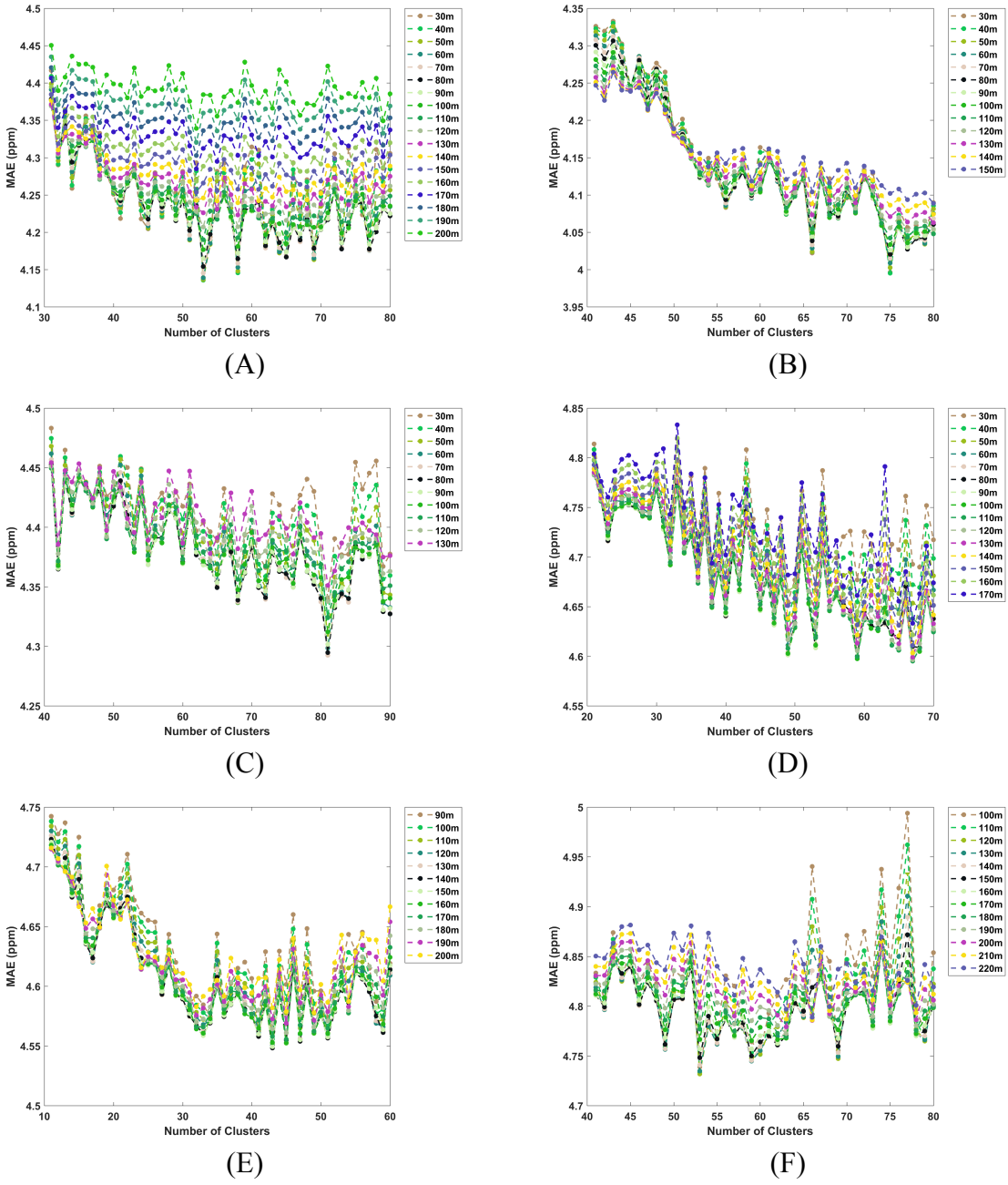
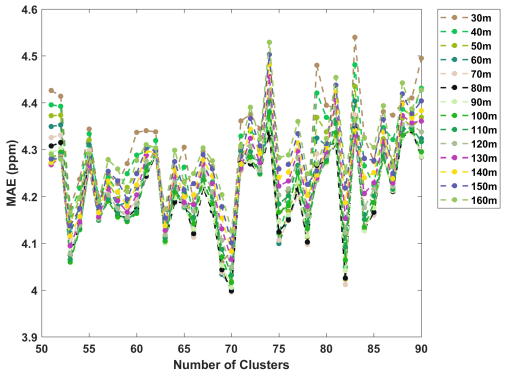
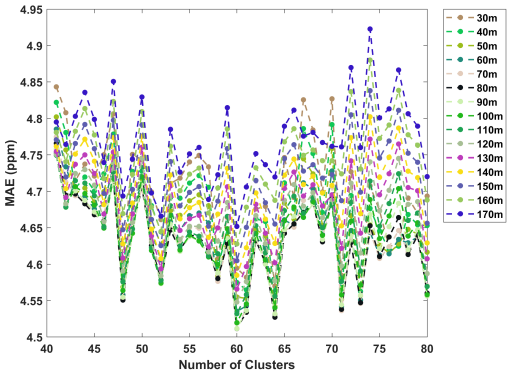


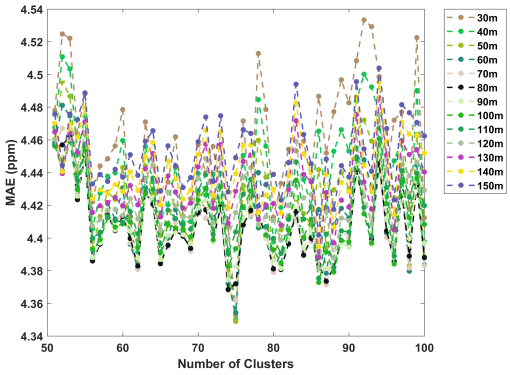
Figure B-15, MAE plotted against number of clusters for ranges of σ tested during the pseudo-optimization for Training sets A-F, subset 2



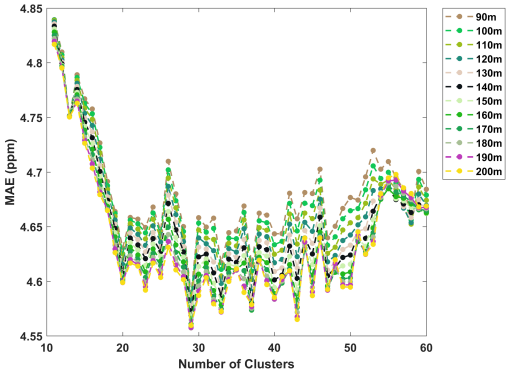
(A)



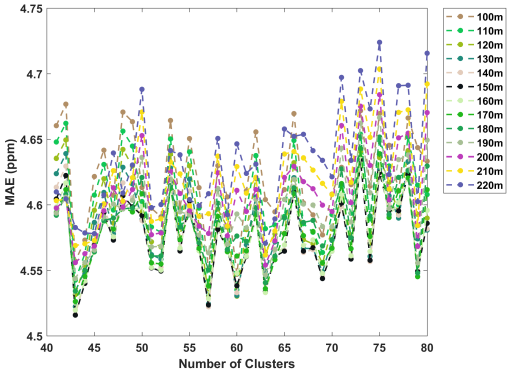
(B)



(D)



(E)



(F)

Figure B-16, MAE plotted against number of clusters for ranges of σ tested during the pseudo-optimization for Training sets A-F, subset 3

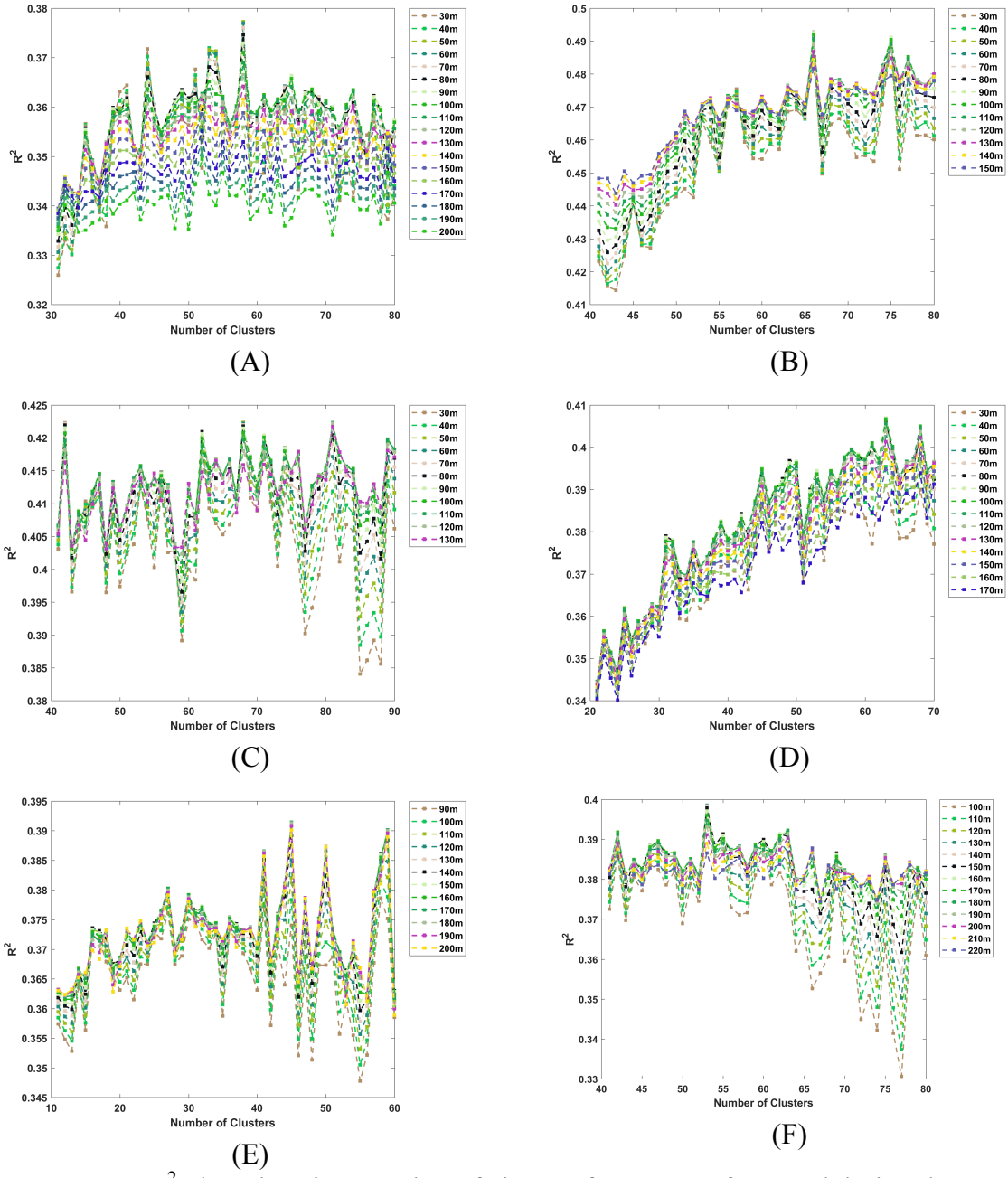
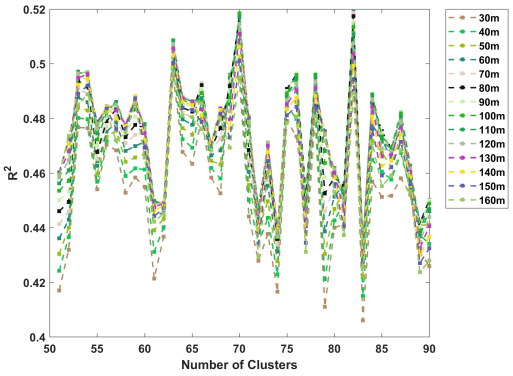
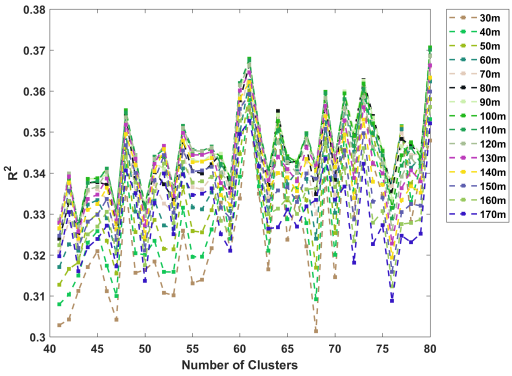


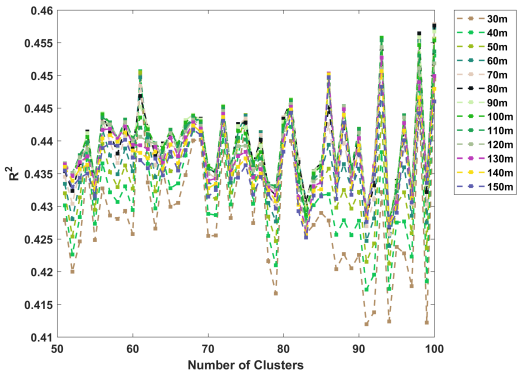
Figure B-17, R^2 plotted against number of clusters for ranges of σ tested during the pseudo-optimization for Training sets A-F, subset 2



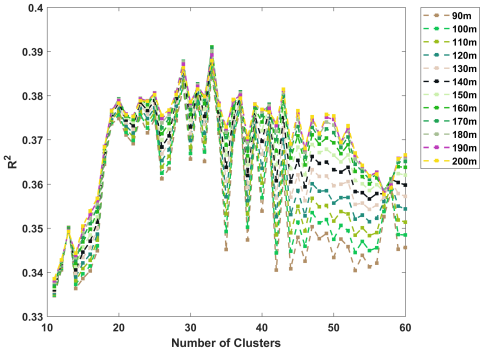
(A)



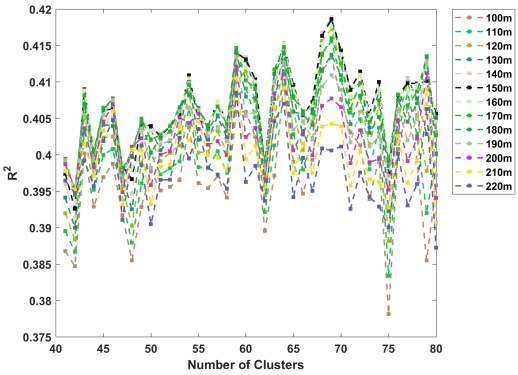
(B)



(D)

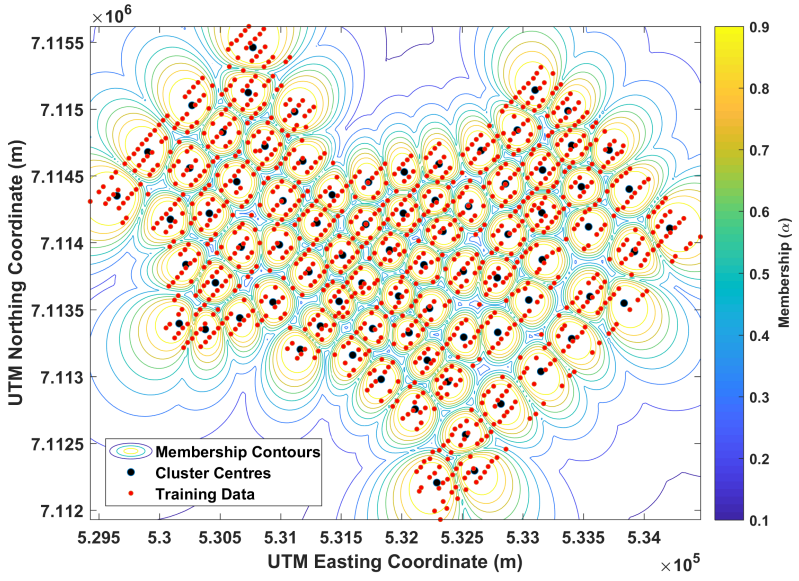


(E)

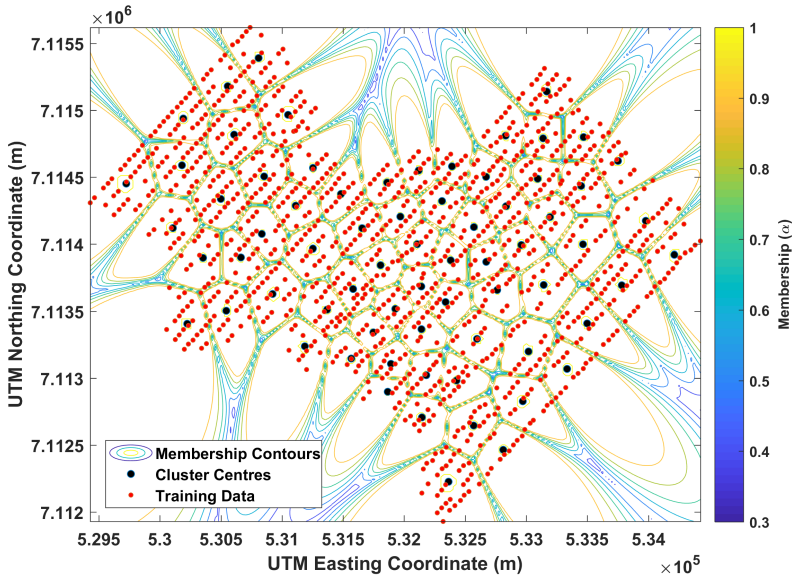


(F)

Figure B-18, R^2 plotted against number of clusters for ranges of σ tested during the pseudo-optimization for Training sets A-F, subset 3

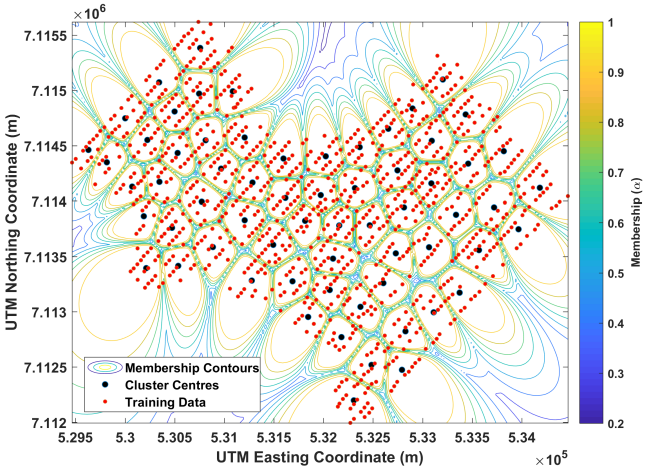


(A2)

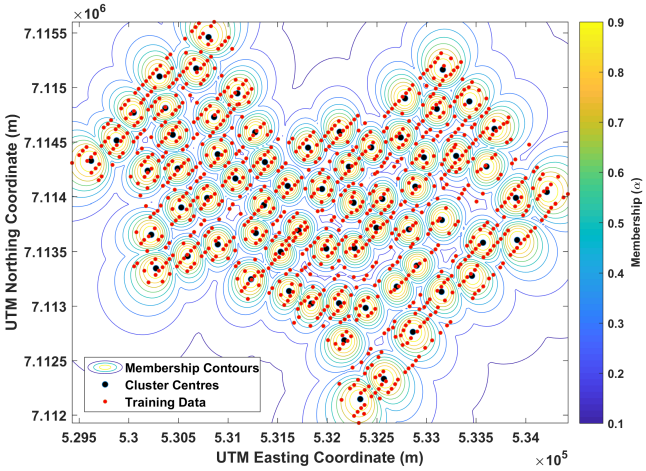


(A3)

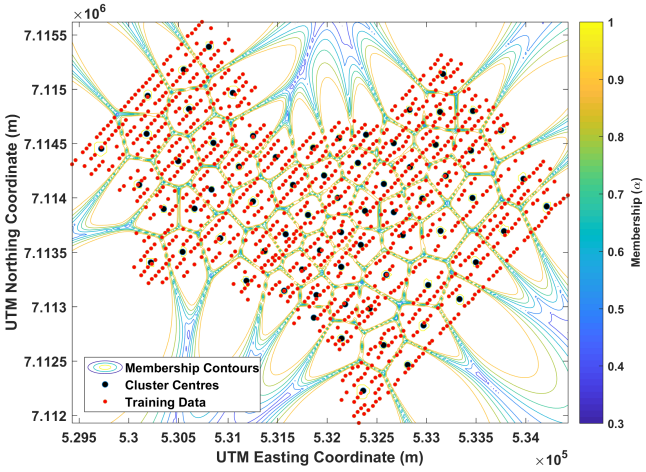
Figure B-19, Membership contour map for data reduction increment A, subsets 2 and 3



(B1)

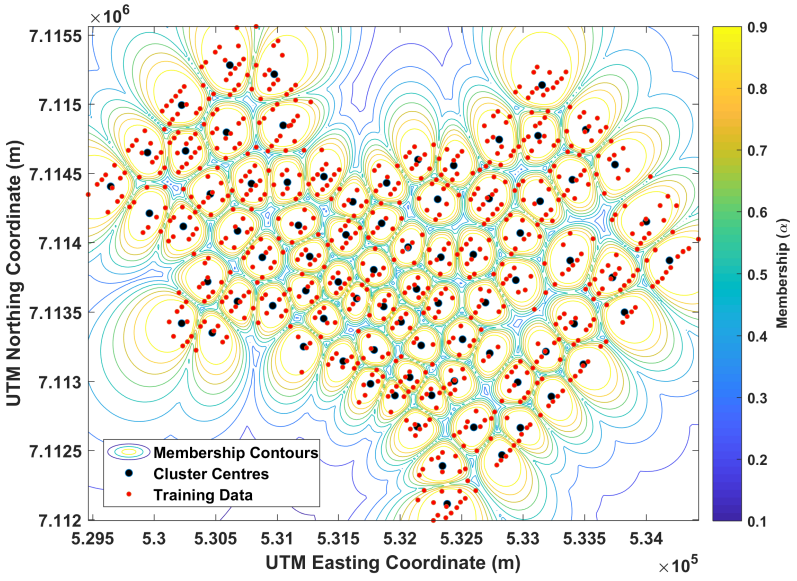


(B2)

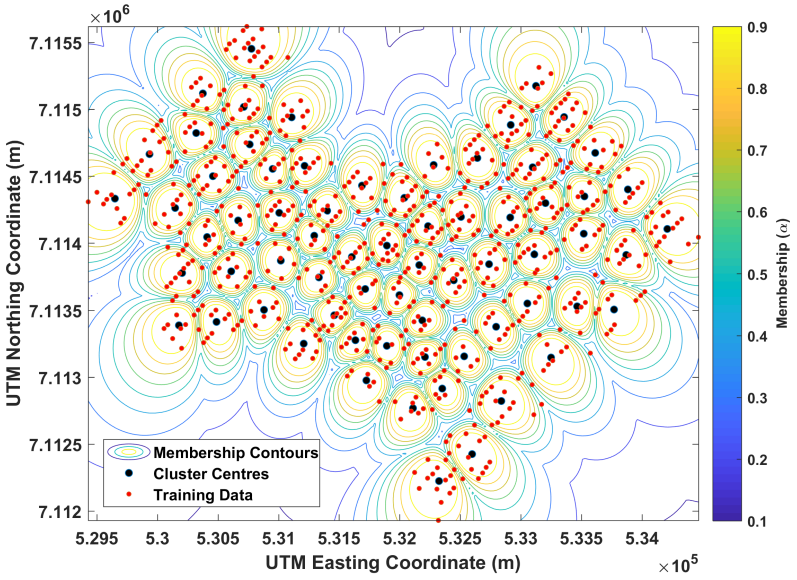


(B3)

Figure B-20, Membership contour map for data reduction increment B, subsets 1-3

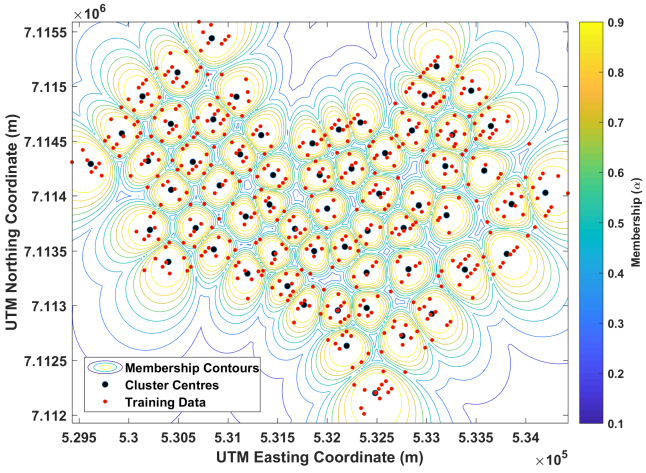


(C1)

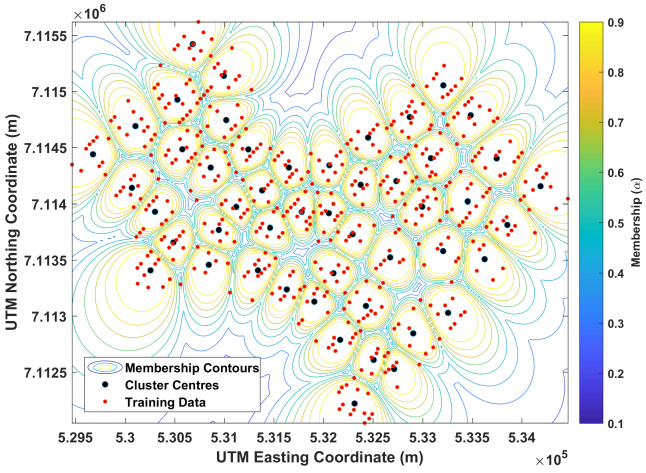


(C2)

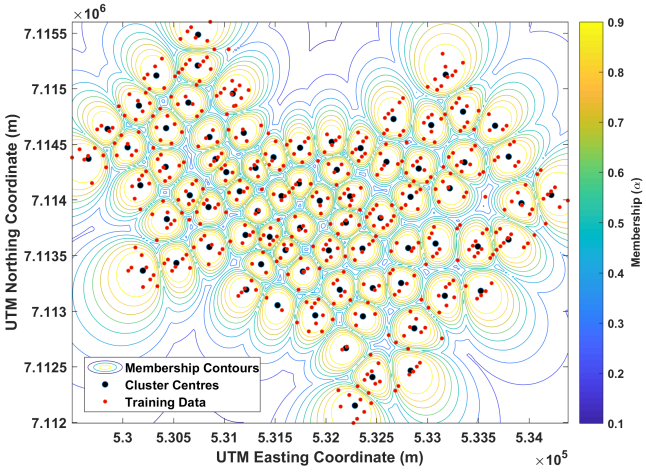
Figure B-21, Membership contour map for data reduction increment C, subsets 1 and 2



(D1)

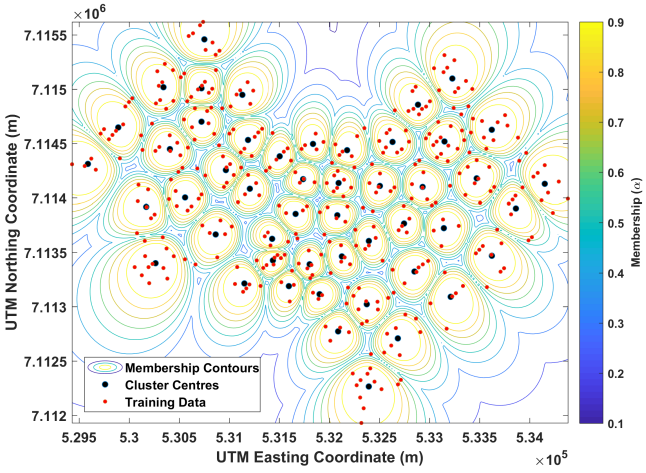


(D2)

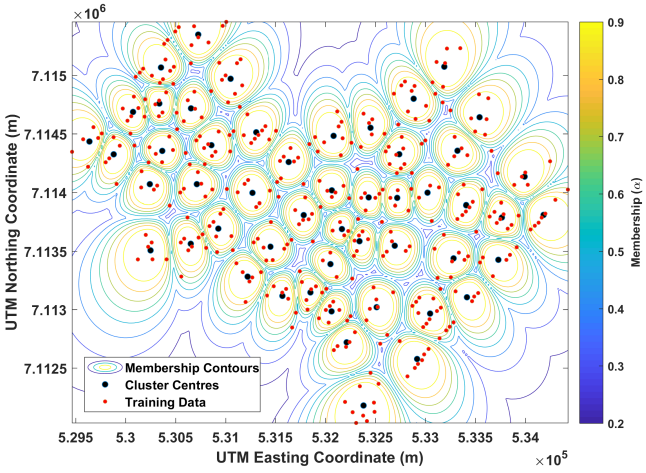


(D3)

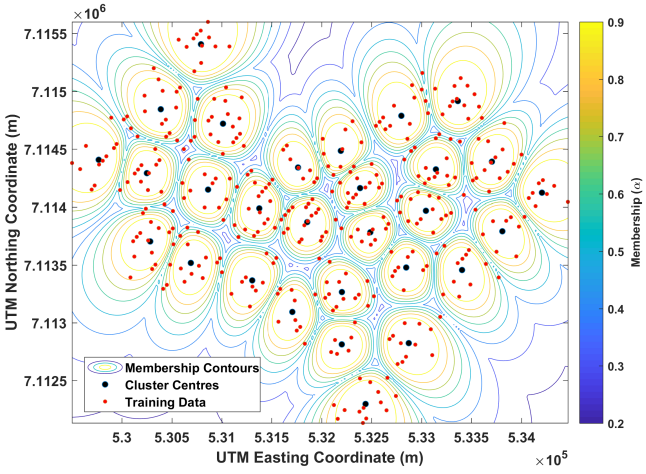
Figure B-22, Membership contour map for data reduction increment D, subsets 1-3



(E1)

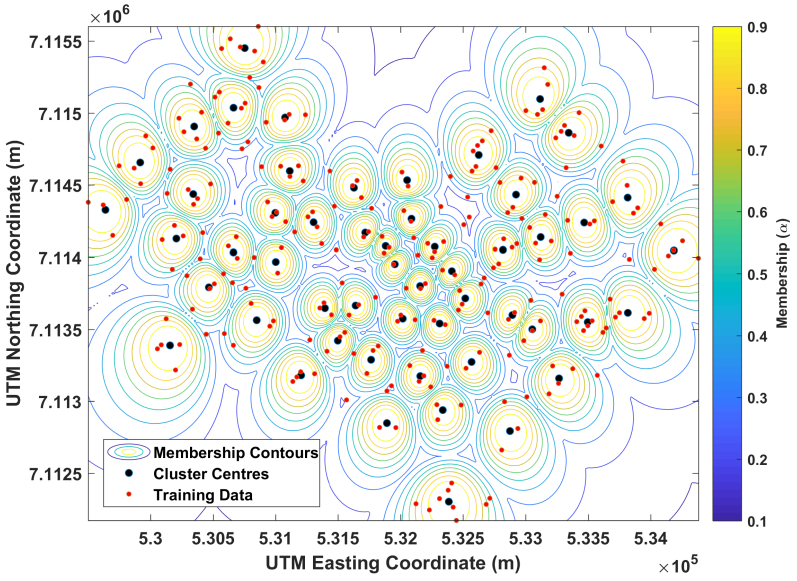


(E2)

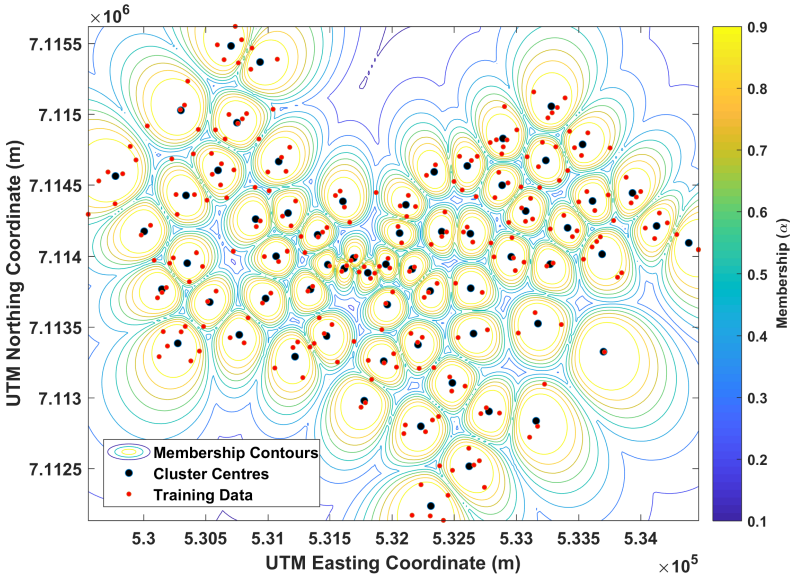


(E3)

Figure B-23, Membership contour map for data reduction increment E, subsets 1-3

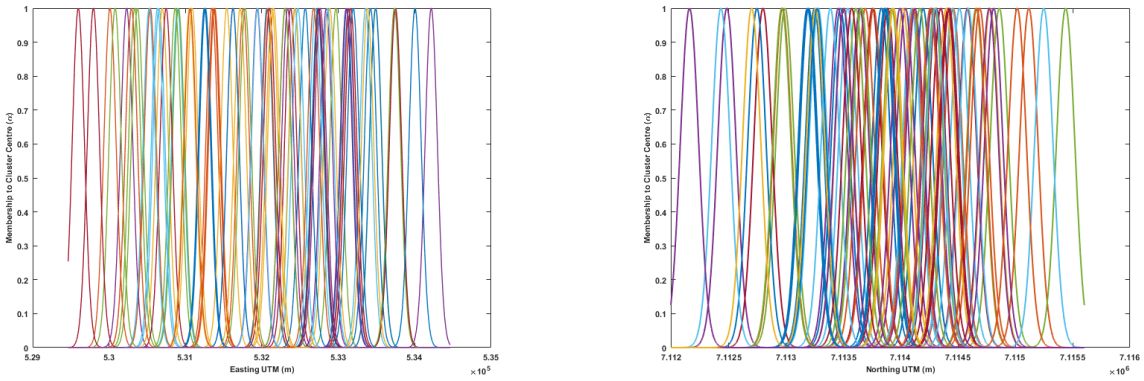


(F2)

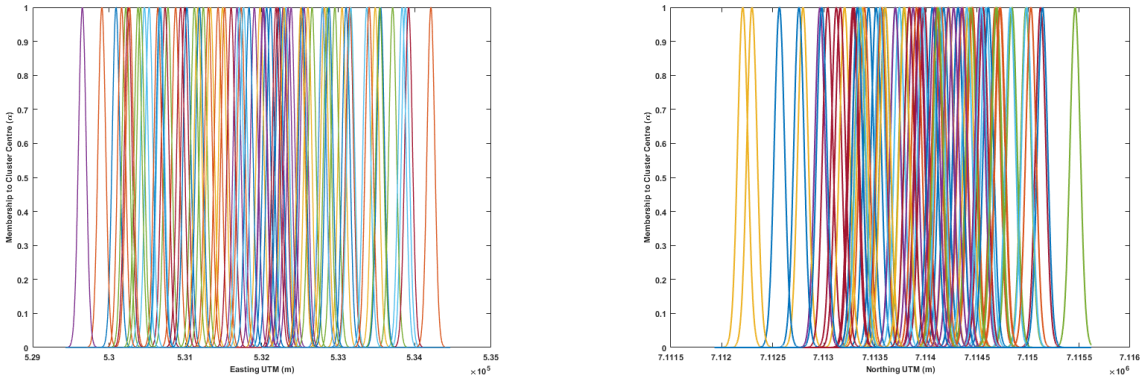


(F3)

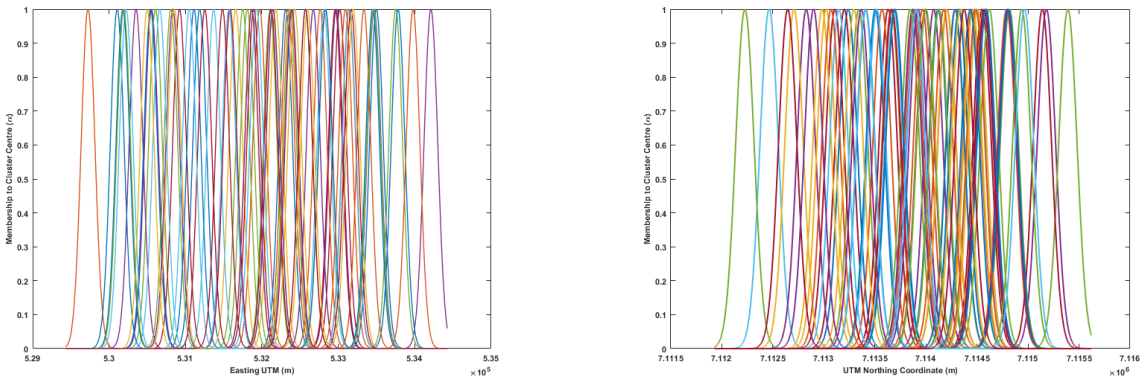
Figure B-24, Membership contour map for data reduction increment F, subsets 2 and 3



(A1)

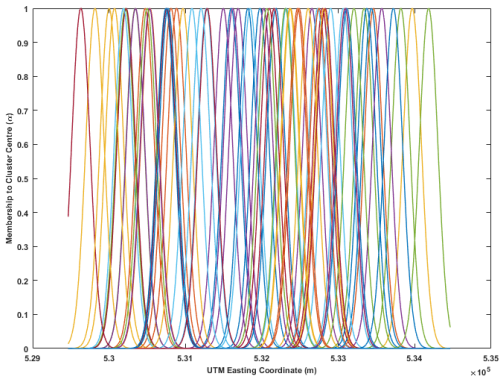


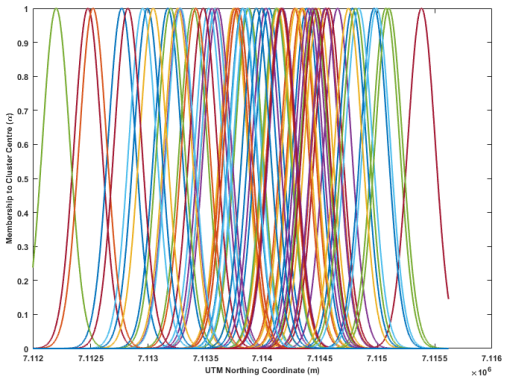
(A2)



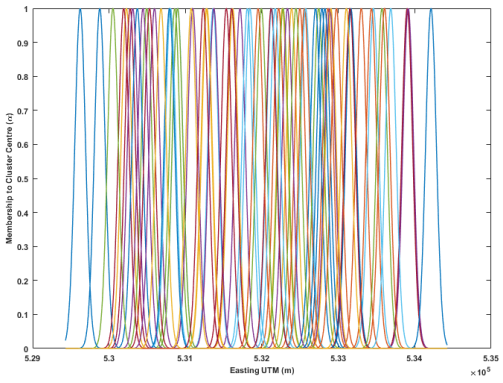
(A3)

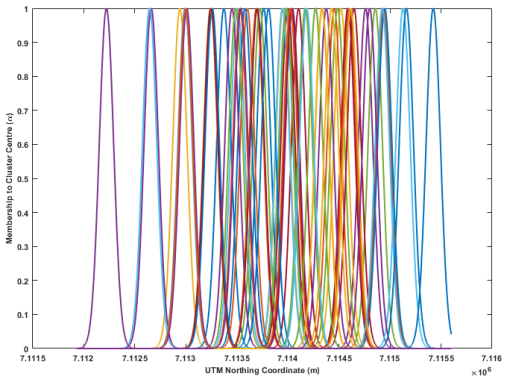
Figure B-25, Membership function for the easting and northing coordinate axes for data reduction increment A, subplots 1-3



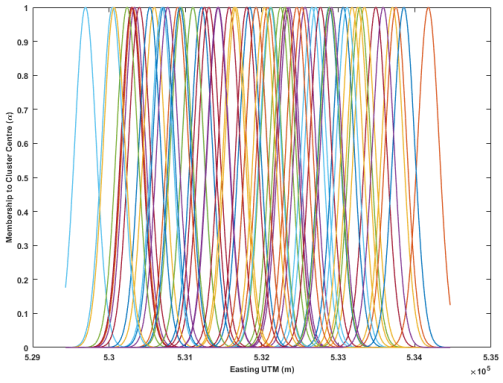


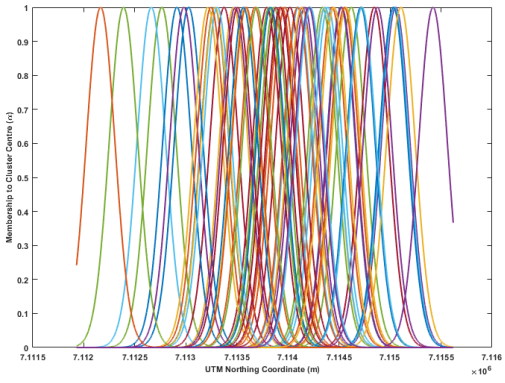
(B1)





(B2)





(B3)

Figure B-26, Membership function for the easting and northing coordinate axes for data reduction increment B, subplots 1-3

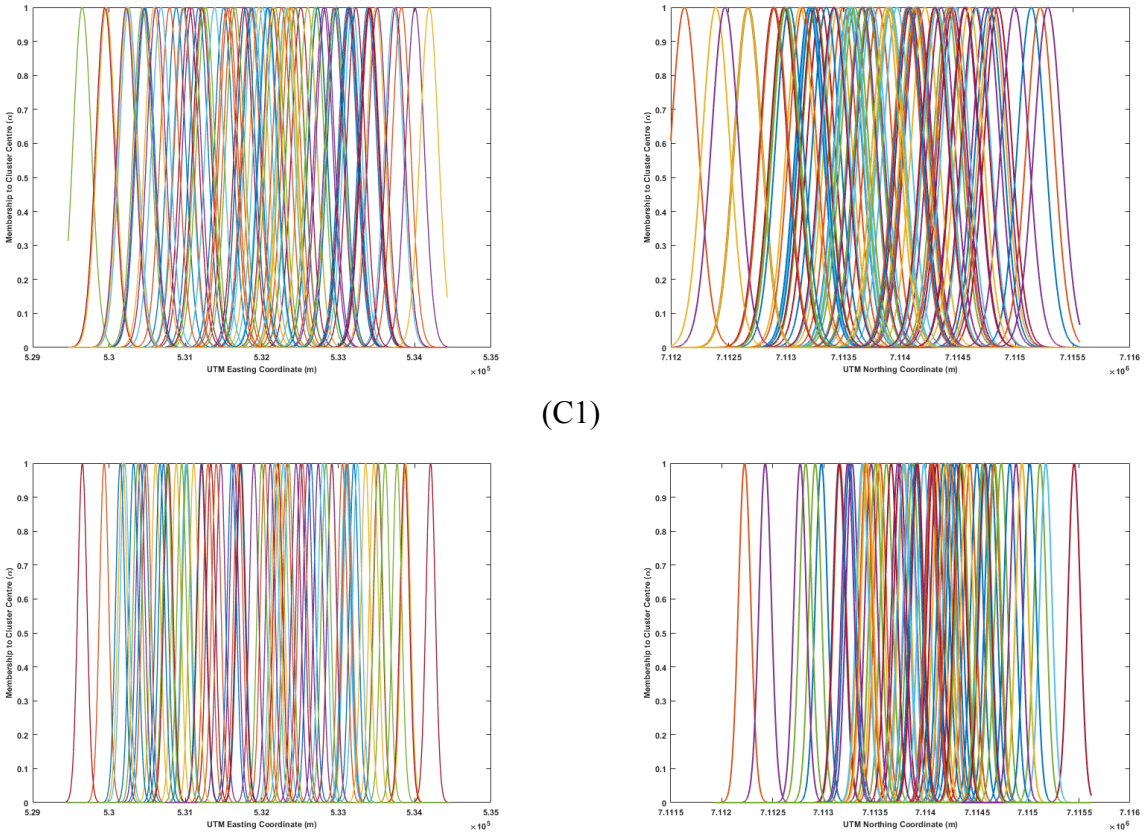
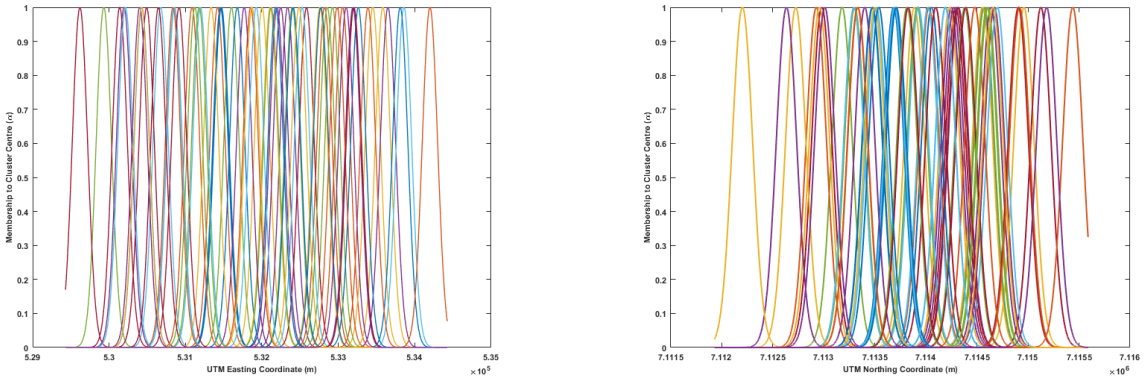
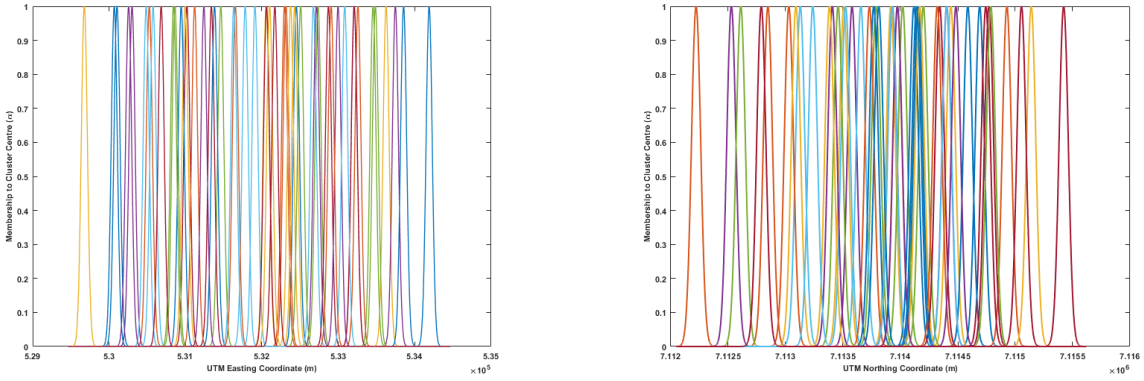


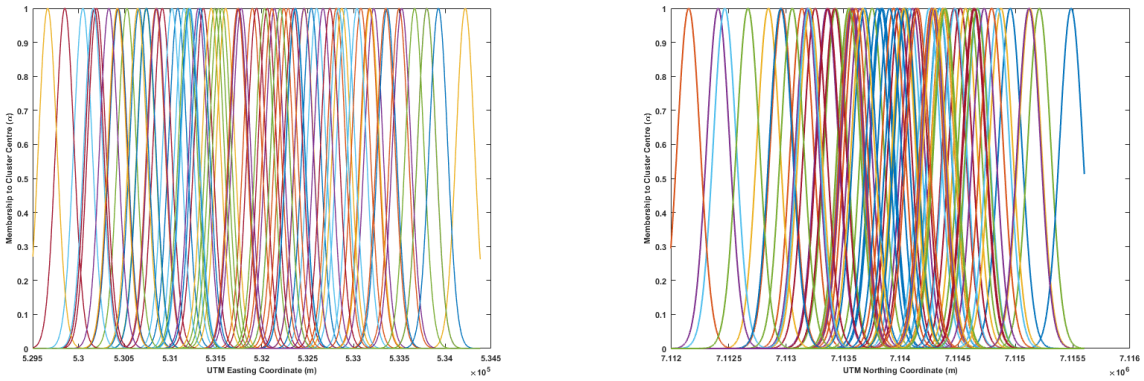
Figure B-27, Membership function for the easting and northing coordinate axes for data reduction increment A, subplots 1-2



(D1)

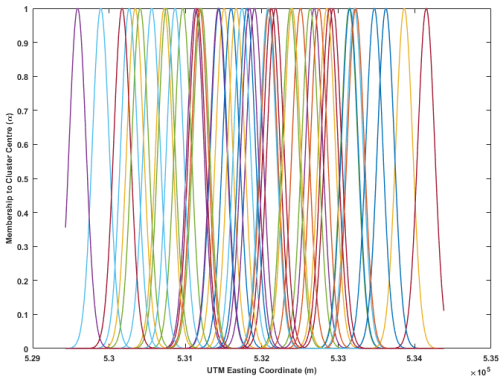


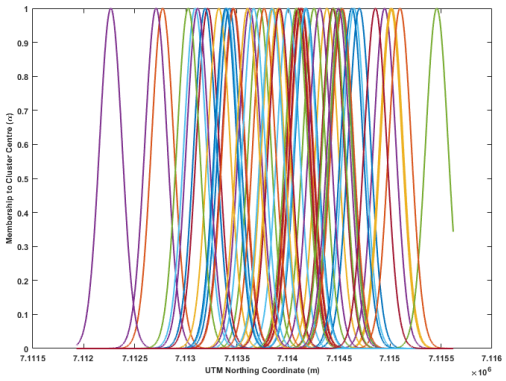
(D2)



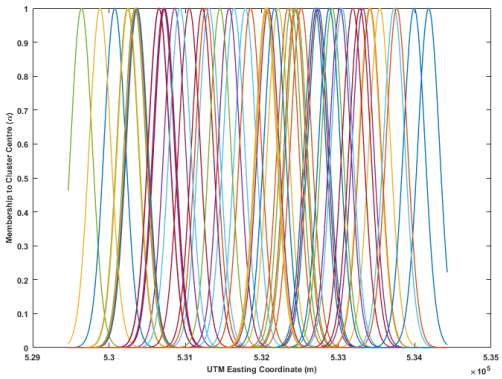
(D3)

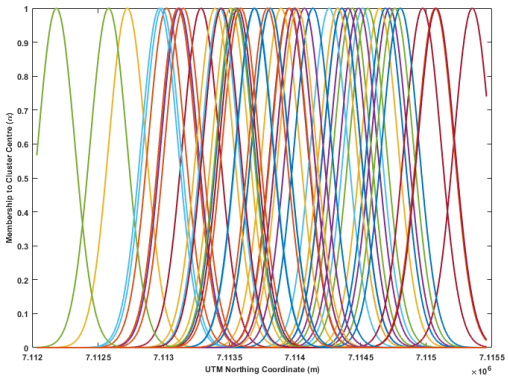
Figure B-28, Membership function for the easting and northing coordinate axes for data reduction increment D, subplots 1-3





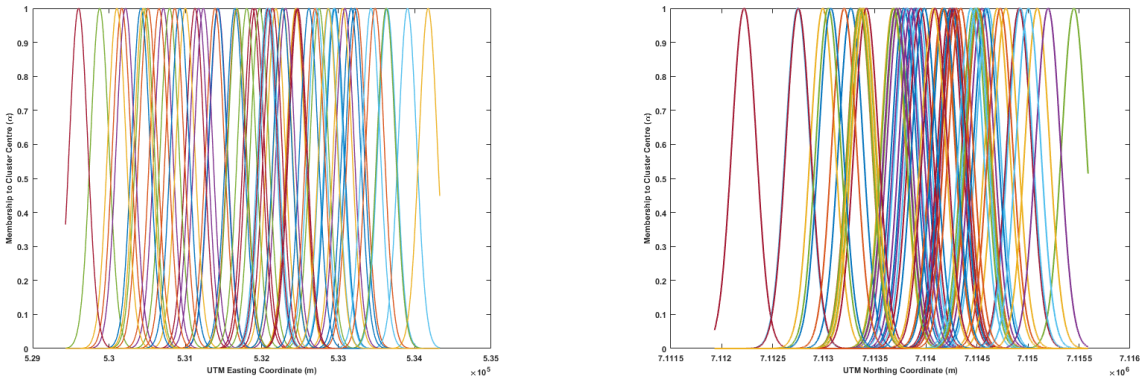
(E1)



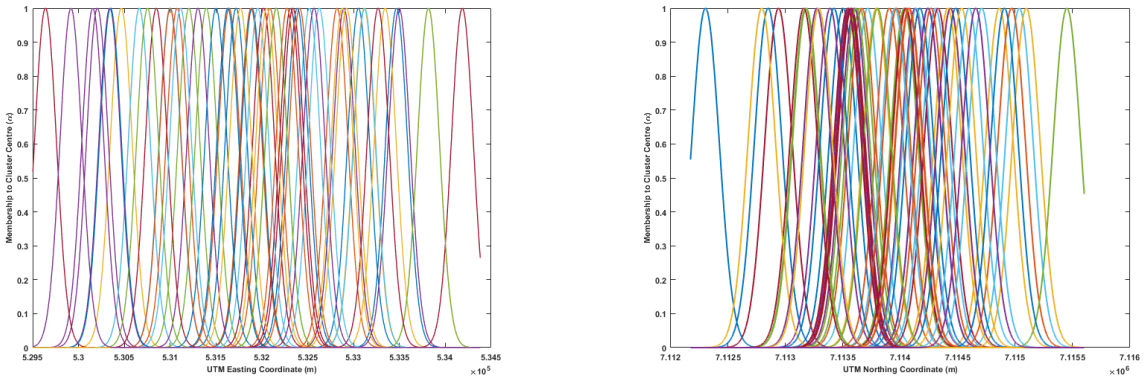


(E2)

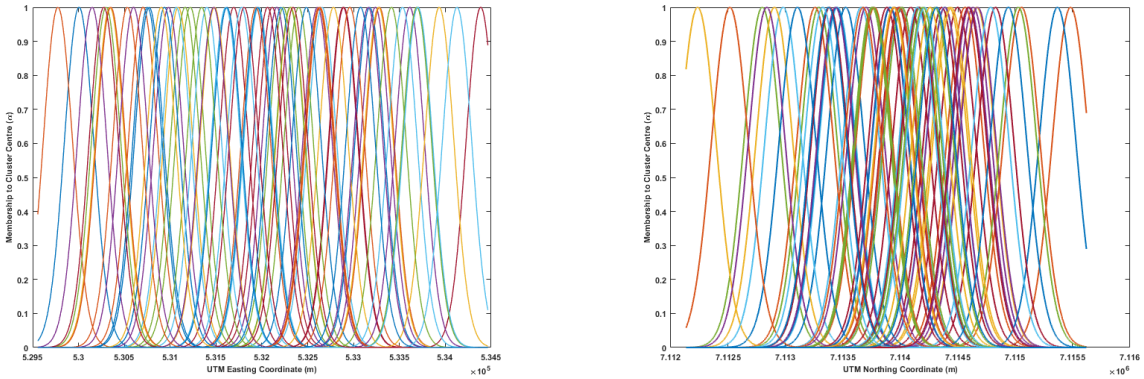
Figure B-29, Membership function for the easting and northing coordinate axes for data reduction increment E, subplots 1-2



(F1)

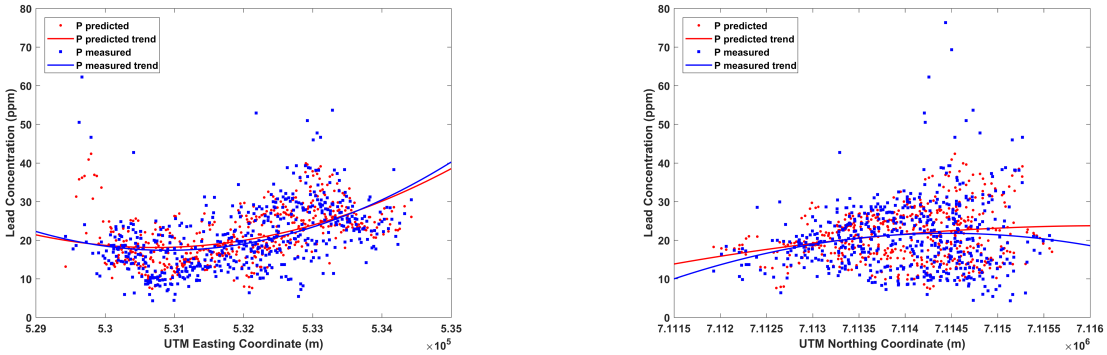


(F2)

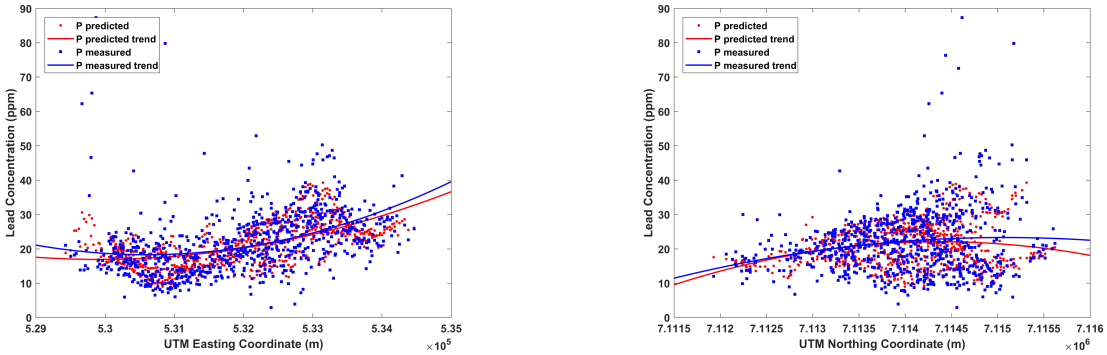


(F3)

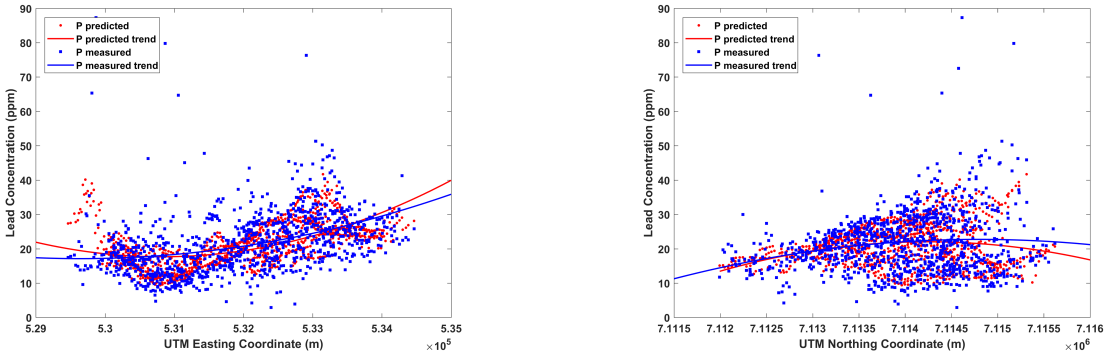
Figure B-30, Membership function for the easting and northing coordinate axes for data reduction increment F, subplots 1-3



(B1)

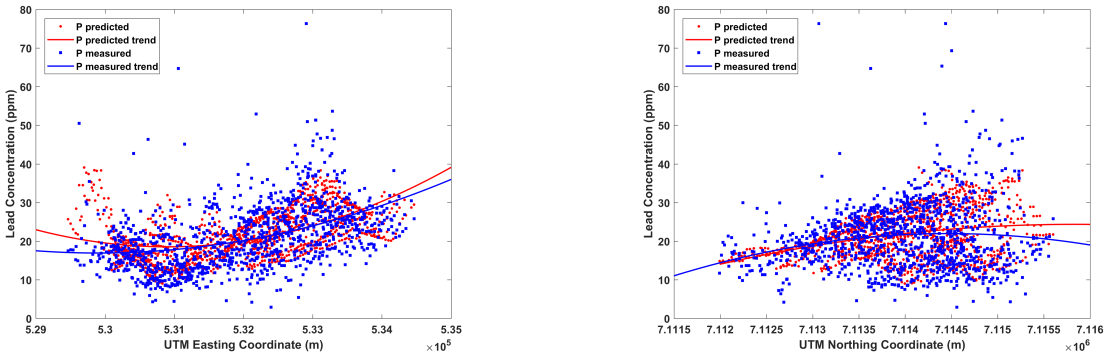


(C1)

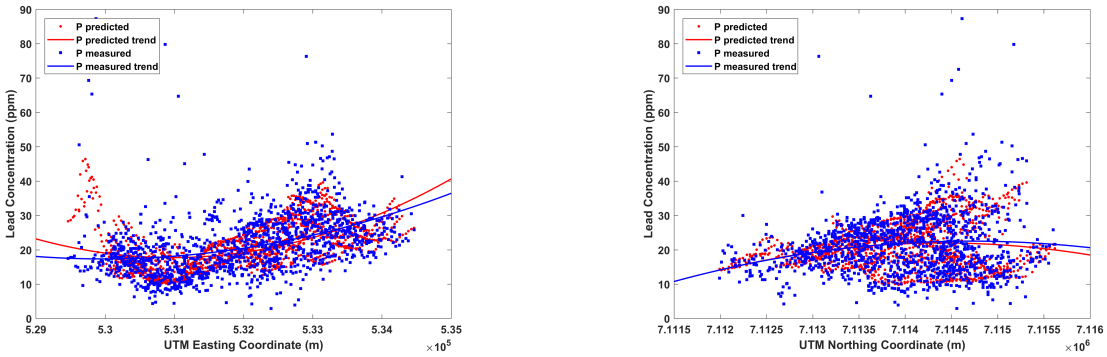


(D1)

Figure B-31, Easting and Northing coordinate axis transect plots of y-measured vs. y-predicted for Data reduction increments B-D for subset 1

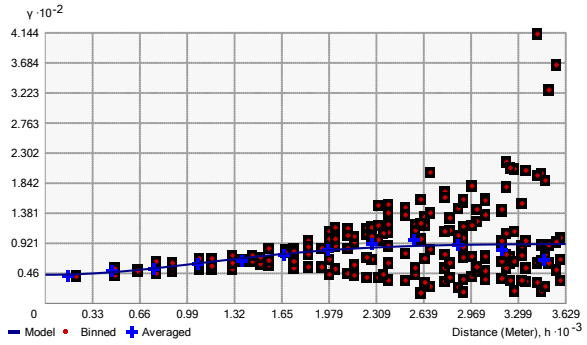


(E1)

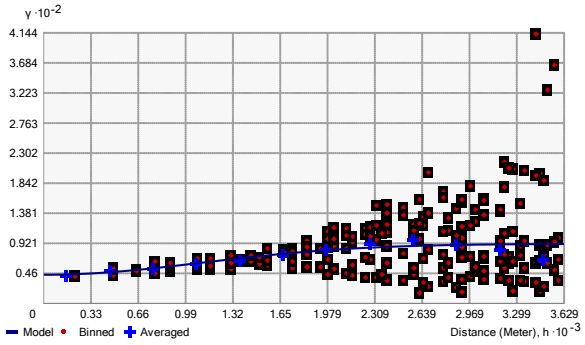


(F1)

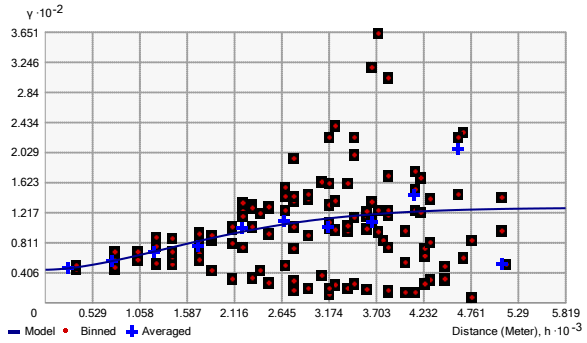
Figure B-32, Easting and Northing coordinate axis transect plots of y-measured vs. y-predicted for Data reduction increments E-F for subset 1



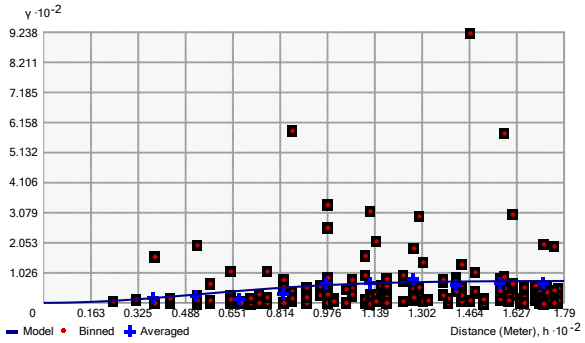
(A)



(C)



(D)



(E)

Figure B-33, Binned experimental variograms for OK, from the Geostatistical Analysis Tool in ArcMap 10.1.2 (ESRI, 2011), for data reduction increments A, C, D, and E for subset 1