OPTIMISING INTERPOLATION AS A TOOL FOR USE IN SOIL PROPERTY MAPPING

By

B. Mtshawu

(2005028077)

Dissertation in partial fulfilment of the requirements

For the degree MSc Geography

Department of Geography

Faculty of Natural and Agricultural Sciences

University of the Free State

BLOEMFONTEIN

Supervisor: Dr C.H. Barker

2014

ABSTRACT

OPTIMISING INTERPOLATION AS A TOOL FOR USE IN SOIL PROPERTY MAPPING

Inverse distance weighting (IDW) and kriging are robust and widely used estimation techniques in earth sciences (soil science). Variance of Kriging is often proposed as a statistical technique with superior mathematical properties such as a minimum error variance. However, the robustness and simplicity of IDW motivate its continued use. This research aims to compare the two interpolation techniques (Inverse Distance Weighting and Kriging), as well as to evaluate the effect of sampling density on mapping accuracy of soil properties with diverse spatial structure and diverse variability in a quest to improve interpolation quality for soil chemical property mapping.

The comparison of these interpolation methods is achieved using the total error of cross-validation and validation statistics. Mean Prediction Error and Root Mean Square Error are calculated and combined to determine which interpolator produced the lowest total error. The interpolator that produced the lowest total error portrays the most accurate soil property predictions of the study area.

The finding of this study strongly suggests that the accuracy achieved in mapping soil properties strongly depends on the spatial structure of the data. This was clearly visible, in that, when the subset training data set was decreased, the total error increased. The results also confirmed that systematic sampling pattern provides more accurate results than random sampling pattern. The overall results obtained from the comparison of the two applied interpolation methods indicated that Kriging was the most suitable method for prediction and mapping the spatial distribution of soil chemical properties in this study area.

DECLARATION

I declare that OPTIMISING INTERPOLATION AS A TOOL FOR USE IN SOIL PROPERTY MAPPING is my own work, and that it has not been submitted for any degree or examination in this or any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

BABALWA MTSHAWU 31 January 2014

Signed ……..

ACKNOWLEDGMENTS

What seemed to be a never-ending endeavor has finally come to a satisfying finish only because of the unselfish contributions of so many. I would like to express my appreciation to my supervisor, Dr C.H. Barker, for his time, encouragement and criticism over the last three years, as well as for reading my numerous revisions and his help in making some sense of the confusions. I owe a great deal of gratitude to Dr Le Roux and his team for their guidance and support. And finally, the most special thanks goes to my parents, and numerous friends who endured this long process with me, always offering support and love. Words cannot explain the appreciation I have for all that you have given me.

This thesis is dedicated to Uminathi O. Mtshawu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

LIST OF APPENDICES

Chapter 1: Introduction

1.1 Motivation of the study

Spatial prediction of both physical and chemical soil properties is becoming a common topic in soil science research. This is intensified by the recent development of technological tools that allow spatial distribution patterns of soil to be easily modelled for the use in environmental management and digital soil mapping (Minasny & McBratney, 2007). Soil properties are not uniform: they greatly depend on factors such as soil type, climate, topography, anthropogenic activities and vegetation, all of which affect the spatial distribution patterns of soil (Wenjiao *et al.,* 2009). Whether or not these soil properties are scattered, converting data into maps that show variation within the region is of prime importance to most earth scientists. According to Oliver (1990), observations that form the basis for analysis with geographical information systems are generally not continuous. Only a given number of the possible infinite locations can be enumerated. Even in situations where near continuous information exists, the amount of information needs to be reduced so as to allow for efficient data handling and analysis in a limited time frame.

Geographical Information Systems (GIS) are increasingly used for the prediction of the spatial distribution of soil chemical properties (Hartkamp *et al.,* 1999). According to Wenjiao *et al.* (2009), a useful tool for soil property mapping in spatial modeling is interpolation. However, the effectiveness of this tool greatly relies on the accuracy of the specific spatial interpolation method which is used to describe the spatial variability of soil properties. It is crucial to study interpolation methods in the quest to find one with high accuracy and improve interpolation quality for soil property mapping.

The interest in spatial distribution of soil, with the wide usage of GIS and the variety of interpolation techniques it provides, has made the comparative investigation of these techniques possible. This comparison is important, because it shows the applicability as well as the accuracy of these interpolation techniques. Although there have been several projects that investigated a substantial number of spatial interpolation techniques, only a small number have provided a comparison and a superiority of some techniques over others.

Interpolation techniques frequently used in soil science include, Inverse Distance Weighting (IDW) and kriging (Kravchenko & Bullock, 1999). Both methods are based on Matheron's

regionalized variable theory. (The theory provides a convenient summary on soil variability in a form of a semi variogram (Wenjiao *et al*., 2009)).

Interpolation methods can be significantly affected by a few factors. These include: spatial structure and variability of the data, the choice of variogram model, search radius, and the number of closest neighboring points used for estimation (Kravchenko & Bullock, 1999). The bases of many geostatistical studies are the sampling scheme, due to the fact that it plays a major role in the quality of spatial predictions. Van Groenigen (2000), states that the most important contributions in spatial modeling are the discussions on spatial sampling for interpolation purposes, aimed at optimal spacing of a regular grid for optimum interpolation accuracy.

This dissertation has two primary objectives. The first objective is to determine the effectiveness of the kriging and IDW interpolation techniques. This will be accomplished by comparing the total error of cross-validation and validation statistics. Soil sampling plays a pivotal role in the quality of spatial prediction, and thus, the second objective is to improve spatial sampling structure for interpolation purposes. This will be achieved by comparing grid sampling and random sampling techniques. The relationship between the input data needed to generate a soil map and the map accuracy will also be examined. The results will determine which interpolator produces the least error using only the subset of the entire data set, as well as which interpolator produces the least error using the entire data set, using both random and grid sampling.

1.2 Specific aims and objectives

The main aim of this thesis is to evaluate the reliability of soil property distribution maps using IDW and kriging interpolation methods. More specific objectives are:

- To compare soil sampling techniques for interpolation by IDW and kriging.
- To compare IDW and kriging interpolation method in order to determine the optimum method for mapping soil properties.
- To analyze the relation between statistical properties of the data and performance of the interpolation technique.
- To assess the accuracy and effectiveness of soil property maps produced by IDW and kriging interpolation methods.
- To suggest the most appropriate technique (if any) for soil property mapping.

Chapter 2: Literature review

2.1 Overview

This chapter presents an introduction to the concepts of soil, soil distribution patterns, as well as soil sampling methods. A discussion on geostatistics is included, supported by the regionalized variable theory. Also discussed are the two spatial interpolation methods (kriging and IDW) used in this research.

2.2 Soil

Bridges (1997: 4) defines soil as "a biologically active, structured and porous medium that has developed over the years on the earth's surface." Scull *et al.* (2003) view soil as a fundamental natural resource which serves as the basis for agriculture and plays an essential role in the biophysical and biochemical functioning of the planet.

Soil means different things to different groups of people. For example, to a mining engineer, soil is the debris covering the rock or minerals which must be quarried; it is a nuisance and must be removed. To the average home owners, the concept of soil is only understood in terms of types that cling to the shoes and eventually to the carpet. A farmer, along with some homeowners, views soil as indispensable, because it is looked upon as a habitat for plants, and a living is made from soil, thereby forcing the farmer to pay more attention to the soil's characteristics.

The science of pedology on the other hand emphasizes the study of soil as a natural phenomenon on the surface of the earth; therefore, a pedologist is interested in the appearance of the soil, its mode of formation, its physical, chemical and biological composition and distribution (Bridges, 1997). This dissertation will look at soil from a pedologist's viewpoint.

2.3 The nature of soil and spatial variation

Understanding the spatial distribution of soil and the complexity of its chemical and physical properties is crucial for managing and maintaining a productive society (i.e. food security). According to Oliver *et al.* (1990), most natural properties on, above and below the earth's surface vary continuously. This variability is a result of the combined effect of physical, chemical and biological processes that occur to different entities at different intensities and scales (Santra *et al.*, 2008). It is this complexity and variability of soil patterns in landscapes that complicates

the already laborious processes of collecting and presenting soil survey data (Oliver *et al.*, 1990).

The nature of spatial data and the only way with which variability of soils can be captured depends on a method that consists of three steps (Scull *et al.*, 2003). The first step is the direct observation of soil profile characteristics and auxiliary data. The second step is the observation of soil attributes which is incorporated into an accepted conceptual model that is used to infer soil variation. The last step involves applying the conceptual model to a survey area to predict soil variation at unobserved sites.

Generally, observations can only be made at a finite number of infinite possible locations. Sampling plays a crucial role in accurately predicting spatial data. According to Yen *et al*. (2007), to expect reliable prediction and produce accurate maps using interpolation methods, one needs an appropriate sampling method that will forecast to the scale and range of spatial variation of that particular area, otherwise the sampling might be more intensive than necessary or too sparse to provide spatially correlated data for any method of spatial variation. Karydas *et al.* (2009), specify that samples should be taken evenly over the study site and that any kind of randomness can lead to uneven distribution of the samples.

2.4 Sampling for the purpose of representing spatial variation

2.4.1 Sample design

Spatial sampling is based on the idea that the variable under study is a stochastic process (Brus & Gruijter, 1997). If the same locations were sampled multiple times, multiple values would result, and could be assembled as probability distributions (Goovaerts, 1997). This is known as regionalized variable theory, which assumes that the spatial variation of any variable can be expressed as the sum of three major components: the first being, a trend or constant mean; the second is a random but spatially correlated component (regionalized variable); and the last is the spatially uncorrelated random noise, or residual error (Burrough & McDonnell, 1998).

Spatial sampling can be defined as those sampling procedures that incorporate the assumption that the variable is stochastic, and rely on estimates of the co-variance in previously collected data to drive sampling campaigns (Mason *et al*., 1988). Both the random and spatial

approaches can produce satisfactorily independent samples for statistical analysis and spatial prediction.

Random sampling has benefits in terms of producing strictly valid, unbiased sample data collection, which is sometimes required for legal or regulatory purposes (Mason *et al*., 1992). However, the lack of bias comes at a cost. Truly random surveys ignore all expert opinion in the sampling design, leading to much greater sampling effort, and resulting in more samples than necessary in some areas and too few in others.

The geostatistical approach rests on several assumptions which are difficult to prove but offers much more flexibility in terms of sample distribution (Brus & Gruijter, 1997). This is often desirable as it can simplify fieldwork logistics, permits spatial analysis, and encourages the incorporation of expert knowledge into the analysis process. The primary concern for the spatial approach is that sampling is adequate to estimate the co-variance structure of the variable of interest (Brus & Gruijter, 1997). Common sampling layouts are discussed below, grouped into random and systematic methods.

It is important to note that even random sampling can lead to samples that are spatially autocorrelated, resulting in a well-known ecological problem called pseudoreplication (Levin, 1992). If an inferential approach is preferred to a geostatistical approach, then it is important to ensure that samples are spaced so that they are not spatially autocorrelated.

Another option is to include the level of spatial autocorrelation as an independent variable in the inferential methodology. This method, commonly referred to as autoregressive or autologistic modelling (Klute *et al.,* 2002), includes a co-variate that allows spatial autocorrelation to influence the prediction. In this case, the co-variate must theoretically be replacing some known physical function.

2.4.2 Random Sampling

The basis of most sampling plans in spatial sampling is the concept of random or probabilistic selection of the sample to be collected and the subsample that is to be analyzed (Mason *et al*., 1992). In random sampling of a site, each sample point within the site must have an equal probability of being selected (Shaffer *et al*., 1979). The same can be said for the selection of

particles within a sample, meaning; each and every particle within the sample must have an equal chance of being selected.

According to South (1982), a properly designed sampling plan based upon the laws of probability provides means of making decisions that have a sound basis and are not likely to be biased. The samples collected randomly often lead to problems. There is no basis for evaluating the validity of the sample, nor is there any means for using these samples in arriving at a sound decision with regards to the site (South, 1982).

The potential for bias introduced by the person taking the sample is great and unknown (Noyes, 2009). These samples, if treated properly, can provide insight into what chemicals may be present on a site, where particular activities have occurred, and the potential source of the pollutant. These deterministic samples are random samples collected for a particular reason. Mason *et al.* (1992) refers to these samples as “purposive samples” in that they are based solely on the collector’s choice of which units are to be collected or analyzed. They are not samples but are, in reality, only specimens. Any specimen that is submitted to the laboratory should be identified in the field records as such. This prevents the sample from being treated in the same manner as those samples that are collected by some probabilistic method (Mason *et al*., 1992) (see Figure 1).

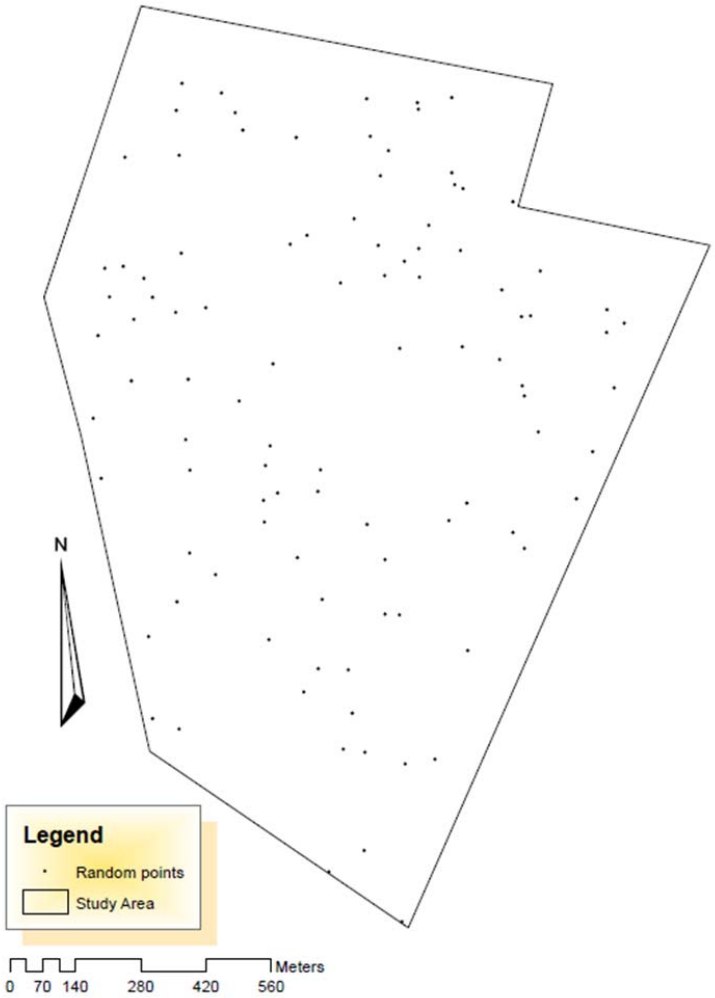


Figure 1: An example of random sampling

2.4.3 Systematic sampling

The various methods of systematic sampling are similar in that once the number and spacing of samples is determined, the distribution of the entire sample is known (Holmes *et al.,* 2004). The grid origin, or rather the starting point, of a systematic sample is drawn randomly. According to Snedecor and Cachran (1989), this method has two advantages over random sampling: the first being that it is easier to design, since only one random number needs to be chosen, and it guarantees that the measurements are evenly spread over the area of interest. The second is that, systematic sampling often gives more accurate estimates than simple random sampling, except in very large homogeneous regions (Dutilleul, 1993). There are also disadvantages. The assurance of intervals calculated from regularly sampled data in space or time for the overall

population estimate may be unreliable and if there is a natural periodic variation in the phenomenon of interest that corresponds with the sampling interval, it may go undetected (Atkinson, 1997). In addition, if the patch scale is much smaller than the sample spacing, the spatial autocorrelation and structure of the patches cannot be determined.

The distribution of the samples in this paper will be defined by the grid systematic sampling. This is defined as a regular, square network of sampling points – ideally, randomly oriented with a randomly selected origin (Holmes *et al.,* 2004*).* (Figure 2).

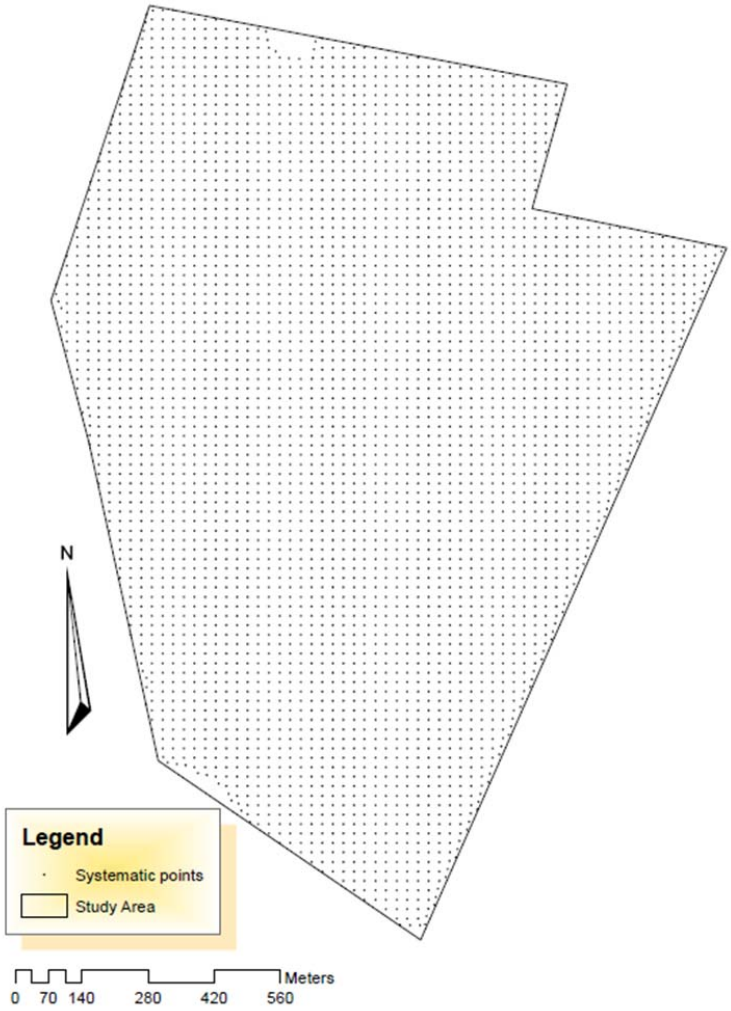


Figure 2: Systematic grid sampling

2.5 The representation of spatial variation in soils

Heuvelink *et al.* (2000), state that there are two principal approaches that represent spatial variation in soils, the first being the earlier discrepancies that has its origins in views of old taxonomies. It is also refers to the traditional method of characterizing soil properties. It does this by breaking down the landscape into discrete regions, to which each is assigned a class (Cadell, n.d.). The boundaries of these soil variations would be fixed lines across regions where the observation suggests the greatest change to have occurred (Figure 3). Thus inside each region, it is assumed that the soil is generally homogeneous.

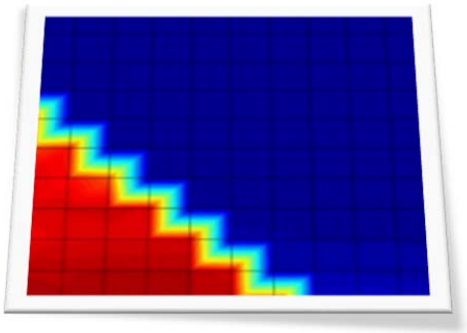


Figure 3: Example of a discrete classification (Zhong *et al*., 2012)

The second principle is the continuous classification (Figure 4), which sees soil as a collection of continuous variables that must be described according to their variation over the land (Heuvelink & Webster, 2000). This approach is quantitative, and has a statistical advantage (Cadell, n.d.) due to the fact that it views soil as an ever-changing medium and represents soil as a continuous surface. This method is statistically complicated and the calculations are very intensive (Cadell, n.d.).

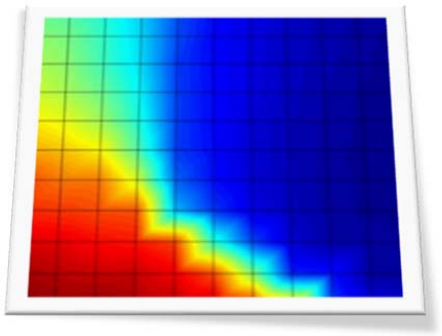


Figure 4: Example of a continuous classification (Zhong *et al*., 2012).

Both of these approaches can be incorporated to form a general model of spatial variation (McBratney, 1992). Recent developments in geostatistics have attempted, with some success, to combine the two approaches into a model that is a more realistic representation of the real world (Heuvelink & Webster, 2000). Hence, the term variation describes actuality rather than variability.

2.6 GIS and Geostatistics

According to Burrough (2001), GIS is based on two concepts: the first being automated map making, and the other being facilitating the comparison of data on thematic maps. GIS is a computerized database management system, capable of assembling, storing, manipulating, and displaying geographically referenced information.

Cressie (1968), defines geostatistics as a subset of traditional statistics that deals with spatial data and accounts for spatial autocorrelation using spatial interpolators. The concept is based on the theory of regionalized variables (Chilés & Delfiner, 1999). Geostatistical methods have been used in predictive soil mapping to spatially interpolate soil properties (Cressie, 1968).

According to Burrough (2001), geostatistics addresses the need to make predictions of sampled attributes at unsampled locations from sparse data. GIS can serve geostatistics by assisting with the geo-registration of data, facilitating spatial exploratory data analysis, providing a spatial context for interpolation as well as providing effective, easy to use interpolation tools for visualization. Scull *et al*. (2003), argue that the value of geostatistics for GIS lies in the provision

of reliable interpolation, up-scaling, and generalization methods with known errors, as well as the provision of multiple realizations of spatial patterns that can be used in environmental modelling.

2.7 Regionalized variable theory

Henly (1981), views the regionalized variable theory as statistics of a particular type of variable, which differs from an ordinary scalar random variable, as well as its usual distribution parameters, in that it has a defined spatial location. Oliver *et al*. (1989), suggests that there are two realizations when it comes to regionalized variables which differ in spatial location – this is a non-zero correlation, in contrast to an ordinary scalar random variable with which its successive realizations are uncorrelated.

Henly (1981), gives a brief explanation of the theory, stating that the regionalized variable can be defined as an ordinary random function of a probability distribution law which it follows, however, it must be normally distributed with a mean and a variance. If this random function fits one of the standard statistical distributions, it will be completely characterized by a small number of such parameters. Cressie (1986), states that in a case of a continuous variable, it is often assumed that the observations represent a particular value of a random function which is normally distributed. In a case of an ordinary random function however, the spatial location is relevant and the most accurate prediction which can be made for any given observation is that which is controlled by the form of the distribution. For the normal distribution, it will be the arithmetic mean.

According to Oliver (1990), the variation from place to place of most soil properties is usually so unpredictable that no simple mathematical expression can describe it. However, most properties appear to be random variables rather than mathematical, but, most variations are not entirely erratic, there is some spatial structure. Regionalized variable theory takes the different aspects of spatial properties into account.

2.8 Interpolation

Spatial interpolation is defined as a procedure for estimating the value of properties at unsampled sites within an area covered by existing observations (Waters, 1988). The primary assumption of spatial interpolation is that points near each other are more alike than those

farther apart (Smith *et al.,* 2013). The principle underlying spatial interpolation is known as the First Law of geography. Formulated by Waldo Tobler, this law states that everything is related to everything else, but near things are more related than distant things (Waters, 1988). The formal property that measures the degree to which near and distant things are related is spatial autocorrelation (Smith *et al.,* 2013). Most interpolation methods apply spatial autocorrelation by giving near sample points more importance than those farther away.

Interpolation methods allow the user to control the number of sample points used to estimate cell values (Waters, 1988). The distance to each sample point varies depending on the distribution of points. The sample size can also be controlled by defining a search radius (Waters, 1988). Like controlling the number of sample points, the number of sample points found within a search radius can vary depending on how the points are distributed.

The physical, geographic barriers that exist in the landscape, such as cliffs or rivers, present a particular challenge when trying to model a surface using interpolation (Waters, 1988). Most interpolators attempt to smooth over these differences by incorporating and averaging values on both sides of the barrier (Goodchild & Lam, 1980). The Inverse Distance Weighted method allows one to include barriers in the analysis. The barrier prevents the interpolator from using samples points on one side of it.

This thesis will only be dealing with IDW and kriging interpolation methods. The reason for investigating these two interpolation methods is that they are similar. Kriging and IDW are local interpolators, and they use weights surrounding measured values to derive a prediction for an unmeasured location (Goodchild & Lam, 1980). There is, however, a principal difference between kriging and IDW. Kriging is a stochastic interpolator, which means that information about the spatial structure of the data is used to predict the value of an unsampled location. IDW, on the other hand, is a deterministic interpolator that uses a mathematical formula to calculate the value of an unsampled location (Goodchild & Lam, 1980).

2.8.1 Kriging

Kriging is a local interpolation technique that uses a geostatistical method developed by G. Matheron and D.G. Krige (Burrough & McDonnell, 1998). Kriging is a local interpolator, because, it only uses the information in the vicinity of the point being estimated, it is exact, in

that, the predicted values at the points for which data values are known will be the known values and it is stochastic, because it provides probabilistic estimates.

Kriging is a two-step process that incorporates random variation in the interpolated surface and also provides standard error of predictions (Johnston *et al.,* 2001). According to Burrough & McDonnell (1998: 133), "Regionalized variable theory assumes that spatial variation of any variable can be expressed as the sum of three components." The first is the standard component with a constant mean; the second component is a random, but spatially correlated component, also known as the regionalized variable; and the last component is the residual component also referred to as the error.

Discussed below are a set of formulas provided by Burrough & McDonnell (1998) to express these three assumptions:

The value of a random variable Z at x is given as:

$Z(x) = m(x) + \varepsilon'(x) + \varepsilon''$

Where m (x) is the structural function describing the structural component, $\varepsilon'(x)$ is the stochastic but structurally autocorrolated residual from m (x) and ε'' is the residual component having a normal distribution with a mean 0 and variance σ^2.

The first step is to decide on a suitable function for m (x). This can be thought of as a flat surface with no trend. The mean value of m (x) is the mean value within the sample area, therefore the difference in the values for two points x and $x + h$ (where h is the distance between points) is zero.

$E[Z(x) - Z(x + Z(x + h)] = 0$

The variance of the differences is then assumed to be a function of the distance between the points.

$E[\{Z(x) - Z(x + h)\}^2] = E[\{\varepsilon'(x) - \varepsilon'(x + h)\}^2] = 2\gamma(h)$

Where $\gamma(h)$ is known as the semi-variance. Under these two assumptions (i.e. the stationary of difference and stationary in the variance difference), the original model can be expressed as:

$Z(x) = m(x) + \gamma(h) + \varepsilon''$

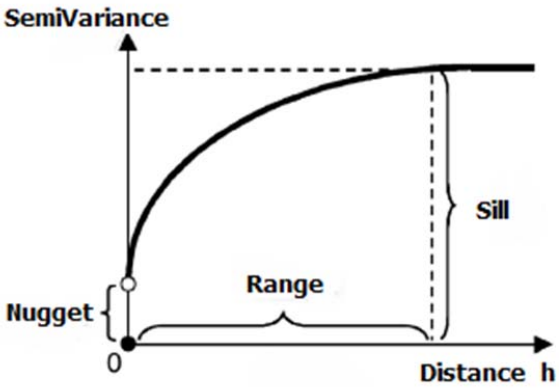
The semi-variance can be estimated from the sample data, using the formula:

$$y(h) = \frac{1}{2n}\sum_{i=1}^{n}\{Z(x_i) - Z(x_i + h)\}^2$$

There are many forms of kriging, but all are firmly grounded on the theory discussed above (Longley *et al*., 2011). This thesis will however be dealing with ordinary kriging, which uses the exact theory discussed above.

2.8.2 Semi-variogram

The semi-variogram is used in kriging to develop a prediction of expected difference in values between pairs of data with similar orientation (Collins, 1995). The semi-variogram is a representation of the average rate of change of a property with distance (Lam, 1983).



Nugget: Noise

Sill: Maximum variance

Range: Distance of correlation is zero

Figure 5. Semi-variogram (Burrough & McDonnell, 1998)

The semi-variogram is important because it provides all the information needed about a regionalized variable, including the size of the zone of influence around the sample, the isotropic or anisotropic nature of the variable, and the consistency of the variable through space (Cressie, 1993).

2.8.3 Inverse Distance Weighting (IDW)

IDW is a deterministic interpolation method in which values at unsampled points are calculated from known points using a weight function in a search neighborhood (Longley *et al.*, 2011).

This interpolation method estimates the data value for each point by calculating a distance weighted average of points within the reach radius (Burrough & McDonnell, 1998). This

technique is not only known to be deterministic, it is said to be local and exact (Johnston *et al.*, 2001). IDW is one of the simpler interpolation methods in that it does not require pre-modeling like kriging (Burrough & McDonnell, 1998). The formula and the weighting function for IDW as provided by Burrough & McDonnell (1998) can be seen in the equations below.

The general formula is:

$$\hat{z}(x_0) = \sum_{i=1}^{n} \lambda_i . z(x_i)$$

Where $z(x_i)$ are the data values for the *n* points ($x_{1.....}x_n$) within the search radius and λ_i are the weights to be applied to the data values for each point. The constraint, however, is that weights must add up to 10. Burrough & McDonnell (1998: 117), define weights as "some function of the distance between the point for which the estimate is being made and the sample points." The IDW predictor, which is the most common function of IDW, is then expressed in the following formula:

$$\hat{z}(x_0) = \sum_{i=1}^{n} [\frac{d_{ij}^{-r}}{\sum_{i-1}^{n} d_{ij}^{-r}} . z(x_i)]$$

Where *j* is the point whose value is being interpolated, d_{ij} is the distance from point *j* to sample i, and r is a random value which can be selected by the researcher. If r is set to be 1, then the interpolation becomes a simple linear interpolation (Burrough & McDonnell, 1998). In most cases, r is set to be equal to 2, "thus, the influence of each sample point is in proportion to the square root of its distance from the point to be interpolated" (Longley *et al.*, 2001). r can be set to higher values if required, higher values give a much higher weight to the nearer sample point (Burrough & McDonnell, 1998).

2.8.4 Validation and cross-validation

The two most popular methods for determining spatial interpolation accuracy are validation and cross-validation (Cressie, 1993). Cross-validation is described as the process of removing parts of the data and interpolating the remaining data to predict the removed data set (Johnston *et al*., 2001). Similarly, validation uses a test and training data set (Johnston *et al*., 2001). Here, a percentage of the data points are removed and used as the test data set, while the remaining data points, known as the training data set are used to predict the removed points. Each interpolation technique is then compared on the basis of mean prediction error (MPE), and root

mean square error (RMSE). The main aim of this comparative effort is to determine which interpolator produces the lowest total error (TE), which is the combination of RMSE and MPE (Krivochko & Bullock, 1999). According to Krivochko & Bullock (1999), the interpolator that produced the lowest total error portrays the most accurate soil property predictions of the study area.

2.9 Conclusion

With new advances in earth sciences and the availability of GIS technologies, it is now possible to accurately divide a field into smaller systematic grids that can be sampled individually to cover the high variability of soils. It is clear from the discussions in this chapter that inconsistencies made during random sampling may lead to an inaccurate representation of the area being studied. Soil test results from each grid cell can be used to prepare soil chemical availability maps. Remediation of pollutants, variability rate fertilizer application as well as lime applications can then be based on these maps. The effectiveness and accuracy of these soil property maps do not only depend on the sampling method, but on the type of interpolation method used. Converting discrete sampled data into a continuous surface requires a clear understanding of the interpolation technique to be used. This chapter discussed the theoretical background and performance of the two interpolation techniques being compared in this study.

Chapter 3: Description of the study area

3.1 Overview

Chapters 1 and 2 provided insight into the defined problem, a brief introduction to soil, and a background of the interpolation techniques used in this thesis. This chapter describes the research study area in detail, including topology, climatic conditions, geology, soils, as well as soil hydrology.

3.2 The Study area

This survey was carried out in the small farming community of Reipan, which is located at 26˚59.03’ 65” S and 25˚19.21’ 74” E. Reipan is situated 60 kilometers north-east of Vryburg, in the Naledi local municipality of Dr Ruth Segomotsi Mampati District Municipality, North West Province, South Africa.

Figure 6: See PDF document provided, it will be inserted here, as an A3 document due to loss of resolution at A4

Figure 7: Spot image of the study area (CSIR, SAC, 2012).

3.3 The physical environment

3.3.1 Topography

The North West province is known to have the most uniform terrain of all the provinces in South Africa, with an altitude ranging between 920 and 1782 meters above sea level (Masigo & Matshego, 2002). The central and western regions are characterized by flat or gently undulating plains. Dunes associated with the arid environment of the Kalahari desert occur in the far western region (Masigo & Matshego, 2002).

3.3.2 Climate

The North West province is characterized by well-defined seasons with hot summers and cool, sunny winters. The rainy season usually occurs from October to March (Masigo & Matshego, 2002). The climate and rainfall vary significantly: the more mountainous and wetter eastern region receives on average 600mm of rainfall per annum; the central region receives around 550mm rain per annum; while some areas in the drier semi-desert plains of the western Kalahari receive less than 300mm per annum (Desmet & Seymour, 2009). These figures are known to vary greatly from year to year. The North West province, therefore, has a higher average rainfall per annum than the South African average. Thus, the province has enormous potential in agriculture (Masigo & Matshego, 2002). See figure 8.

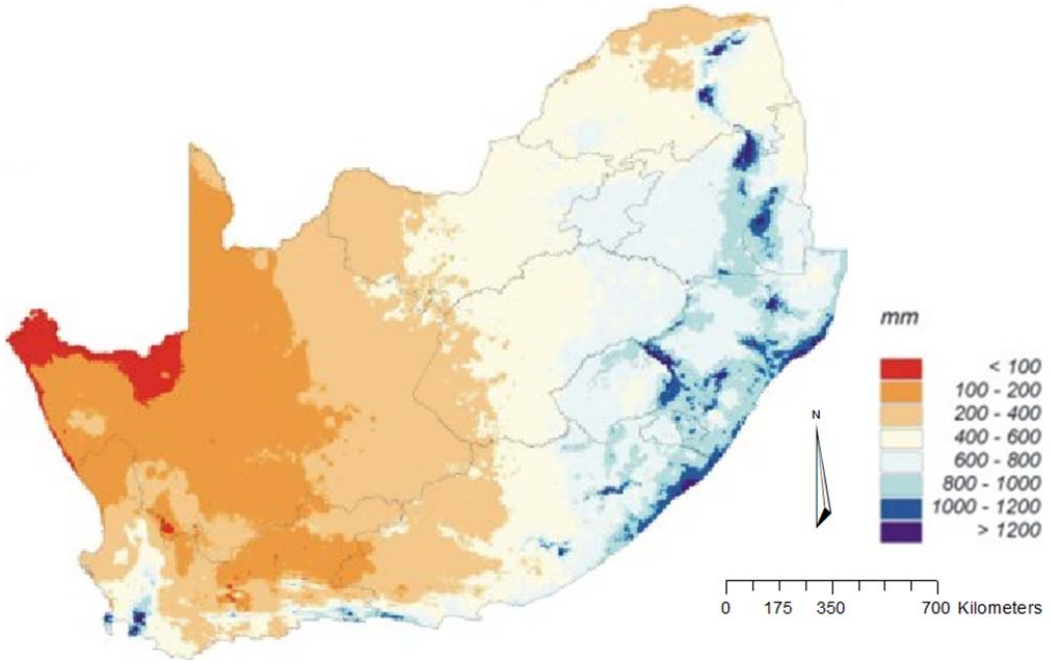


Figure 8: Map of Mean Annual Precipitation (mm) (Schulze, 1997).

Figure 8 shows a map of mean annual precipitation in South Africa. This gives an indication of areas where prolonged droughts exist because of a below low rainfall recorded over a period of a year. The most remarkable is the severe-to-extremely dry regions along the North West, Free State and Northern Cape borders, while the central regions, extending to the eastern parts of the Eastern Cape are experiencing moderate rainfall conditions.

3.3.3 Hydrology

Water is one of the most critical and limited natural resources in the North West province (Mapukule, 2009). The sources of water available in the province are surface water and groundwater, including rivers, dams, pans, wetlands and dolomite eyes fed by underground springs (Schulze, 1997). Apart from highly variable precipitation from year to year, one of the most important factors affecting surface water in the province is the highly variable but low actual runoff. Runoff as a percentage of the precipitation ranges from less than 1% in the west to approximately 7% in the eastern region. The average runoff for the province is 6%, which is below the average of 9% for Southern Africa (Schulze, 1997).

3.3.4 Geology

Geologically, the north-eastern and northern central regions of the North West province are largely dominated by an igneous rock formation as a result of the intrusion of the Bushveld complex (Mapukule, 2009). Sedimentary rocks dating back to the Quaternary period occur in the north-western corner of the province (Keyser & Du Plessis, 1993). Outcrops of granites occur in the south-eastern portion of the province and further west as far as the north-central portion of the Vryburg region (Keyser & Du Plessis, 1993).

The north-eastern portion of the Vryburg region, covering the study area, is largely made up of Ventersdorp supergroup rocks, which include Brecias, Conglomerates, Feldspars and Porphyrites (Keyser & Du Plessis, 1993), as well as low grade metamorphic rocks such as granite gneiss (de Villiers & Mangold, 2002) (Figure 9).

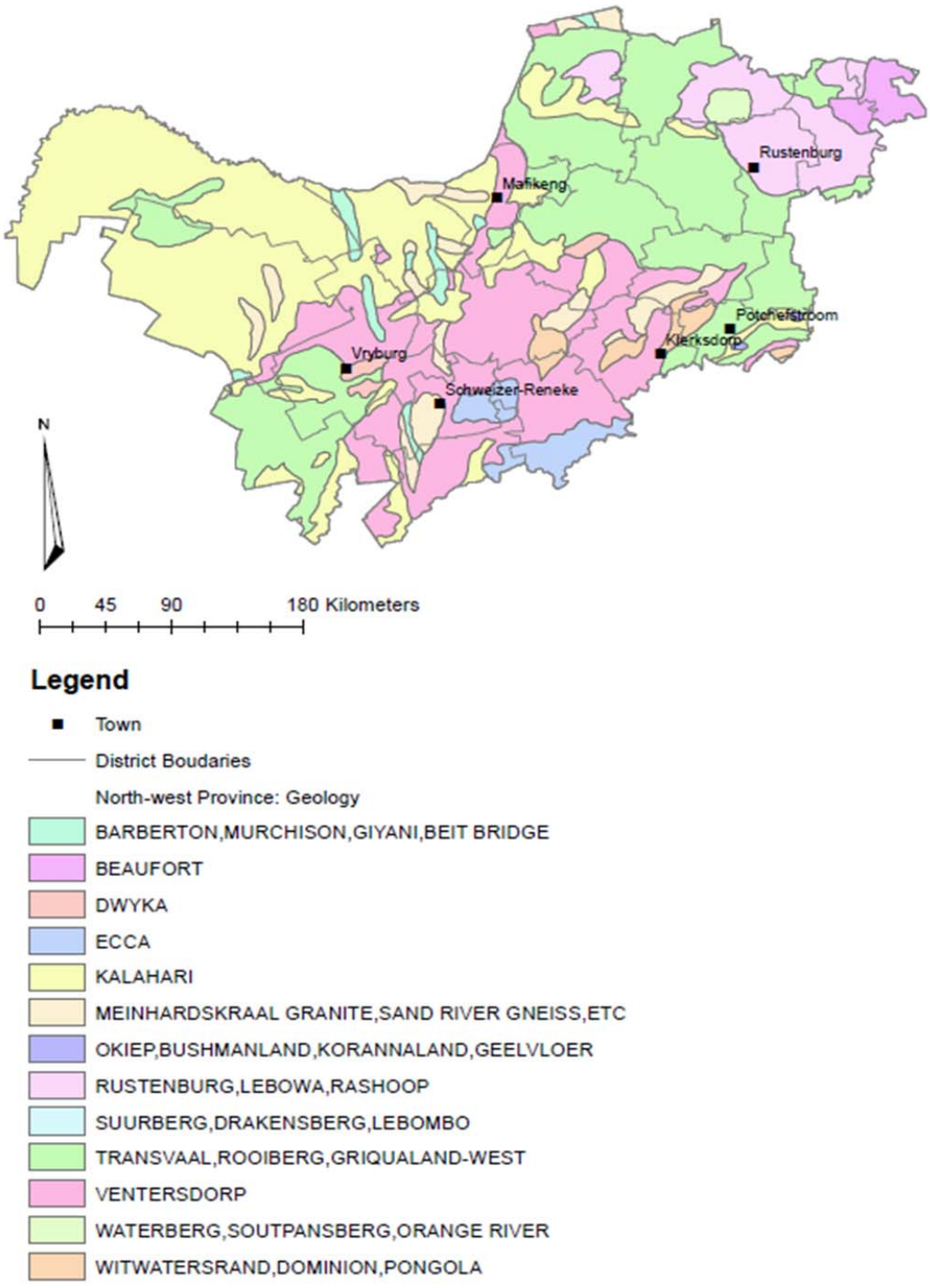


Figure 9: Simplified geological map of North West province

3.3.5 Soil

A soil map of the area can be seen in Figure 10. A very interesting soil sequence was found on the site. Predominately an unspecified material with signs of a wetness layer was found covering most of the area, at varying depths. This horizon of unspecified materials with signs of wetness is characterized by soil which has undergone iron reduction and bleaching due to prolonged saturation with water. Although unspecified material with signs of wetness soils are to some degree water impenetrable, they form under conditions of a fluctuating water table.

The soil forms dominating the study area are Pinedene and Avalon. Avalon soil forms are deep soils, and in some instances one can find a deeper hard plinthic layer. This soil form is hydrologically classified as recharge soils. Table 1 describes the role of recharge soils. The Pinedene soil form is found to have compacted layers at shallower depths. This is important, because although these layers are not water impenetrable, it does retard the flow of water through the profile and it is expected that some soil properties be retained slightly above these layers.

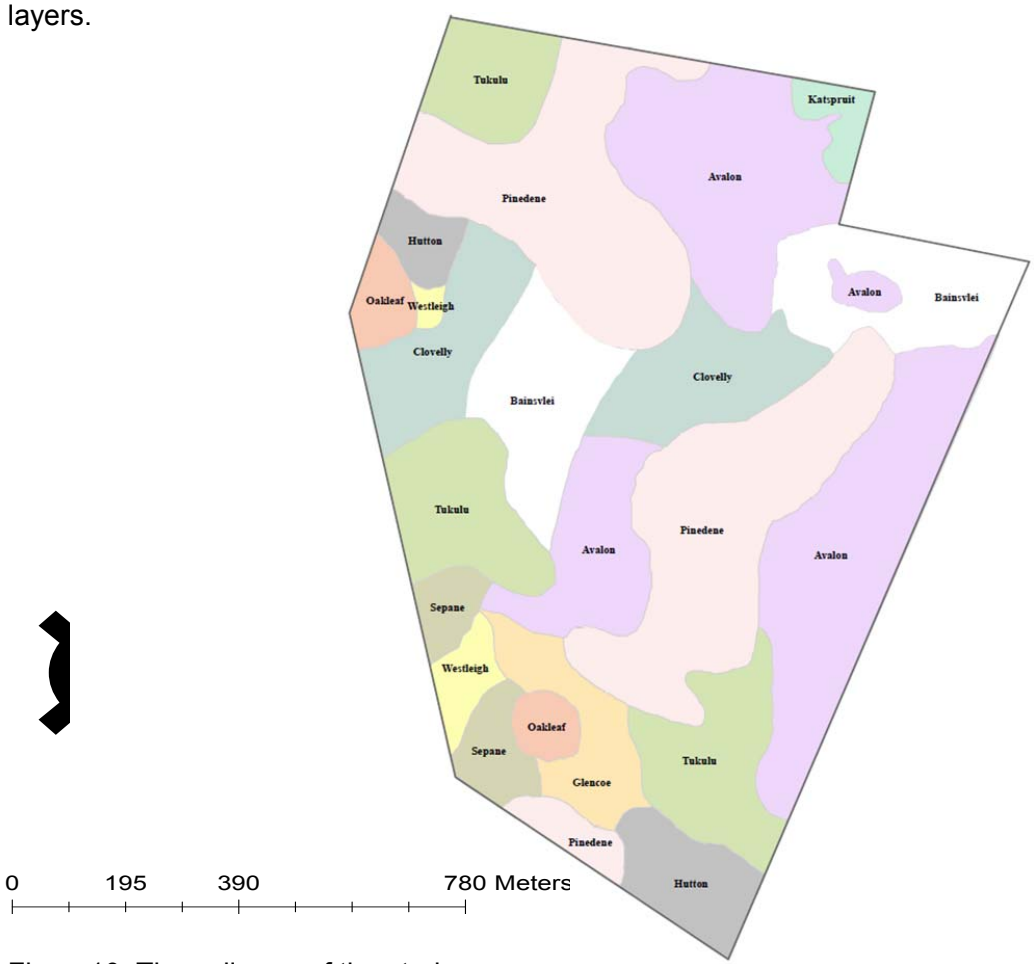


Figure10: The soil map of the study area

3.3.6 Soil Hydrology

Soils are divided into three hydrological classes, each with an expected hydrological behavior. Below is a short description of these classes.

1) Recharge soils

Recharge soils are so named because water moving through them will recharge the groundwater. The dominant flow of water through the profile is vertically downwards. Precipitation will flow vertically through the profile under gravitational forces until it reaches an impermeable layer. The water then either filters slowly through the layer into the groundwater or it will move laterally downslope on top of the layer. These soils are generally found on the crests of hill slopes, which are gentle slopes. Recharge soils are very important as they contribute largely to base flow which is a major water source to rivers and streams (Van Tol, 2008).

2) Interflow soils

In interflow soils, the dominant flow of water is horizontally through the profile. Water will infiltrate the soil and move vertically through the soil. For interflow to occur, a deeper layer must have a lower hydraulic conductivity (i.e. a clayey layer below a sandy layer or soil and rock interface) than the above layer and a slope must be present (Van Tol, 2008). The water will then stop moving vertically when it reaches the layer with the lower hydraulic conductivity and be diverted laterally in a downslope direction.

3) Responsive soils

Responsive soils carry the name due to the fact that soon after these soils are saturated, a response can be seen in the water flow of the streambed. These are soils in which infiltration does not occur and water flows on the surface. This could be due to a saturated soil profile or a very shallow soil that has a very low water holding capacity. Due to the need for saturation, responsive soils are typically found on lower positions in the landscape and concave positions where water can accumulate. Shallow responsive soils on the other hand often occur on top of hill slopes. Therefore overland flow is found on these soils when precipitation occurs.

Table 1: Soil hydrological classes of the study area

Soil Form	Diagnostic Horizon	Hydrological Soil Type	Description
Avalon	Orthic A Yellow-Brown Apedal B Soft Plinthic B	Recharge	Freely drained soil, but may have deep water impenetrable layer
Bloemdal	Orthic A Red Apedal B Unspecified material with signs of wetness	Interflow	As Avalon, but these are shallow soils and offer resistance to root and water penetration
Cloverlly	Orthic A Yellow-Brown Apedal B Unspeciefied	Interflow	Freely drained soil, but with an impenetrable C horizon layer
Glencoe	Orthic A Yellow-Brown Apedal B Hard Plinthic B	Interflow	Freely drained soil, but with shallow hard plinthic horizon
Hutton	Orthic A Red Apedal B	Recharge	Freely drained soil, but may have deep water impenetrable layer
Katspruit	Orthic A G Horizon	Responsive	Commonly found in wetlands, high clay percentage with very little drainage through the soil
Oakleaf	Orthic A Neocutanic B Unspecified	Interflow	Same as Cloverlly, without the presence of carbonates within 1500mm of the surface Neocutanic B would have qualified as diagnostic yellow-brown
Pinedene	Orthic A Yellow-Brown Apedal B Unspecified material with signs of wetness	Interflow	Weakly developed cutans, C horizon is evidence of water freely draining through A and B horizons but will not flow through the C horizon
Sepane	Orthic A Pedocutanic B Unspecified material with signs of wetness	Interflow	The stongly developed cutans found in the B horizon suggests that water flows freely in horizon A and B however the weakly developed cutans in C horizon suggests that water will not flow through C horizon
Tukulu	Orthic A Neocutanic B Unspecified material with signs of wetness	Interflow	As Pinedene, without the presence of carbonates within 1500mm of the surface Neocutanic B would have qualified as diagnostic yellow-brown Apedal B
Westleigh	Orthic A Soft plinthic B	Interflow	Soft plinthite is evidence of a fluctuating water table Water will move horizontally through this soil

3.4 Soil survey sampling

Three methods were followed with soil survey and sampling. Using the Fishnet tool in ArcMap, a 20×20m grid was constructed. Using the Calculate Geometry tool, the coordinates of each point (centroids), created by the Fishnet were extracted, these field points' xy coordinates were important for the collection of surface samples. In total 3896 sample points were created over the 153ha study area (refer to Figure 11). Samples were taken just below the surface to avoid contamination from other sources. Secondly, samples were taken using a soil auger, up to the depth of the limiting layer of that particular sample point for soil classification purposes. Lastly, the classification of these soils described in detail with special reference to morphological indications of the hydrological behavior of the soils. Soils were classified according to the South African soil classification system (Soil Classification Working Group, 1991).

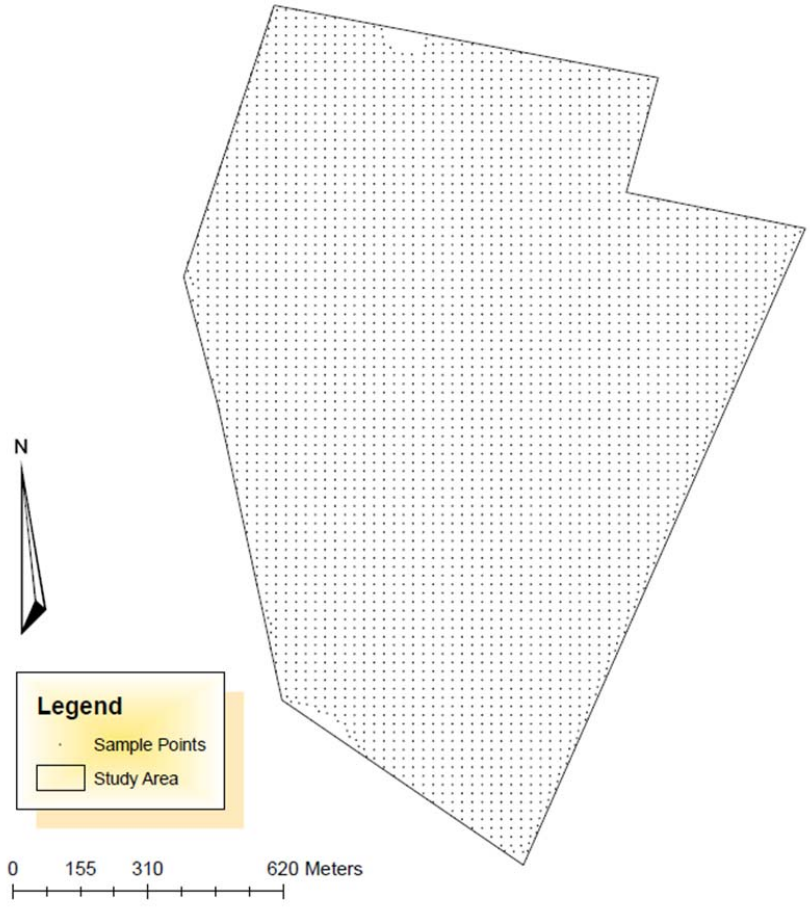


Figure 11: Sampling Points of the study area.

3.5 Conclusion

This section described the location of the study area and discussed the existing physical conditions in the area. The discussion on the conditions in the study area primary focused on the topography, climate, water resources, geology, soil, and soil hydrology.

Chapter 4: Soil analysis methodology

4.1 Overview

This chapter presents a short discussion on the soil analysis technique used in this research, as well as the calibration of the equipment used in analyzing the soil properties resulting in the study area.

The chapter is divided into three sections. The first section deals with biographic and background information on soil analysis by looking at the previously used techniques, comparing them to the Fourier transform MIR spectrometer (Bruker Optics) used in this thesis. The second section deals with the calibration and analysis of the soil samples. The third section presents the summary of the soil analysis results.

4.2 Soil analysis

The equipment used for soil characterization procedures can be very expensive¸ with processing times of 6 to 12 months, Brown *et al.* (2006). As a result, relatively few locations are fully characterized. Soil landscape models and soil maps have been constructed largely on the basis of field observation. These include: Munsell colors, hand Texturing, pH indicators and an acid reaction. Recent advances in soil analysis demonstrate that diffuse reflectance spectroscopy is a strong analytical technique suited for rapid and simultaneous analysis of biological, chemical and physical attributes of soil (Awiti *et al.*, 2008). Researchers have successfully predicted several soil fertility parameters, including organic carbon (SOC), inorganic carbon, total nitrogen (TN), cation exchange capacity (CEC), pH, potassium (K), magnesium (Mg), calcium (Ca), zinc (Zn), iron (Fe) and manganese (Mn) with various levels of prediction accuracy (Bro, 2003). According to Shepherd and Walsh (2004), Infrared spectroscopy, both near-infrared (NIR) and mid-infrared (MIR) are by far the most cost effective and producible analytical techniques available for the 21st century. The analysis of samples for this project was done with a Fourier transform MIR spectrometer (Bruker Optics, 2006) (see Figure 12).

Figure 12: Fourier transform MIR spectrometer from Bruker Optics

4.3 Infrared (IR) spectroscopy applications

IR spectroscopy has only recently been investigated for routine use, including in soil analysis and quality control in cash crops such as tea, coffee and sugar cane. The potential of IR spectroscopy has perhaps been least exploited in integrative fields such as agroforestry and landscape ecology, which includes the study of tree, crop and livestock production in farms and landscapes, and their interactions with the ecosystem (Shepherd and Walsh, 2002). According to Brown *et al.* (2005), infrared diffuse reflectance has an interconnected effect, responding to mineral composition, iron oxides, organic matter, water, carbohydrates, soluble salts and particle size distribution. Thus, properties largely determine functional capacity of soil, an example being the ability to support plant growth and hydraulic regulation (Shepherd and Walsh, 2002).

4.4 Mid-IR spectroscopy

"Mid IR spectroscopy provides richer information on soil properties" (Shepherd and Walsh, 2007: 13), this is due to the fact that essential vibrations of organic and mineral compounds are detected. For this reason, mid-IR spectroscopy is better suited for organic matter research because absorption features associated with various organic functional groups can be identified. Furthermore, mid-IR spectroscopy may provide more stable calibrations across soil types. Lastly, MIR may be advantageous where surface features are of interest, for *in situ* characterization of soil profiles, for remote sensing applications, and in precision agriculture.

4.5 MIR spectroscopy calibration

According to Bruker Optics (2006), modern analytical chemistry has changed over the last few years, due to the introduction of chemometric evaluation techniques. The term chemometrics encompasses all multivariate calibration methods used in analytical chemistry. Compared to the classical univariate calibration, this technique uses not only one spectral data point for the calibration, but the whole spectral structure (Bruker Optics, 2006). The advantage of this type of calibration is the amount of spectral information used so that even minor differences in the sample spectra can be identified.

4.5.1 Multivariate calibration techniques

Generally speaking, every quantitative analytical method aims to determine a system property (Y) quantitatively from a measured system parameter (X) (Bruker Optics, 2006). This determination requires two steps: the calibration and the analysis.

During calibration, a correlation of the measured quantity (X) and the system property (Y) is sought. This correlation is described in the calibration model:

$$y = x \cdot b$$

with the calibration function b^1 , which is often called "regression coefficient" or "b -coefficient":

$$b = (x^t \cdot x^{-1}) \cdot x^t \cdot y$$

In this equation, the parameters X and Y are written in matrix form. If they were to represent a spectroscopic measurement, for example, the spectral intensities would be written into the X matrix in rows, point by point. Each additional sample would, therefore, correspond to an additional row in the matrix. The corresponding component values would then be written into the rows of the Y matrix. T represents the transposition of the associated matrices. After the calibration, the analysis is performed. By connecting the calibration model to the measured parameter (X), the system property (Y), of an unknown sample is determined. This is depicted schematically in Figure 13.

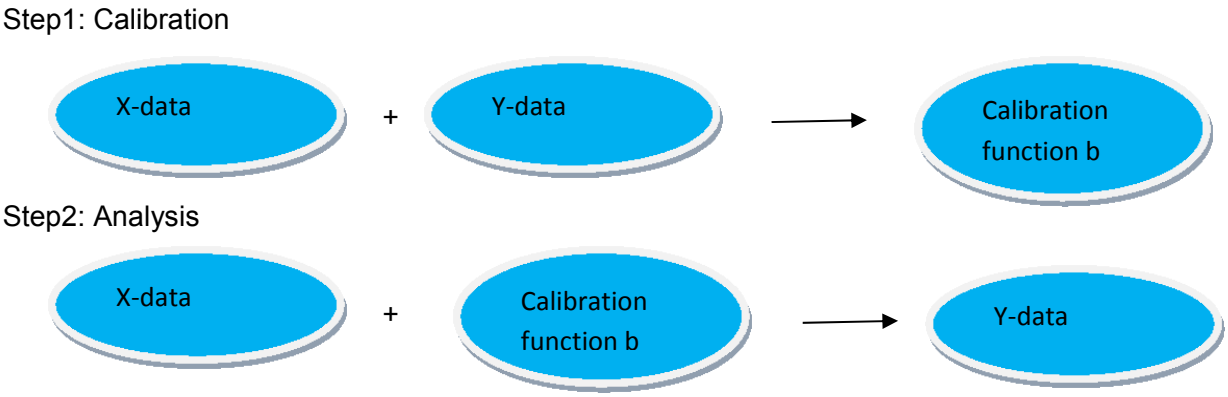


Figure 13: Schematic procedure of the quantitative determination (Bruker Optics, 2006).

Where there is a quantitative evaluation of infrared, the measured value is normally an absorption or emission of a spectrum, and the system value to determine is the concentration of the analyte. Bruker Optics (2006), further describes two methods of setting up a calibration model: firstly, univariate calibration, and secondly, the increasingly popular method of multivariate calibration.

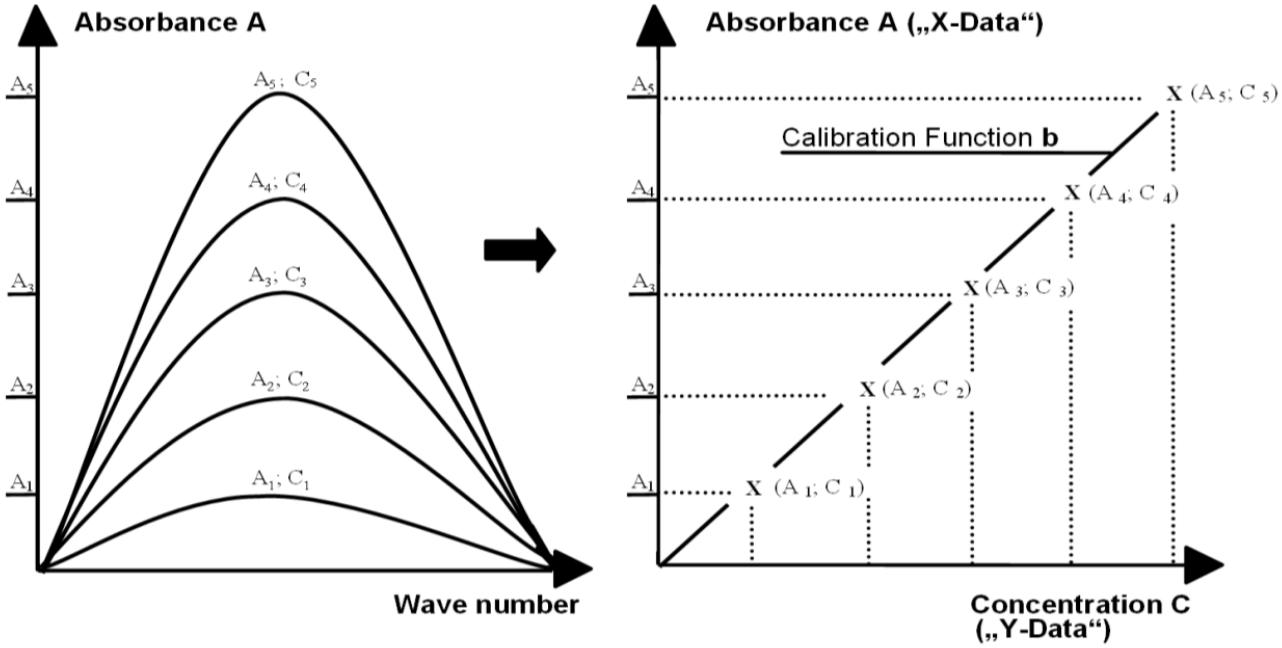


Figure 14: Calibration of absorbance spectra (Bruker Optics, 2006).

Figure 14 show a univariate evaluation of an absorption band. Five samples with the concentrations C1 to C5 were measured. These values correspond to the "*Y* values". The measurements result in five absorption values A1 to A5; the "X values".

In a univariate calibration, the absorbance values of the peak maximum are plotted versus the concentration of the analyte. The fit function calculated from the absorbance data then allows calculating the concentration from the measured absorbance values and vice versa. The analysis of a new, unknown sample is carried out by measuring it spectroscopically and determining the absorbance value *Ap* at the peak maximum. This value is then correlated with the calibration function b, which was calculated earlier, and results in the analyte value (see Figure 15).

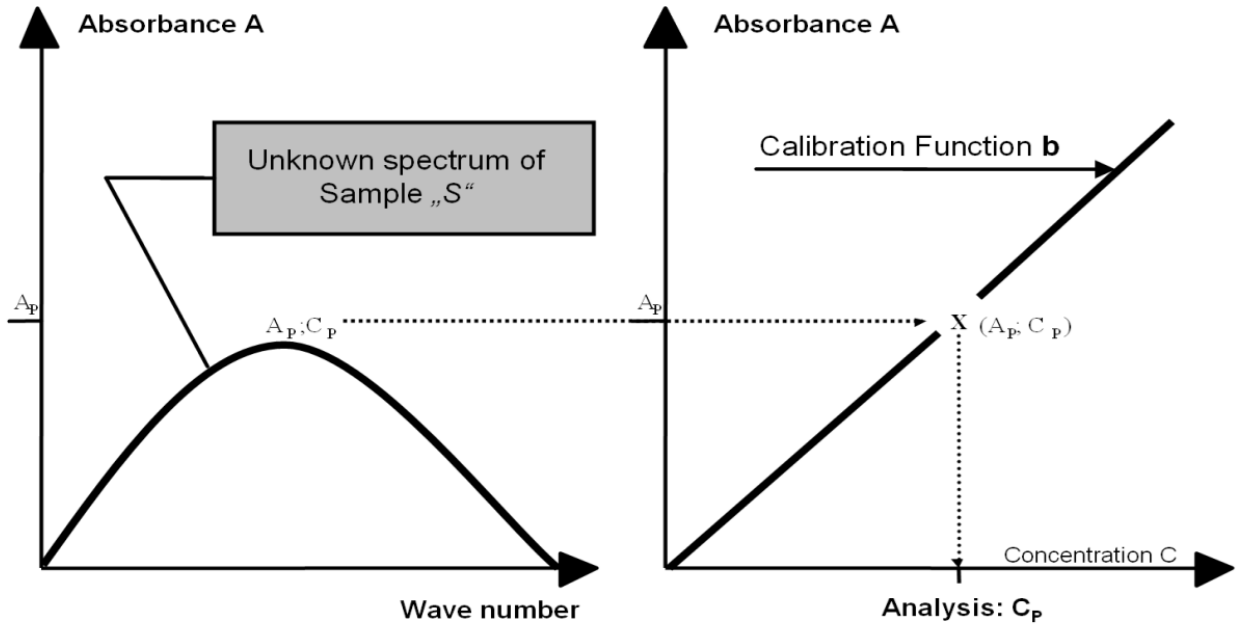


Figure 15: Analysis of absorbance spectra (Bruker Optics, 2006).

4.5.2 Partial Least Squares (PLS) Regression

PLS is a method for creating predictive models when there are many highly collinear factors (Shepherd and Walsh, 2002). It is important to understand that the emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables. When a prediction is the objective and there is no practical need to limit the number of measured factors, PLS can be a useful tool (Bro, 2003).

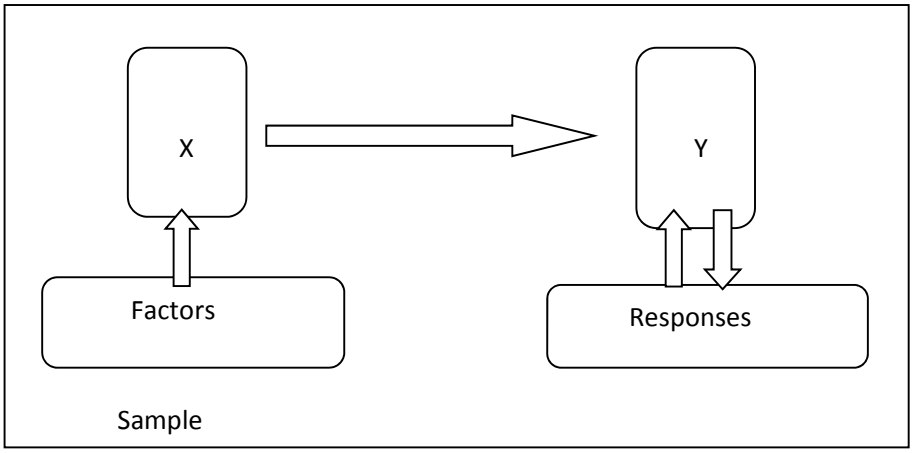


Figure 16: A schematic representation of the process of extracting latent variable X and Y from sampled factors and responses (Shephard & Walsh, 2002).

The overall goal is to use the factors to predict the responses in the sample. This is achieved indirectly by extracting latent variable X and Y from sampled factors and responses. The predicted Y scores are used to construct predictions for the responses. (See Figure 16).

According to Shepherd and Walsh (2002), the X and Y scores are chosen so that the relationship between successive pairs of scores is as strong as possible. In principle, this is like a strong form of redundancy analysis, seeking directions in the factor space that is associated with high variation in the responses but biasing them toward directions that are accurately predicted.

In order to carry out a PLS regression for a given system, the information of the substance spectra must be compared to the corresponding concentration values (Bruker Optics, 2006). In addition to the changes that occur, both data points must be recognized and correlated with each other. For this purpose, a large number of samples must be measured. The acquired information can then be used for the prediction of concentrations instead of the original spectra, because it contains relevant information of the investigated system (Janik *et al.*, 2007). The advantage here is obvious: the analytically relevant information from large data sets is compressed into factors and can then be used for the calibration (Shepherd & Walsh, 2004).

The selection of the optimum number of factors is of central importance for the quality of the PLS model (Shepherd & Walsh, 2007). With too few factors, the spectral structures are not recognized sufficiently. The corresponding regression can therefore never lead to satisfactory analysis results (Bruker Optics, 2006). This is described as underfitting. If the number of factors is too large, this leads to a deterioration of analysis, as too many parts of the disturbing spectral noise are incorporated. This is called overfitting.

4.5.3 Practically setting up a model

This is a description of practical considerations from collecting the spectra to analyzing the unknown samples. Usually six steps are necessary (Bruker Optics, 2006).

Step 1. Entering spectral data and concentration data

It is of prime importance that, prior to calculating the model, the spectra should be loaded and the corresponding concentration values for the individual components must be entered. It is necessary to define the calibration and validation spectra sets.

Step 2. Data preprocessing

Here, a method for preprocessing the spectra data is selected. This is important because it often eliminates disturbing baseline drifts. Practically, subtraction of a straight line, vector normalization or taking the first derivative of a spectrum often leads to an optimized PLS model.

Step 3. Defining an appropriate frequency range

The description of an appropriate frequency range is of crucial importance for the quality of the PLS model. Therefore, when setting a model, one should use the frequency range of the spectrum where a good correlation between the changes in the spectral and concentration data can be found. The degree of correlation can be evaluated easily by the coefficient of determination R^2.

Step 4. Validation and optimization of the method

During validation, the suitability of the chosen data preprocessing methods and the frequency range for the given measurement task is evaluated. Therefore, important parameters such as the coefficient of the determination R^2 and the root mean squared error of cross-validation (RMSECV) or the root mean squared error of prediction (RMSEP) are calculated. In addition, automatic outlier recognition is carried out.

Step 5. The calibration

After all outliers have been removed from the calibration data set, and after the optimum system parameters have been found, the final version of the model is constructed. During the calibration, the scores and loading vectors are calculated and the calibration function Y is determined. These values are stored and are available for the analysis of the new samples.

Step 6. The analysis

Here, the optimized chemometric model is used to analyze new samples. At the same time, the reliability of the analysis is checked by using characteristic parameters. There are two ways of doing this. One option is the calculation of the *Mahalanobis* distance. Here, the spectral structures of the complete calibration data set are compared to the structure of the analyte spectrum. If the spectrum contains structures which do not fit, or if the component values of the analyte are outside the calibration range, an increase of the *Mahalanobis* distance can be observed.

The second option which is frequently used to determine outliers is the calculation of the spectral residuae. This is done by calculating the difference between the measured spectrum and the spectrum which is theoretically expected from the factor analysis of the calibration spectra. The smaller this difference, the more credible the analysis result.

It is important to note that, the analysis delivers two pieces of information: it indicates the analysis value of the sample; and it provides an outlier determination. This ensures that the user is alerted if, by mistake, an erroneous measurement causes incorrect analysis results.

4.5.4 Application

Soil spectral libraries are used to generalize results of soil assessments that are conducted at a limited number of sites (Shepherd and Walsh, 2004). This in turn increases the efficiency of expensive, time consuming soil related studies. The variability of soils in a study area is thoroughly sampled. Soil properties are measured on only a selection of soils that are designed to sample the variation in the spectral library, and then calibrated to soil reflectance (Bruker Optics, 2006). The soil functional attributes can then be predicted for the entire library as well as for new samples from the study area (Vagen *et al.,* 2006). New samples that classify spectral outliers to the library are characterized and added to the calibration library, increasing the predictive value of the library.

The spectral library approach has potential for direct and simultaneous prediction of soil property attributes; this is due to the soil reflectance providing an integrated measure of the number of fundamental soil properties (Shepherd and Walsh, 2004). Such calibrations perform better and are more rapid than pedotransfer functions based on conventional measurements of soil properties. According to Bruker Optics (2006), this rapid nature of measurement allows soil variability to be more sufficiently sampled than with conventional approaches.

4.5.5 Spectral measurements

All samples were collected, dried, sieved through a 2mm sieve and placed in containers for safe storage. Soil MIR diffuse reflectance spectra were recorded for all samples, using a Fourier transform MIR spectrometer (Bruker Optics). The detector was a liquid N_2 cooled HgCdTe detector. The measured wavebands ranged from 400 to 4000cm^{-1} with a resolution of 4cm^{-1} and zero filling of 2, which resulted in 1763 data points at a waveband distance of about 2cm^{-1}. A special feature of the instrument optics is the specula reflectance, which is shielded and can

distort the shape of MIR spectra strongly (Vagen *et al*., 2006). This is the only optical setup that combines a high throughput measurement (1000 samples d^{-1}) with the exclusion of specula reflectance.

The air-dried soil samples were finely ground to powder (approximately <100µm), using a crusher. The samples were loaded into Aluminium (Al) microtiter plates (A752–96 by Bruker Optics) using a microspatula to fill the 6mm diameter well and level the soil, taking care to avoid spillage. Al is suitable as a reference material because it does not absorb infrared light. A background measurement of the first empty well was taken before the measurements, to account for changes in temperature and air humidity. Each soil sample is then loaded into a well, which is in turn scanned 64 times. The scanning process was repeated three times (see Figure 12 and 17).

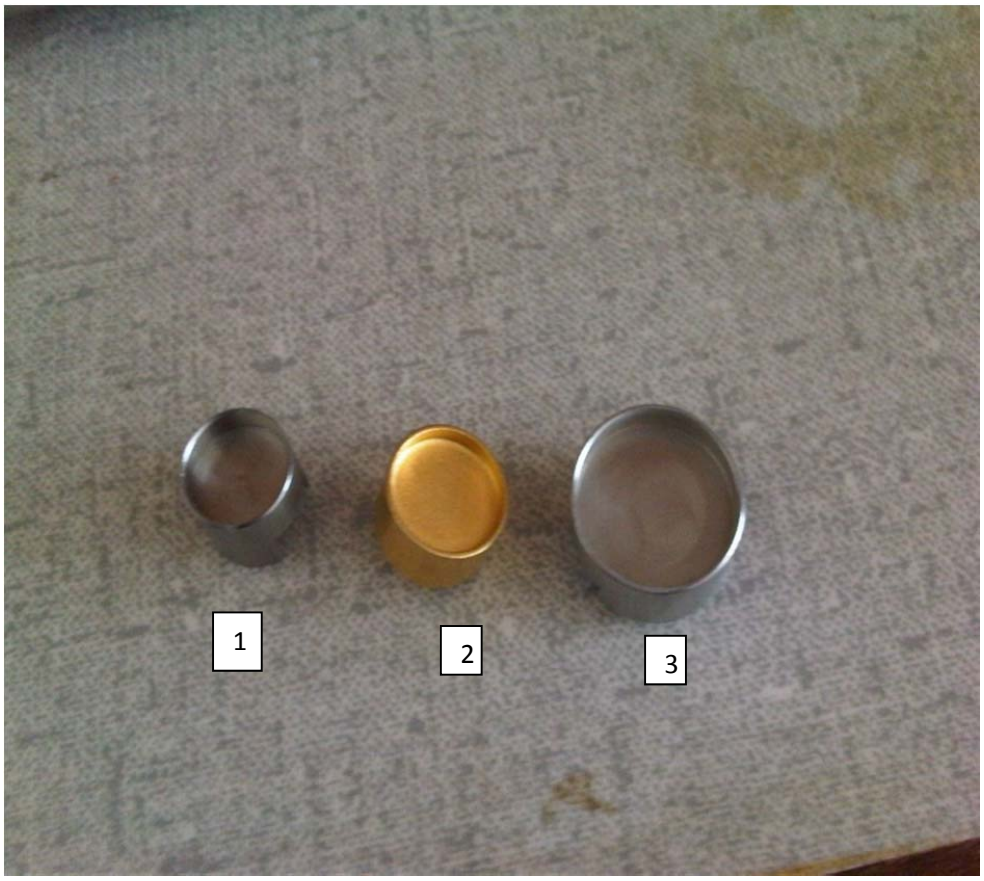


Figure 17: An illustration of a Gold background plate (2) and Aluminium microtiter plates (1 and 3).

4.6 Results (soil properties)

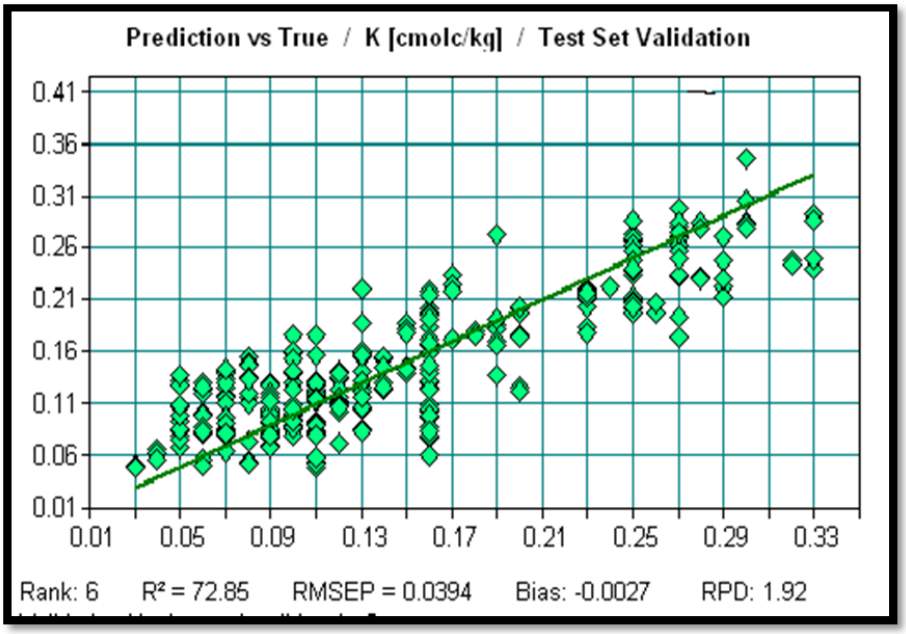
During the spectral measurement, the scanning procedure reappears three times (as described above) for each sample. This is to note any changes that might occur in the reading for each sample. Some samples are accounted for within the calibration curve, meaning all three scans fall within the calibration curve, while some samples partially fall within the calibration – in that one or two of the scans are not accounted for within the calibration curve. In some cases, all three scans are not accounted for within the calibration curve, therefore falling within the outlier category. This is presented in Table 2.

In summary, for S, Na, K, N, and C:N, 70% of the samples fitted onto the calibration curve. The calibration curve only accounted for less than 40% of the samples for pH, Ca, OrgC, Mg, and base saturation. This means that for the first group of ions only 30% of the samples have to be analyzed using wet chemistry, and even for the second group laboratory costs will be cut by 60%. The reason for the inability of the calibration curve to account for all the samples lies in an incomplete spectral library. However, when wet chemistry values are available for the outliers, they can be included in the spectral library and the calibration curve can be updated. Thus, with time, all samples should be able to be included in the calibration curves.

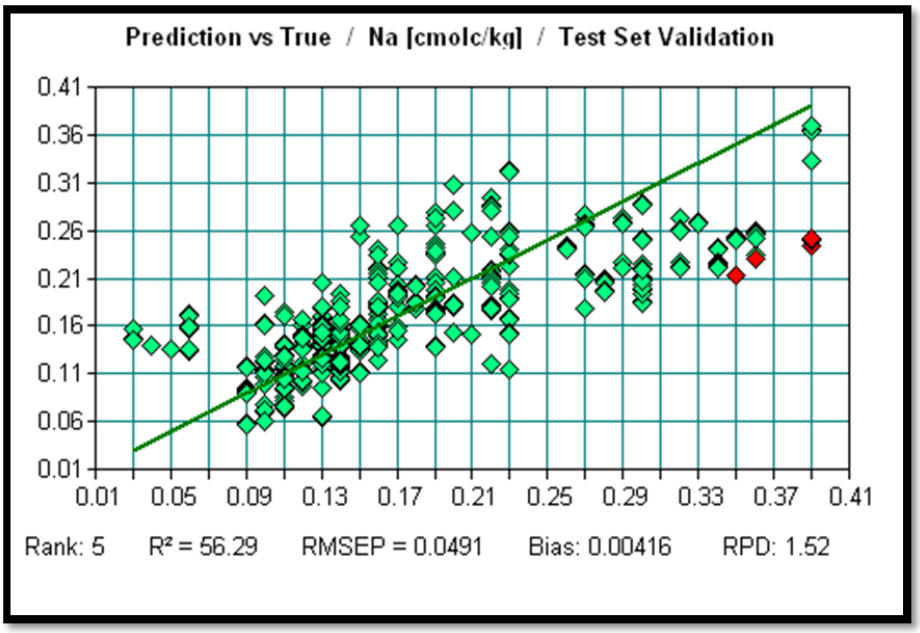
Table 2: Tabulated number of samples that were: calibrated, partially calibrated and not calibrated samples for each property.

Property	Calibrated	Partially calibrated	Not calibrated
Ca	8	5	93
C:N	79	17	10
K	51	12	43
Mg	6	8	92
N	64	11	31
Na	59	11	36
OrgC	31	25	50
S	62	16	28
BaseSat	33	13	60
pH (Kcl)	1	1	104
pH(H2O)	31	2	73

Graphs 1 and 2 are examples of the linear regressions for the validation set predicted against measured soil property values, for samples which were fitted into the calibration curve, (in red are outliers).



Graph 1: Linear regression for Potassium (K).



Graph 2: Linear regressions for Sodium (Na).

4.7 Conclusion

The MIR has proven to be rapid, timely, less expensive, non-destructive and more straightforward than conventional analysis. It shows great potential for large-scale application in South Africa. However, to be able to use it effectively, the spectral library needs to be expanded to accommodate all possible soil property values.

Even though most of the soil properties fell within the calibration curve, it is important to note that, for the purposes of this paper, only K, Ca, Na, pH-Kcl and Mg will be interpolated in hopes of achieving the main study objectives.

Chapter 5: Geostatistical analysis

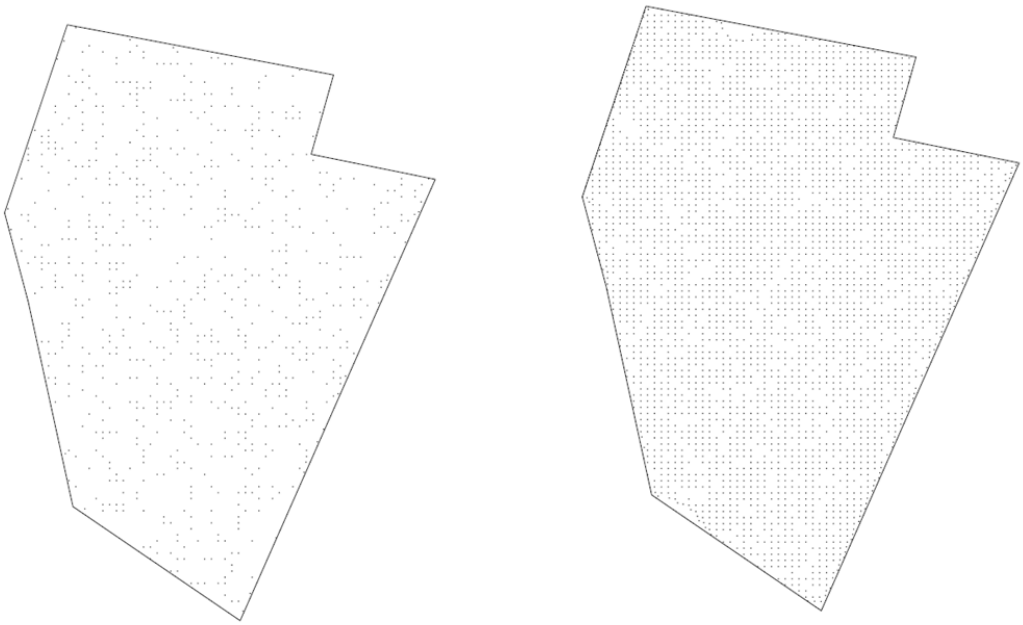
5.1 Overview

The preceding chapters presented the background of the study, defined the research problems, and explained the findings of relevant literature related to the study. This chapter describes the methodology used in testing the two different interpolation techniques – Inverse Distance Weighting and kriging (ordinary) – to determine which interpolation method would produce the most accurate surface with the least amount of prediction errors. The interpolation was conducted using ESRI's ArcGIS 10, with the Geostatistical Analyst extension.

5.2 Testing and Training Data Set

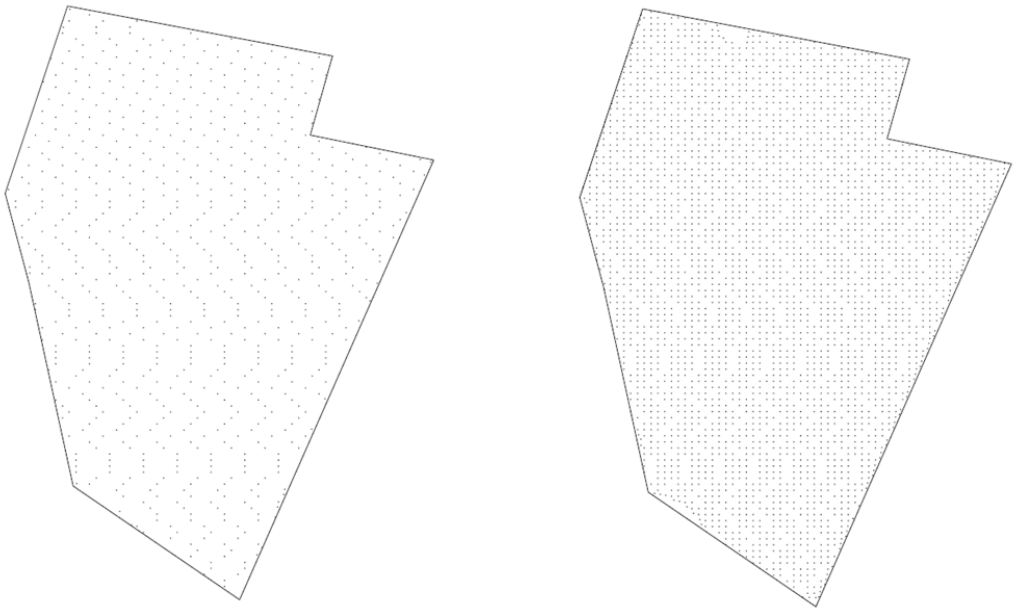
The test and training data sets were created using the ArcGIS Geostatistical Analyst extension Subset feature. The total number of surveyed points amounted to 3897. From the grid of these surveyed points, the subset of training and testing data set was randomly and systematically created. The test and training data representation of the randomly selected subsets started at: 20% training and 80% testing, 30% training and 70% testing, 40% training and 60% testing, as well as 50% training and 50% of the entire data set respectively. The test and training data representation of systematically selected subsets started at: 20% training and 80% testing (i.e. the removal of every firth data point), 25% training and 75% testing (i.e. the removal of every fourth data point), 33% training and 67% testing (i.e. the removal of every third data point), as well as 50% training and 50% testing (i.e. the removal of every second data point) (refer to Figures 18 and 19).

It is of prime importance to note that due to the systematic nature of this data set, one does not have control over the percentages of the training and testing data sets, hence, some of the training and testing data set percentages for the systematic and random data are not the same (These are: random data set: 30% training and 70% testing, 40% training and 60% testing with systematic data sets; 25% training and 75% testing (i.e. the removal of every fourth data point), 33% training and 67% testing (i.e. the removal of every third data point). The random and systematically selected data sets will provide insight into which sampling technique is better suited for interpolating using IDW and ordinary kriging. The training and testing data sets were created for use in validation and cross-validation. It is important to note that, all interpolation surfaces were created using the ArcMap geostatistical analyst tool's default parameters. The last step of the interpolation techniques shows the results of cross-validation.



(a) 20% training (b) 80% testing

Figure 18: The spatial distribution of the randomly selected training (20%) and testing (80%) data set.

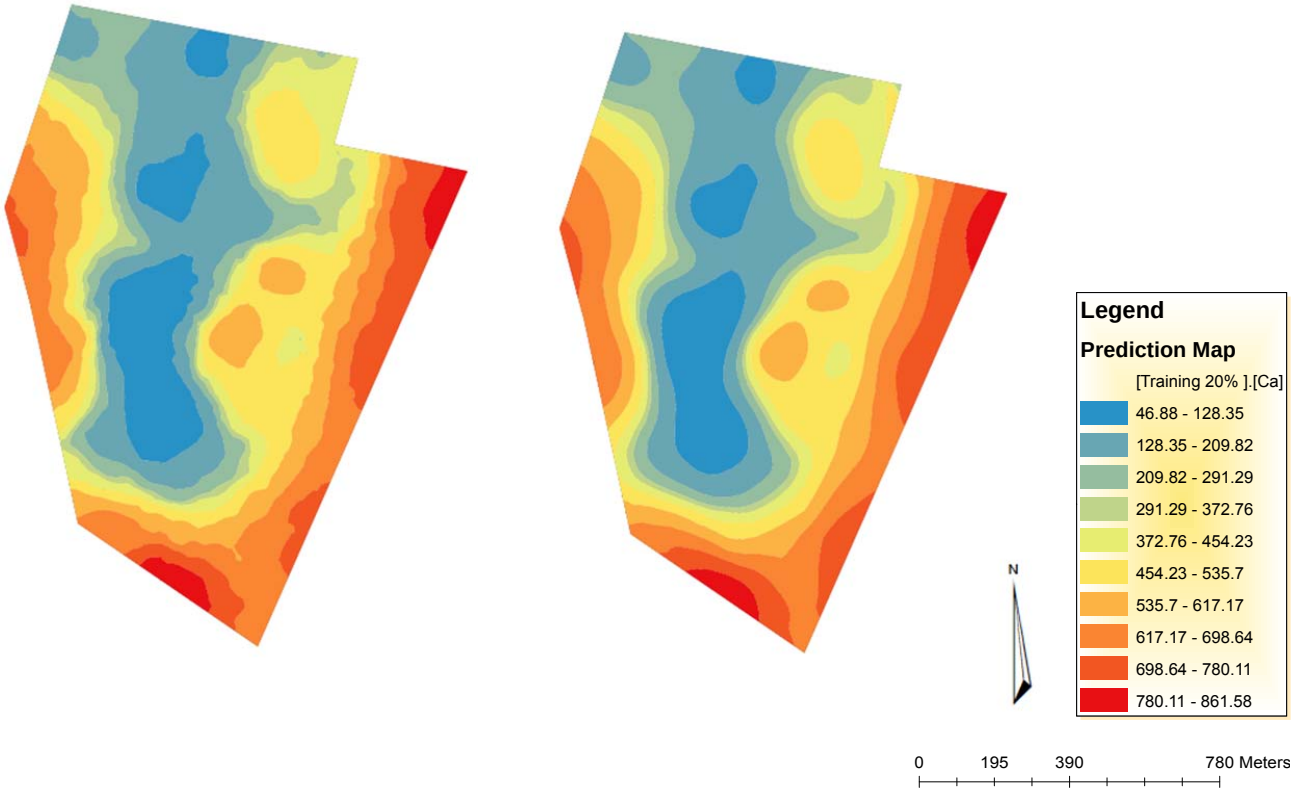


(a) 20% training (b) 80% testing

Figure 19: The spatial distribution of the systematically selected training (20%) and testing (80%) data set.

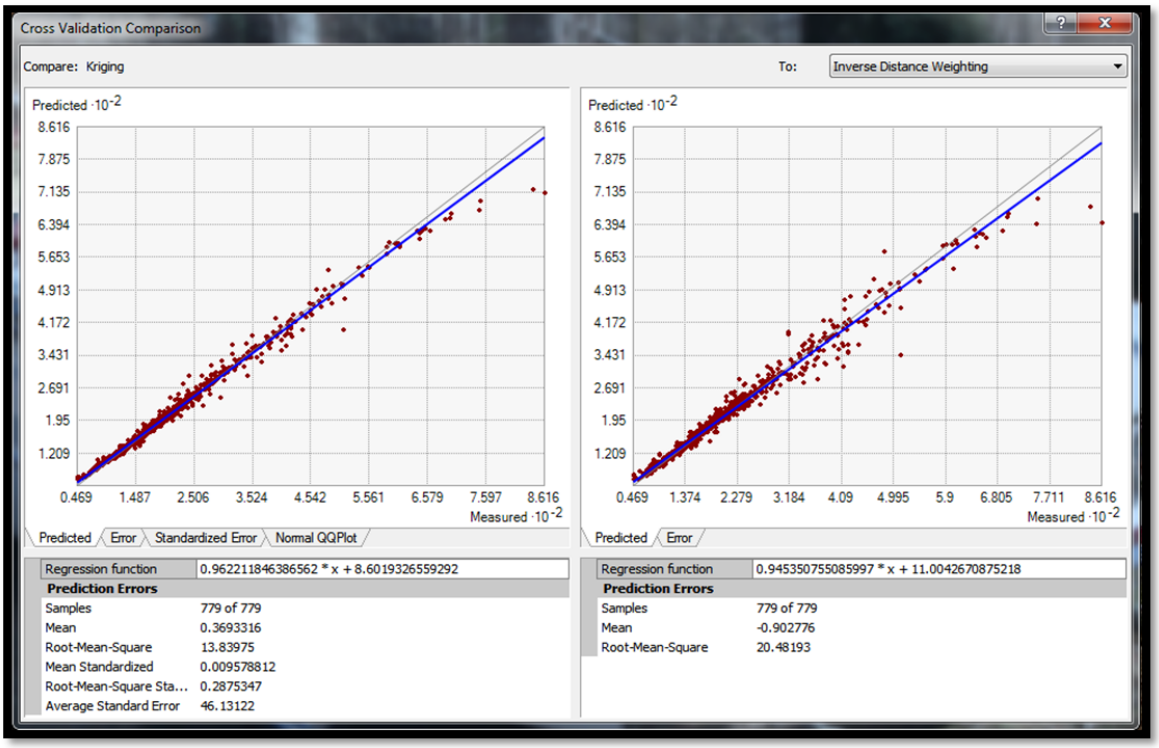
5.3 Cross-validation

Cross-validation uses all of the data to estimate the autocorrelation model. This process removes one data point at a time and predicts the associated value (Tomczak, 1998). The predicted and actual values at the location of the removed data point are compared (Esri, 2013). This procedure is repeated for a second point, and so on. For all the data points, cross-validation compares the measured and predicted values (Johnston *et al.*, 2001). After completing cross-validation, some data points may be set aside as outliers, requiring the trend and autocorrelation models to be refitted. For the purpose of this study, cross-validation was performed automatically and the results are represented on a graph of predicted versus measured soil properties. Then the two models were compared visually. Refer to Figure 20 and 21 as well as Graph 3 and 4.

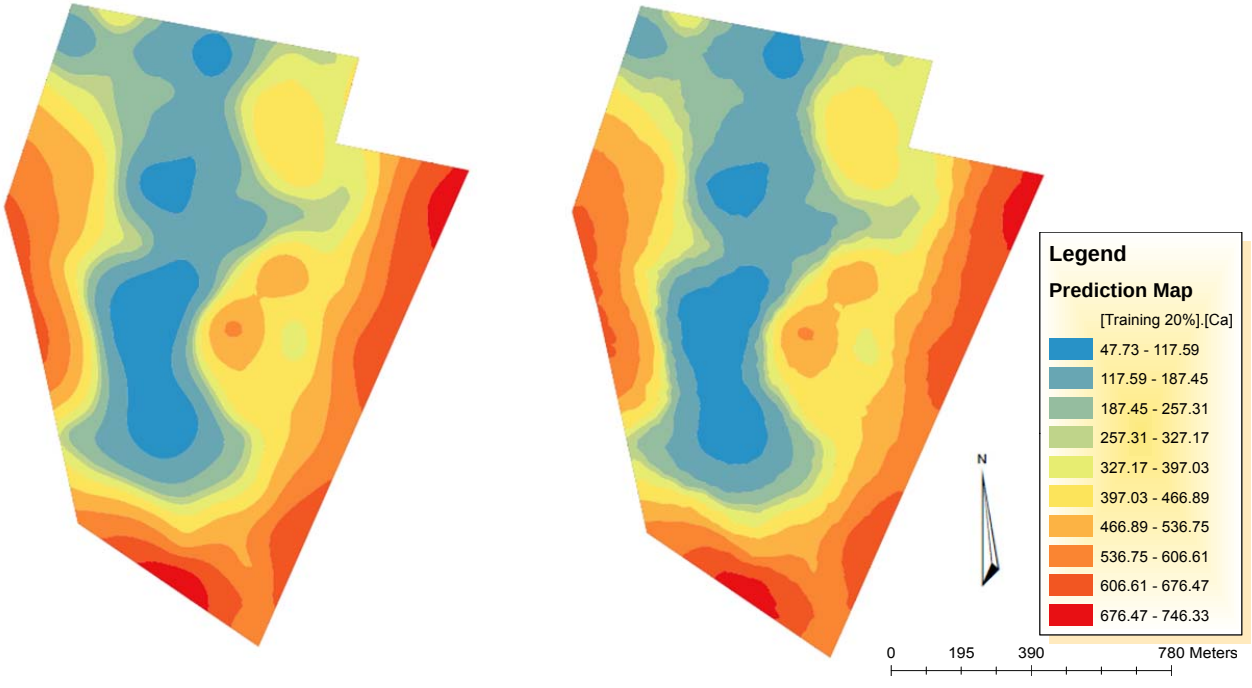


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 20: Prediction maps of randomly selected training (20%) and testing (80%) data sets for Calcium (Ca).

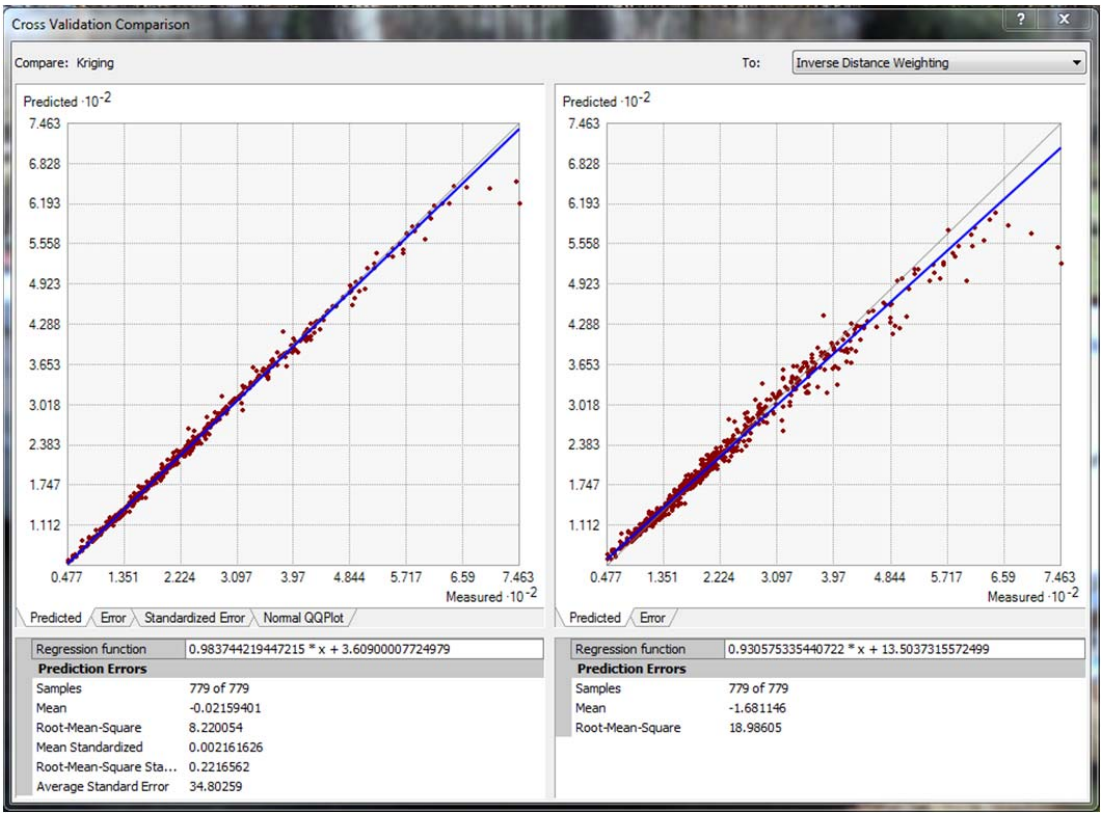


Graph 3: Cross-validation comparison of predicted errors for the randomly selected Ca training (20%) data set.



(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 21: Prediction maps of systematically selected training (20%) and testing (80%) data sets for Calcium (Ca).



Graph 4: Cross-validation comparison of predicted errors for the systematically selected Ca training (20%) data set.

5.4 Validation

Validation uses the training and testing data set to develop the trend and autocorrelation model used for the prediction (Esri, 2013). The types of graphs and summary statistics used to compare predictions to true values are similar for both validation and cross-validation. Validation creates a model for only the subset of the data, so it does not directly check the final model, which should include all available data. Rather, validation checks whether a protocol of the decision is valid – for example: the choice of a semi-variogram model, lag size, and search neighborhood.

For the purpose of the study, validation models were performed using the Geostatistical Analyst (GA) layer to Points geoprocessing tool. All the models were assessed using validation. This was done after an interpolation model (surface layer) was created. The training data set was used to create the interpolation model, then validation was run using the testing data set. The field to validate on was the same attribute field used to create the interpolation model.

The RMSE was calculated on validation attribute tables. To calculate the RMSE, a field was added to each validation attribute table, defined as a double and called Error_Squared. A field calculator was used to calculate the values of this newly created field with an expression of “Error * Error”. The statistics tool was used to obtain the mean of these squared errors and finally, the square root of the mean was calculated, resulting in the value of the Root Mean Square Error (RMSE).

5.5 Conclusion

This chapter discussed the geostatistical methodology followed to acquire results for the two different interpolation techniques, Inverse Distance Weighting and kriging (ordinary), to determine which interpolation method will produce the most accurate surface with the fewest prediction errors.

Chapter 6: Results

6.1 Overview

In this chapter, the results of the data analysis are presented. The data were collected and then processed in response to the problems posed in chapter 1 of this dissertation. Two fundamental objectives drove the collection of the data and the subsequent data analysis. The first objective was to determine the effectiveness of the kriging and IDW interpolation techniques. This was to be accomplished by comparing the total error of cross-validation and validation statistics. The second objective was to improve the spatial sampling structure for interpolation purposes. This would be achieved by comparing grid sampling and random sampling techniques. These objectives were accomplished. The findings in this chapter demonstrate which interpolator produced the least TE using both random and grid subset data of the entire data set.

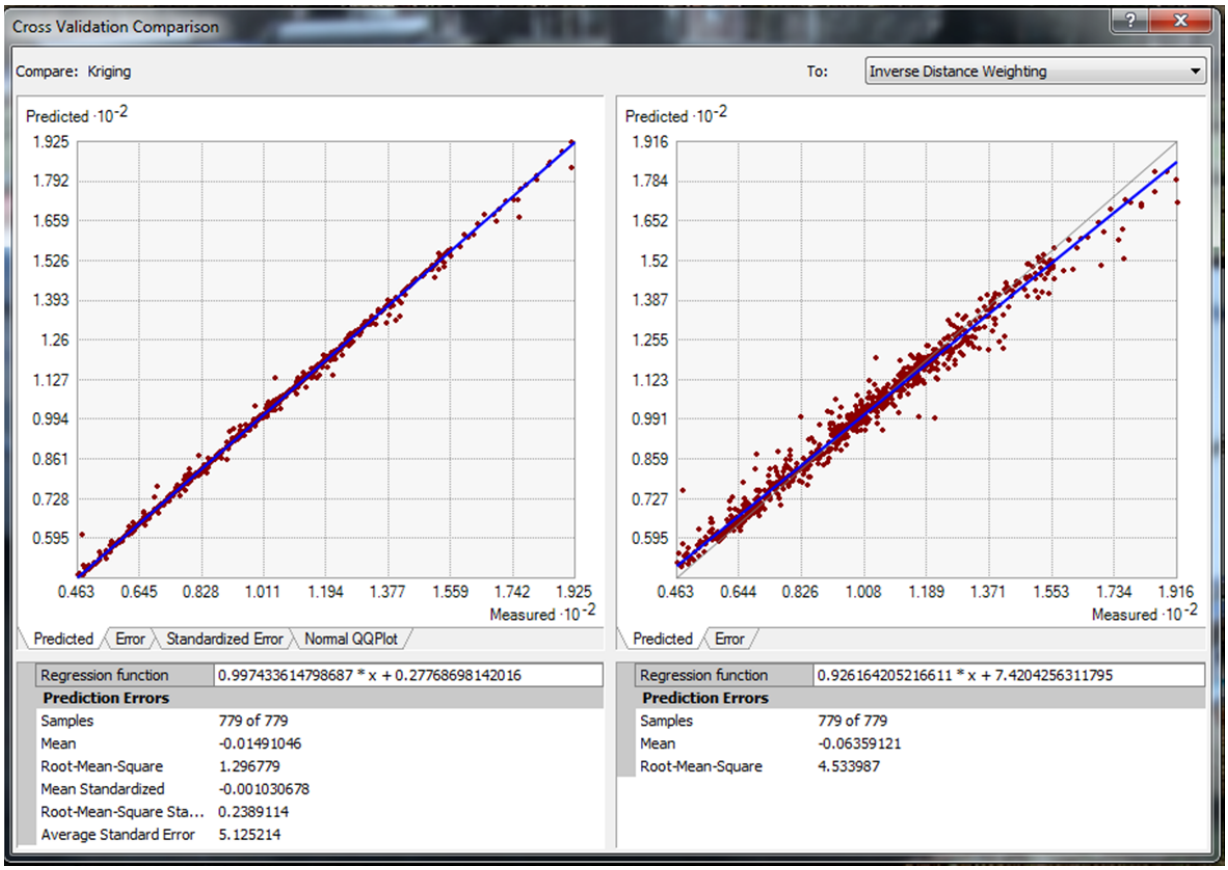
6.2 Results

The accuracy of the two interpolators to predict soil chemical properties was performed on randomly and systematically selected training and testing data sets. A total of 40 interpolation surfaces and 40 cross-validation scatter plots were produced. These are presented in Appendix A and B. Validation was performed for each data set using the training and testing subsets. The MPE and RMSE produced by each interpolation technique were tabulated, and the TE were calculated. The interpolator that produced the lowest total errors (TE) was then considered the most accurate interpolator between the two (ordinary kriging and IDW).

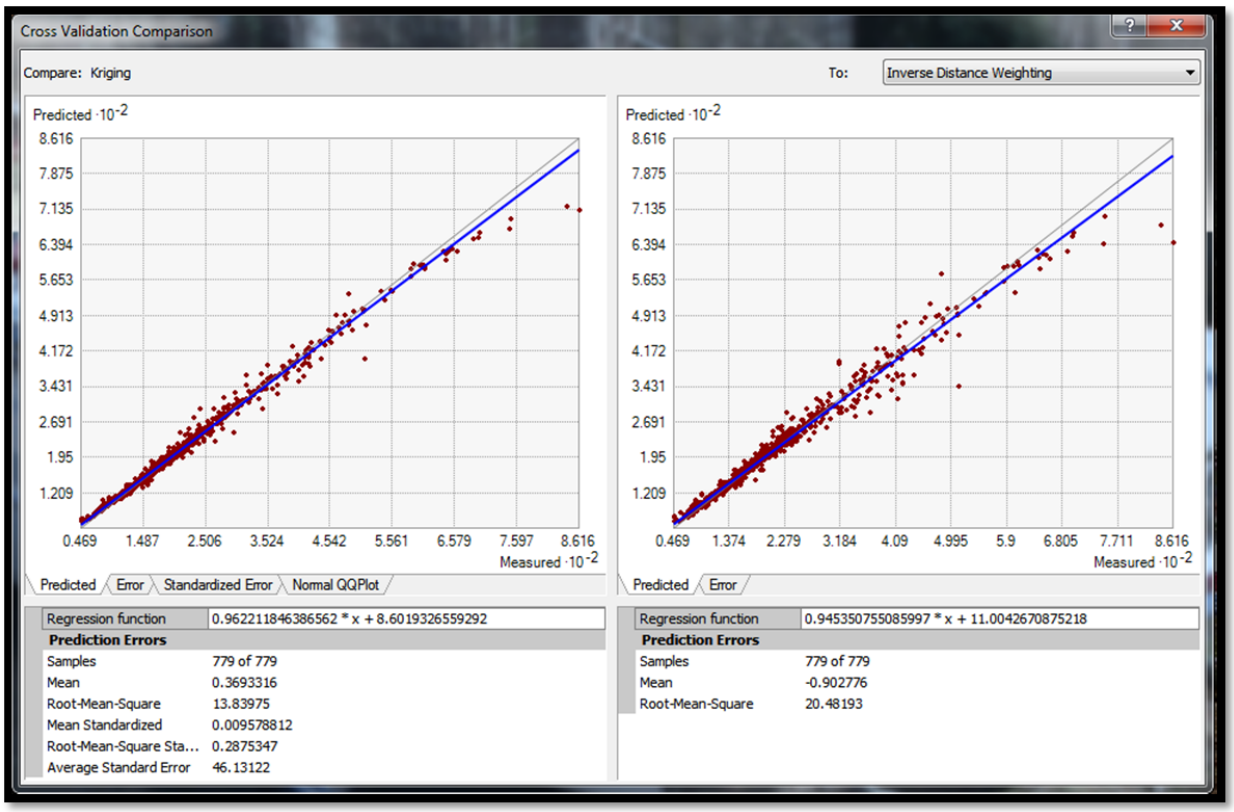
6.3 Scatter plots

Geostatistical Analysis gives several graphs and summaries of the measured values versus the predicated values. Among these graphs, a scatterplot of predicated versus measured true soil chemical properties were produced for the purpose of this study. These scatter plots are expected to be on a line of about 1:1 (the gray line). Though this expectation is not always the case due to the slope, which is usually less than 1. The fitted line though the scatter of points is given in blue. These graphs show how well the interpolators predicted. If all the data was independent (no autocorrelation), the blue line should be horizontal. With autocorrelation and good interpolation method, the blue line should be closer to the 1:1 (gray) line. For the systematic subset data, using the ordinary kriging method, all the scatter plot graphs were closer to the 1:1 line and only a few of the random subsets using ordinary kriging method were closer to the 1:1. This proves that systematic sampling with equal distances between soil

samples is more suitable for interpolation purposes, as it provides better autocorrelation. On the other hand, systematic subset data using IDW showed a poorer autocorrelation than that of ordinary kriging. This is because IDW uses a simpler averaging logarithm. Randomly selected data sets showed ordinary kriging and IDW to have almost the same precisions in that, in some cases, IDW samples were closer to the 1:1 line. (See Graph 5 and 6).



Graph 5: Cross-validation comparison of predicted error for the systematically selected 20% K Training data set



Graph 6: Cross-validation comparison of predicted error for the randomly selected 20% Ca Training data set

6.4 Evaluation performance tables

Listed in the evaluation performance tables are the mean prediction errors (MPE) and root mean squared error (RMSE) for both cross-validation and validation (see Tables 3–10). Tables 3–10 include the total errors (TE), which are the sums of MPE and RMSE for each data set. These scores show the overall ability of each interpolation technique during the cross-validation and validation process. An unbiased prediction is indicated by a mean prediction error closest to 0. Ordinary kriging showed MPEs closer to zero in the entire random and systematic performance tables as compared to IDW. This then proved beyond doubt that ordinary kriging is a more unbiased predictor of soil chemical properties. Ordinary kriging also produced the lowest RMSE and TE values in comparison with IDW. (See Tables 3–10). All evaluation performance tables show the RMSE as well as the TE for both random and systematic data set to be smaller for ordinary kriging. For this reason, ordinary kriging is the most suitable method for prediction and mapping the spatial distribution of soil chemical properties.

The systematic subset data produced lower TE values as compared to random subset data. This is evident on Table 3 and Table 7, where both tables show similar percentages of 20% training and 80% testing subsets for random and systematic data. Similarly, Table 6 and Table 10, which are also representing similar percentages of 50% training and 50% testing subsets, also showed systematic data sets outperforming random data sets by producing lower TE values. For this reason, systematic sampling is much more suited for the purpose of interpolation.

Both systematically and randomly selected data sets produced the lowest TE values at 50% training and 50% testing subset data for ordinary kriging and IDW. This is an indication that 50% or more of the study area should be sampled for optimum interpolation results. Thus, for the purpose of interpolation, there is a relationship between data needed to generate a map and map accuracy.

Table 3: Evaluation performance of ordinary kriging and IDW of soil properties through cross-validation and validation statistics for the random subset data of 20% training and 80% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.36933	14.09510	14.46443	-0.90277	20.48193	19.57916
K	0.00459	1.22361	1.22820	0.02414	4.26228	4.286414
Mg	-0.00517	1.87471	1.86954	-0.31897	4.50938	4.190411
Na	0.00035	0.19479	0.19514	-0.00739	0.69169	0.684296
pH-KCl	-0.00038	0.02193	0.02156	-0.00565	0.05741	0.051761

Table 4: Evaluation performance of ordinary kriging and IDW of soil properties through cross-validation and validation statistics for the random subset data of 30% training and 70% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.22922	10.88430	11.11352	-1.06327	13.06520	12.00193
K	0.01279	0.97646	0.98925	0.08414	3.75691	3.84105
Mg	-0.00628	0.76441	0.75813	-0.27379	4.23909	3.96530
Na	-0.00001	0.14991	0.14990	-0.00854	0.61949	0.61095
pH-KCl	-0.00006	0.01947	0.01941	-0.00164	0.03984	0.03820

Table 5: Evaluation performance of ordinary kriging and IDW of soil properties through cross-validation and validation statistics for the random subset data of 40% training and 60% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.07117	4.00559	4.07676	-0.56829	11.01720	10.44891
K	0.00906	0.63885	0.64791	-0.00102	2.91403	2.91301
Mg	-0.00166	0.45287	0.45121	-0.01877	2.97545	2.95668
Na	-0.00058	0.09711	0.09653	-0.00834	0.50286	0.49452
pH-KCl	0.00007	0.01855	0.01862	-0.00093	0.03379	0.03286

Table 6: Evaluation performance of ordinary kriging and IDW of soil properties through cross-validation and validation statistics for the random subset data of 50% training and 50% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.07062	5.78091	5.85153	-0.77244	14.62990	13.85746
K	-0.00299	0.50967	0.506676	-0.05266	2.56034	2.50768
Mg	-0.00206	0.33082	0.32876	-0.15719	2.50444	2.34725
Na	-0.00039	0.07509	0.0747	-0.00015	0.46801	0.46786
pH-KCl	-0.00007	0.01783	0.01776	-0.00092	0.03143	0.03051

Table 7: Evaluation performance of ordinary kriging and IDW of soil properties through cross-validation and validation statistics for the systematic subset data of 20% training and 80% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	-0.02159	8.38757	8.36598	-1.68115	18.62088	16.93973
K	-0.01491	0.91854	0.90363	-0.06359	3.47751	3.41392
Mg	-0.00782	0.74979	0.74197	-0.30971	3.62169	3.31198
Na	0.00284	0.18129	0.18413	-0.00621	0.61095	0.60474
pH-KCl	-0.00027	0.02112	0.02085	-0.00194	0.04494	0.043

Table 8: Evaluation performance of ordinary kriging and IDW of soil properties through cross-validation and validation statistics for the systematic subset data of 25% training and 75% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.18140	7.24447	7.42587	-1.24343	12.8851	11.64167
K	-0.00711	0.74231	0.73520	-0.04953	2.94828	2.89875
Mg	-0.01641	0.49773	0.48132	-0.25859	2.74152	2.48293
Na	-0.00077	0.10269	0.10192	-0.01224	0.46523	0.45299
pH-KCl	-0.00002	0.01844	0.01842	-0.00201	0.03667	0.03466

Table 9: Evaluation performance of Ordinary Kriging and IDW of soil properties through cross-validation and validation statistics for the systematic subset data of 33% training and 67% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.12709	5.47369	5.60078	-0.96880	11.76567	10.79687
K	-0.00594	0.64410	0.63816	-0.05817	2.54758	2.48941
Mg	-0.00599	0.41258	0.40659	-0.17284	2.61854	2.44570
Na	-0.00184	0.09948	0.09764	-0.00467	0.42819	0.42352
pH-KCl	-0.00007	0.01881	0.01874	-0.00166	0.03484	0.03318

Table 10: Evaluation performance of Ordinary Kriging and IDW of soil properties through cross-validation and validation statistics for the systematic subset data of 50% training and 50% testing.

	Ordinary Kriging			**Inverse Distance Weighting**		
Properties	**MPE**	**RMSE**	**TE**	**MPE**	**RMSE**	**TE**
Ca	0.06595	3.99649	4.06244	-0.56978	10.24499	9.67521
K	0.00557	0.39299	0.39856	0.01075	1.88788	1.89863
Mg	-0.00319	0.3251	0.32191	-0.15186	2.21169	2.05983
Na	-0.00063	0.05544	0.05481	-0.00675	0.32089	0.31414
pH-KCl	0.00005	0.01616	0.01621	-0.00056	0.02472	0.02416

6.5 Conclusion

With the application of GIS and geostatistics mapping, spatial distribution of soil chemical properties in the study area was mapped and evaluated based the errors produced by the two interpolation techniques being compared on random as well as systematic data sets. The results of the validation suggest that the best model for generating soil chemical property prediction is ordinary kriging, using 50% systematic subset data. This means that for optimal results for the study area, 20x20m had to be sampled, and interpolated using ordinary kriging.

Chapter 7: Study conclusion and recommendations

7.1 Conclusion

In this study, IDW has proven to be a simpler procedure and has fewer steps in comparison to kriging. As discussed in the previous chapters, IDW estimates cell values by averaging the values of nearby sample data points. The closer a point is to the center of the cell which is being estimated, the more weight it is given. IDW is usually better when sample data points representing the minimum and maximum values of your surface are known. Because IDW is an averaging process, all interpolated values are within the sample range. Line data may interrupt the interpolation process – for example: one would not be able to interpolate soil chemical property data over a cliff, or across a river. IDW also allows the user to set interpolation barriers of this kind (line data). Lastly, IDW is ideal for interpolating large sets of samples. For more informative data, like the data used in this study, kriging is preferable.

Like IDW interpolation, kriging forms weights from surrounding measured values to predict values at unmeasured locations. As with IDW interpolation, the closest measured values usually have the most influence. However, the kriging weights for the surrounding measured points are more sophisticated than those of IDW. IDW uses a simple algorithm based on distance, but kriging weights come from a semi-variogram that was developed by viewing the spatial structure of the data. To create a continuous surface or map of the phenomenon, predictions are made for locations in the study area based on the semi-variogram and the spatial arrangement of measured values that are nearby. Even though kriging proved to be a more reliable interpolator in this study, IDW also has its advantages.

The production of soil property maps is the most important and first step in site-specific management. These maps display spatial variability and provide the basis to control it. The main objectives of this study were to evaluate the effect of sampling density on mapping accuracy of soil properties with diverse spatial structure and variability, as well as to compare the performance of ordinary kriging and IDW for interpolating soil properties using cross-validation and validation statistics. Three sensitivity analyses were performed in this research. The interpolation methods to produce the soil property surface map in this section follow the exact steps as listed in section 5. Cross-validation and validation MPE and RMSE were calculated, combined to determine which interpolator produced the lowest TE.

The findings demonstrated that the accuracy achieved in mapping soil properties strongly depends on the spatial structure of the data. This was shown when the subset training data set was decreased, and the resulting total error (TE) increased. The results also confirmed that using a systematic sampling pattern provides more accurate results than a random sampling pattern, and that precision of the interpolation increases with increasing data size. The overall results obtained from the comparison of the two applied interpolation methods indicated that kriging was the most suitable method for prediction and mapping the spatial distribution of soil chemical properties in this study area, and that systematic sampling at 20x20m on a 153ha area was the most reliable sampling scheme for the purpose of interpolation.

7.2 Recommendations

There are other interpolation techniques available under the ArcGIS Geostatistical Analyst which were not used in this study. These include: universal, simple, probability and disjunctive interpolation techniques, which are available under kiging or co-kriging methods. These techniques could be applied in future comparative studies to estimate soil chemical properties.

References

ANONYMOUS, 2010. Geographic Information Science and spatial reading. http://map.sdsu.ed/geo104/lecturer/unit-6.htm [Accessed: 04/03/2012].

ANONYMOUS, Undated. www.personal.psu.edu/cab38/GEO321/08_isolines02/semivariogram.giff [Accessed: 04/03/2013].

ATKINSON, P.M., 1997. Scale and spatial dependence. In: van Gardingen, P.P., Foody, G.M., and Curran, P.J., (eds.). Scaling up from cell to landscape. Society for Experimental Biology seminar series 63. Cambridge: Cambridge University Press.

AWITI, A.O., WALSH, M.G., SHEPHERD, K. D., & KINYAMARIO, J., 2008. Soil condition classification using infrared spectroscopy: a proposition for assessment of soil condition along a tropical forest-cropland chronosequence, *Geoderma*, Vol. 143: 73–84.

BAILEY, T.C., & GATRELL, A.C., 1995. Interactive spatial data analysis. Essex, England: Longman.

BRIDGES, E. M., 1997. World soils. 3rd edn. Cambridge, UK: Cambridge University Press.

BRO, R., 2003. Multivariate calibration: what is in chemometrics for the analytical chemistry? Analytica Chimica Acta, Vol. 500: 185–194.

BROWN, D.J., SHEPHERD, K.D., & WALSH, M.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, Vol. 132: 273–290.

BRUKER OPTICS, 2006. Multivariate calibration: a practical guide for the method development in the analytical chemistry, 2nd English edn.

BRUS, D.J., & DE GRUIJTER, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model based sampling strategies for soil (with Discussion). *Geoderma*, Vol. 80: 1–59.

BURROUGH, P.A., & MCDONNELL, R.A., 1998. Principles of geographical information systems. New York: Oxford University Press.

BURROUGH, P.A., 2001. GIS and geostatistics: essential partners for spatial analysis. *Environmental and Ecological Statistics*, Vol. 8: 361–377.

CADELL, W., Undated. Notes on: report on geostatistics for spatial interpolation and sampling efficiency in site characterization. Available at: www.macualay.ac.uk/LADSS/documents/Geostatistics-notes.pdf [Accessed: 28/06/2012].

CHILÈS, J. P., & DELFINER, P., 1999. Geostatistics: modeling spatial uncertainty. 2nd edn. New York: Wiley.

CLARK, I., 1979. Practical geostatistics. Great Yarmouth: Gilliard Printers.

CRESSIE, N., 1986. Kriging nonstationery data. *Journal of the American Statistical Association*, Vol. 81, No. 395: 625–634.

CRESSIE, N., 1990. The origins of Kriging. *Mathematical Geology*, Vol. 22, No. 3: 239–252.

CRESSIE, N., 1993. Statistics for spatial data: statistics and mathematical statistics. New York: John Wiley.

CSIR, SAC, 2012. 2624DC, 2625CB, 2724DC, and 2725CB. Pretoria, CSIR.

DE SMITH, M.J., GOODCHILD, M.F., & LONGLEY, A.P., 2013. Geospatial analysis: a comprehensive guide to principles, techniques and software tools. 4th ed. Winchelsea: Winchelsea Press.

DE VILLERS, B. & MANGOLD, S., 2002. The biophysical environment. State of the environment report 2002 North West Province, South Africa. North West province department of Agriculture Conservation and Environment, Mmabatho

DESMET, P. & SEYMOUR, C., 1999. Coping with drought – do science and policy agree? *South African Journal of Science*, Vol. 105: 18..

DUTILLEUL, P., 1993. Spatial heterogeneity and the design of ecological field experiments. *Ecology*, Vol. 74, No. 6: 1646–1658.

ESRI, 2013. ArcGIS Desktop 10 Help**.** An overview of the Interpolation toolset. Redlands, CA. http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/An_overview_of_the_Interpolation_tools/009z00000069000000/ [Accessed: 06/05/2013].

GOODCHILD, M.F., & LAM, N., 1980. Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, Vol. 1: 297–312.

GOOVAERTS, P., 1997. Geostatistics for natural resources evaluation. New York: Oxford University Press: 483.

HARTKAMP, A.D., DE BEURS, K., STEIN, A., & WHITE, J.W., 1999. Interpolation techniques for climate variables. Natural Resources Group GIS, Series 99-1. Mexico, D.F.: CIMMYT: 1–268.

HENGL, T., HEUVELINK, B.M.G., & STEIN, A., 2004. A generic framework for spatial prediction of soil prediction of soil variables based on regression-kriging. *Geoderma*, Vol. 120: 75-93.

HENLEY, S., 1981. Non-parametric geostatistics. Great Yarmouth: Gilliard.

HEUVELINK, G.B.M., & WEBSTER, R., 2000. Modelling soil variation: past, present, and future. *Geoderma*, Vol. 100, No. 3: 269–301.

HOLMES, K.W., VAN NIEL, K., & BAXTER, K., 2004. Designs for marine remote sampling: a review and discussion of sampling methods, layout, and scaling issues. University of Western Australia: Crawley: 1–37.
Available at: http://www.thsoa.org/pdf/h01/4_5.pdf (Accessed: 19/02/2003).

JANIK, L.J., SKJEMSTAD, J.O., SHEPHERD, K.D., & SPOUNCER, L.R., 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Australian Journal of Soil Research*, Vol. 45, No. 2: 73–81.

JOHNSTON, K., VER HOEF, J. M., KRIVORUCHKO, K., & LUCAS, N., 2001. Using ArcGIS geostatistical analyst. USA: ESRI.

KARYDAS, G.C., GITAS, Z.I., KOUTSOGIANNAKI, E., LYDAKIS-SIMANTIRIS, N., & SILLEOS, N.G., 2009. Evaluation of spatial interpolation techniques for mapping agricultural topsoil properties in Crete. *EARSel eProceedings*, Vol. 8, No. 1: 28–39.

KEYSER, N. & DU PLESSIS, C.P., 1993. The geology of the Vryburg area. Explanation to 1: 250 000 geology sheet 2624 Vryburg, *Council for Geoscience*, Pretoria: 28.

KLUTE, D.S., LOVALLO, M.J., & TZILKOWSKI, W.M., 2002. Autologistic regression modeling of American woodcock habitat use with spatially dependent data. In: Scott M.J., Heglund, P.J., & Morrison, M.L. (eds). Predicting species occurrences: issues of accuracy and scale. Washington, D.C: Island Press: 335–344.

KRAVCHENKO, A.N., & BULLOCK, D.G., 1999. A comparative of interpolation methods for mapping soil properties, *Agronomy Journal*, Vol. 91: 393–400.

KRIGE, D.G., 1966. Two-dimensional weighted moving average trend surface for ore-valuation in symposium on mathematical statistics and computer applications in ore-valuation. Journal South African Institute of Minerals and Metals: 13–79.

LEVIN, S.A., 1992. The problem of pattern and scale in ecology. *Ecology*, Vol. 73: 1943–1967.

LU, Y.G., & WONG, W.D., 2008. An adaptive inverse-distance weighting spatial interpolation technique, earth system and geoinformation sciences. *Computers and Geosciences Journal*, Vol. 34, No. 9: 1044–1055.

MAPUKULE, L.E., 2009. Interpretation of regional geochemical data as an aid to exploration target generation in the North West province, South Africa. Unpublished MSc thesis. Alice, Eastern Cape, South Africa. University of Fort Hare.

MASIGO, A., & MATSHEGO, C., 2002. Provincial report on education and training for agriculture and rural development in North West Province. North West: North West Department of Agriculture, Conversation and Environmental.

MASON, B.J., BROWN, K.W., & SCHUMACHER, B.A., 1992. Preparation of soil sampling protocols: sampling techniques and strategies. USA, Nevada: Environmental Monitoring Systems Laboratory Office of Research and Development U.S. Environmental Protection Agency: 1–169.

MATHERON, G., 1963. Principles of geostatistics. *Economic Geology*, Vol. 58: 1246–1266.

MAURO, R., ROSAMARI, S., & ROBETO, S., 2011. Earth and planetary science: planning air pollution monitoring networks in industrial area by means of remote sensed images and GIS techniques. *National research council of Italy*, Vol. 10, No. 1: 16341–16416.

McBRATNEY, A.B., 1992. On variation, uncertainty and informatics in environmental soil management. *Australian Journal of Soil Research*, Vol. 30, No. 6: 913–935.

McBRATNEY, A.B., WHELAN, B.M., WALVOORT, D.J.J., & MINASNY, B., 1999. A purposive sampling scheme for precision agriculture. Sheffield, UK: Sheffield Academic Press.

MINASNY, B., & McBRATNEY, A.B., 2005. The Matérn function as a general model for soil variograms, *Geoderma*, Vol. 128, No. 3–4: 192–207.

MINASNY, B., & McBRATNEY, A.B., 2007. Spatial predictions of soil properties using EBLUP with the Matérn covariance function. *Geoderma*, Vol. 140, No. 4: 324–336.

NAOUM, S., & TSANIS, K.I., 2004. Ranking spatial interpolation techniques using a GIS based DSS. Global Nest*: the Int. Journal,* Vol. 6, No. 1: 1–20.

NOYES, P.D., MCELWEE, M.K., CLARK, B.W., VAN TIEM, L.A., WALCATT, K.C., ERWIN, K.N., & LEVIN, E.D., 2009. The toxicology of climate change: environmental contaminates in a warming world. *Environment International*, Vol. 10: 971–986.

OLIVER, M.A., 1990. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, Vol. 4, No. 3: 313–332.

OLIVER, M.A., WEBSTER, R., & GERRARD, J., 1989. Geostatistics in physical geography. *Transactions of the Institute of British Geographers*, New Series, Vol. 14, No. 3: 259–269.

PANAGOPOULOS, T., JESUS, J., ANTUNES, M.D.C., & BELTRAO, J., 2006. Analysis of spatial interpolation for optimizing management of a salinized field cultivated with lettuce. *European Journal of Agronomy*, Vol. 24, No. 1: 1–10.

ROBINSON, T.P., & METTERNICHT, G., 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. *Computer and Electronics in Agriculture*, Vol. 50, No. 2: 97–108.

SANTRA, P., CHOPHRA, U.K., & CHAKRABORTY, D., 2008. Spatial variability of soil properties and its application in predicting surface map of hydraulic parameters in an agricultural farm. *Current Science*, Vol. 95, No. 7: 937–945.

SCHULZE, R.E., 1997. South African atlas of agrohydrology and climatology. Water research commission, Report TT82/96. Pretoria.

SCULL, P., FRANKLIN, J., CHADWICK, O.A., & McARTHUR, D., 2003. Predictive soil mapping: a review. *Progress in Physical Geography*, Vol. 27, No. 2: 171–197.

SEN, Z., & SAHIN A.D., 2001. Spatial Interpolation and estimation of solar irradiation by cumulative semivariogram. *Solar Energy*, Vol. 71, No. 1: 11–21.

SHAFFER, K.A., FRITTON D.D., & Baker D.E., 1979. Drainage water sampling in a wet, dual-pore soil system. *Journal of Environmental Quality*, Vol. 8, No. 2: 241–245.

SHEIKHHASAN, H., 2006. A comparison of interpolation techniques for spatial data prediction. Master's Thesis in Computer Science. Amsterdam, Netherlands: Faculty of Science, Universiteit van Amsterdam: 1–60.

SHEPHERD, K.D., & WALSH M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America*, Vol. 66: 988–998.

SHEPHERD, K.D., & WALSH M.G., 2004. Diffuse reflectance spectroscopy for rapid soil analysis, Encyclopedia of Soil Science. Kenya: Marcel Dekker Inc.

SHEPHERD, K.D., & WALSH M.G., 2007. Infrared spectroscopy-enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. *World Agroforestry Centre*, Vol. 15: 1–19.

SHEPHERD, K.D., 2010. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. *Soil Science Society of America*, Vol. 74, No. 5: 1792–1799.

SNEDECOR, G.W., & COCHRAN, W.G., 1989. Statistical methods. USA: Iowa State University Press.

SOIL CLASSIFICATION WORKING GROUP, 1991. Soil classification: a taxonomic system for South Africa. Pretoria, South Africa: Department of Agricultural Development.

SOUTH, J.B., 1982. Selecting an acceptance sampling plan that minimizes expected error and sampling cost. *Quality Progress*, Vol. 15, No. 10: 18–22.

TOM, V.W., 2000. Mapping With multibeam data: are there ideal model settings? Unpublished Document. University of Ghent, Geography Department.

TOMCZAK, M., 1998. Spatial interpolation and its uncertainty using automated anisotropic Inverse Distance Weighting (IDW) cross validation/jackknife approach. *Journal of Geographic Information and Decision Analysis*, Vol. 2, No. 2: 18–33.

VAGEN Tor-G., SHEPHERD, K.D., & WALSH, M.G., 2006. Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy. *Geoderma,* Vol. 133, No. 3–4: 281–294.

VAGEN Tor-G., WALSH, M.G., & SHEPHERD, K.D., 2006. Stable isotopes for characterization of trends in soil carbon following deforestation and land use in the highlands of Madagascar. G*eoderma*, Vol. 135: 133–139.

VAN GROENIGEN, J.W., 2000. The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma*, Vol. 97: 223–236.

VAN TOL, J.J., 2008. Soil indicators of hillslope hydrology in the Bedford and Weatherly catchments. Unpublished M.Sc. Thesis. Bloemfontein, South Africa: University of the Free Sate.

WATERS, N.M., 1988. Expert systems and systems of experts. In Coffey, W.J. (ed.). Geographical systems and systems of geography: essays in honour of William Warntz. Ontario: Department of Geography, University of Western Ontario.

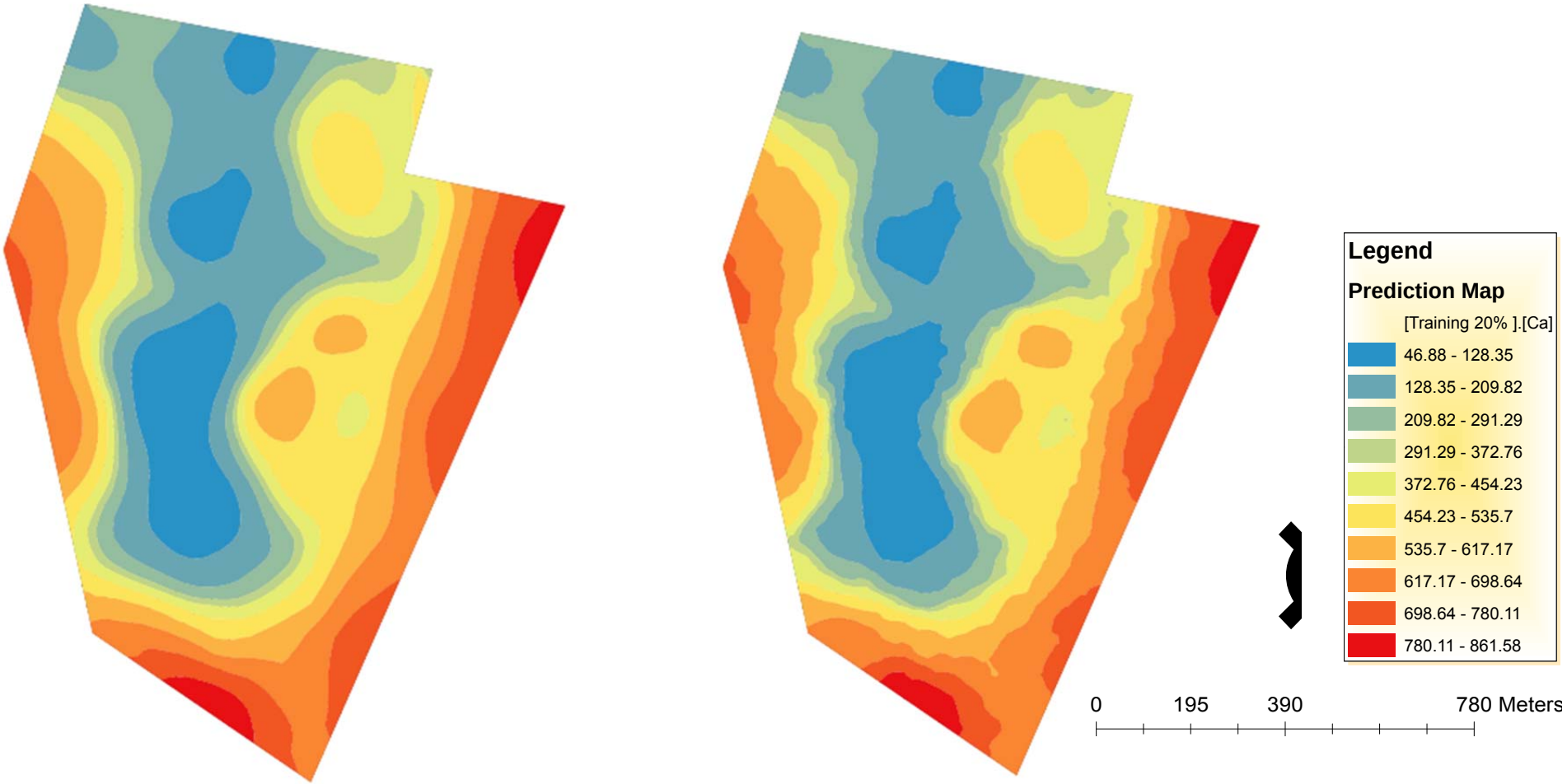
WENJAIO, S., LIU, J., DU, Z., SONG, Y., CHEN, C., & YUE, T., 2009. Surface modeling of soil pH. *Geoderma*, Vol. 150, No. 1–2: 113–119.

YEN, L., ZHOU, S., CI-FANG, W., FENG, L., & HONG-YI, L., 2007. Optimized spatial sampling scheme for soil electrical conductivity based on Variance Quad-Tree (VQT) method. *Agricultural Sciences in China*, Vol. 6, No. 12: 1463–1471.

ZHONG, W., KOKUBO, S., & TANIMOTO, J., 2012. How is the equilibrium of continuous strategy different from that of discrete strategy game? *Biosystems*, Vol. 107, No. 2: 88–94.

Appendix A

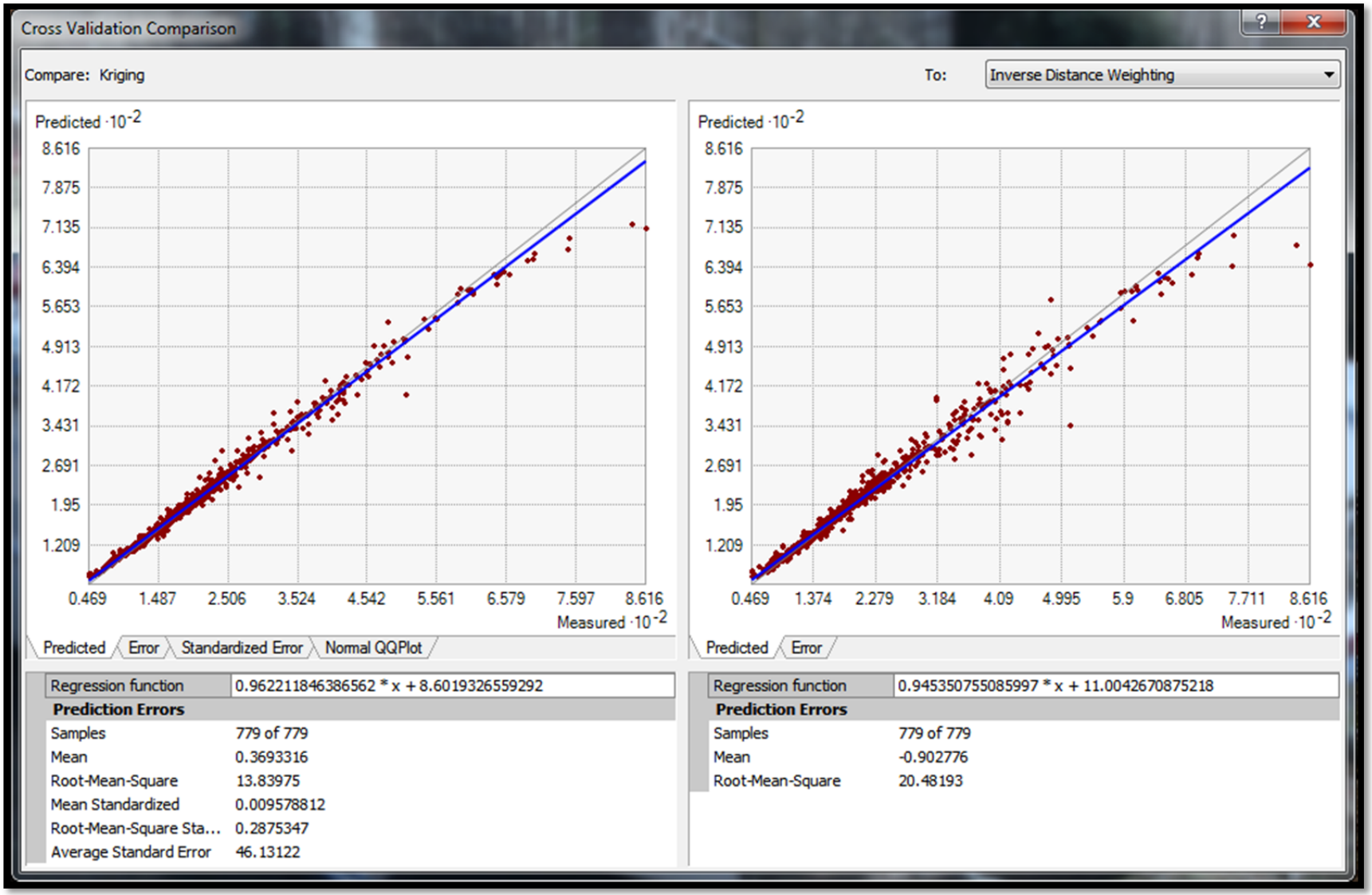
Appendix A contains: Prediction maps of randomly selected training and testing data sets for the soil chemical properties (Ca, K, Mg, Na and pH) as well as a cross-validation scatter plot graphs comparison of ordinary kriging and IDW predicted error for the random data set.



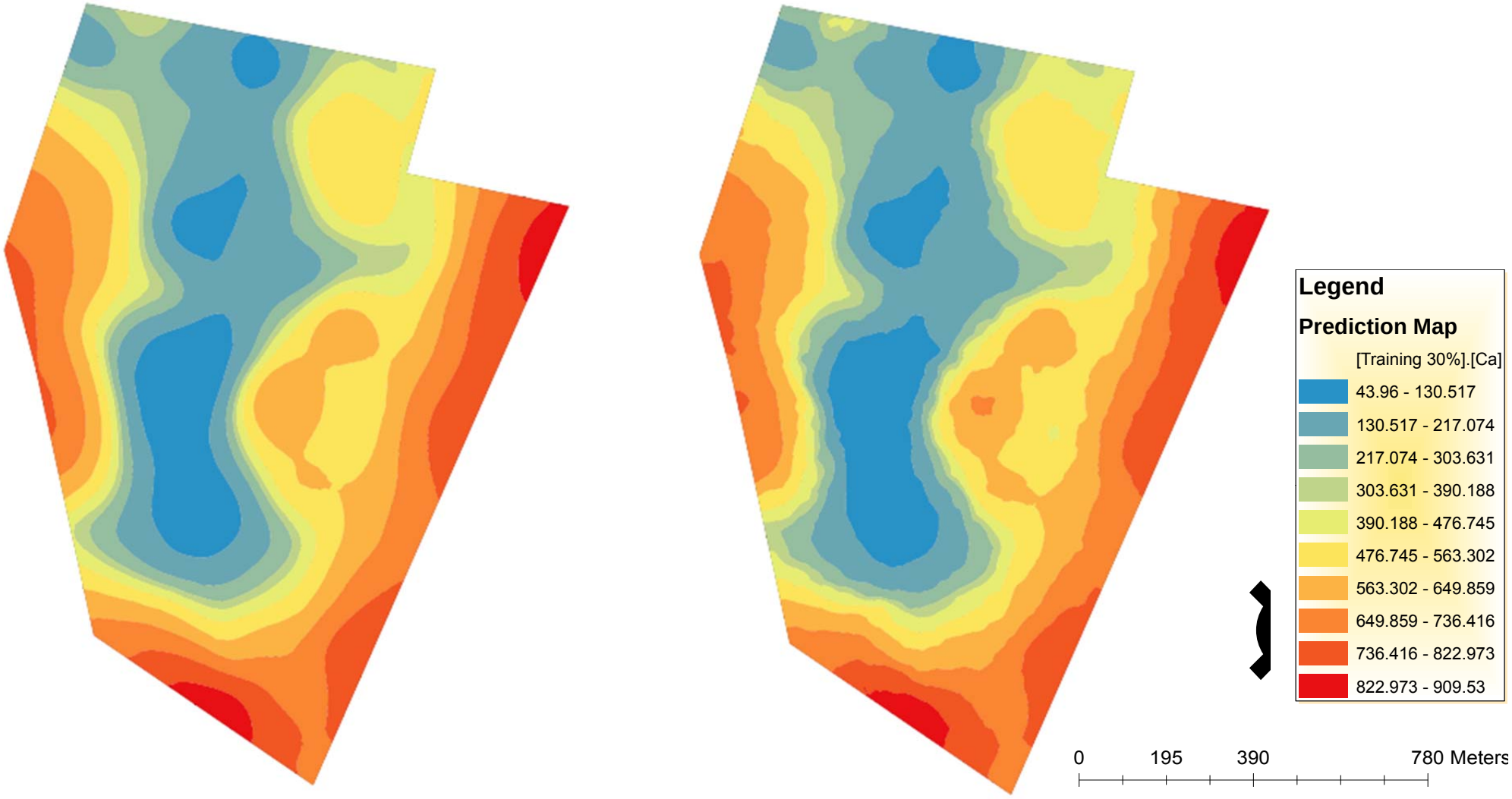
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 22: Prediction maps of randomly selected 20% training and 80% testing data sets for Calcium (Ca).



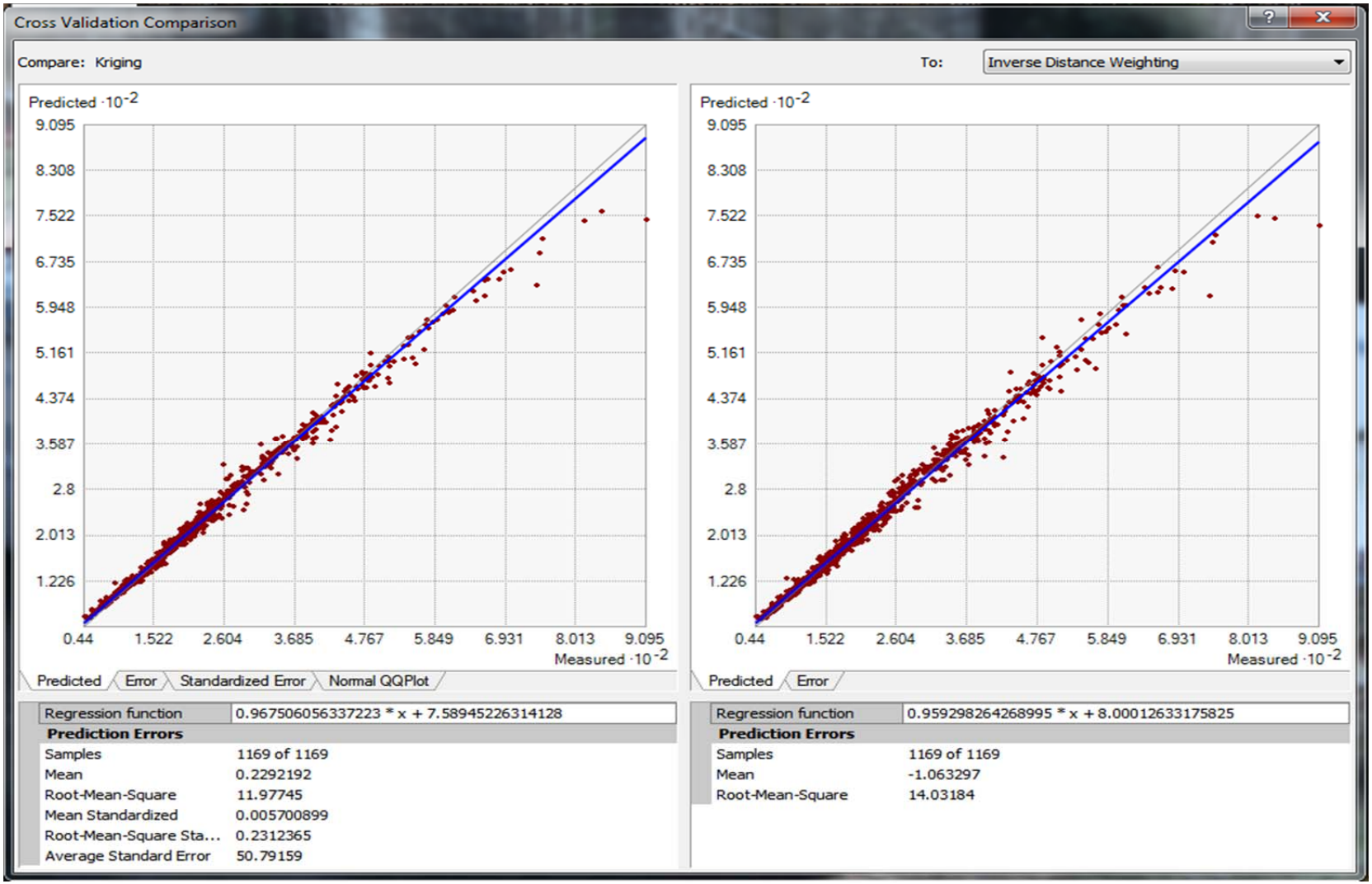
Graph 7: Cross-validation comparison of predicted error for the randomly selected 20% Ca training data set.



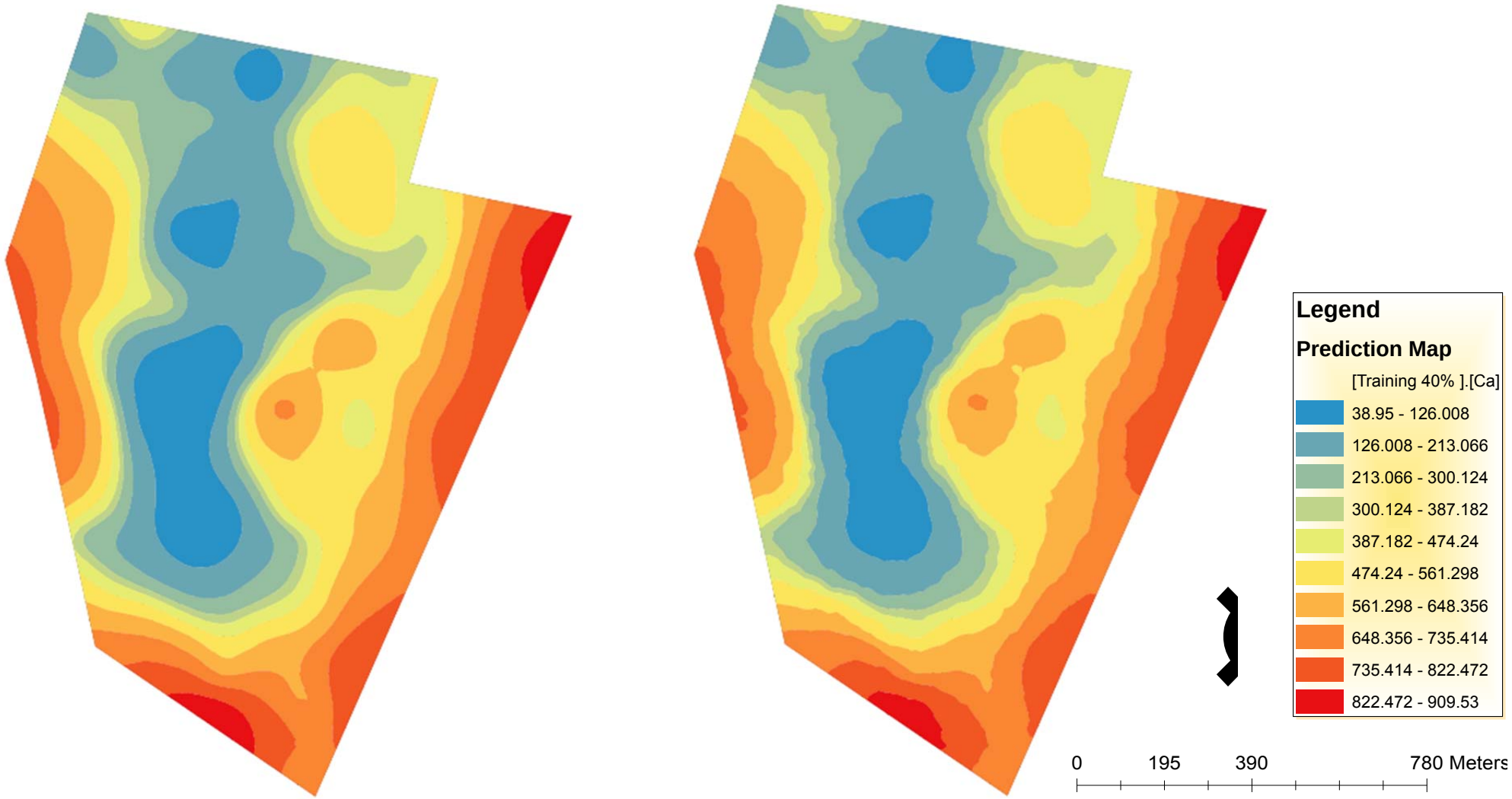
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 23: Prediction maps of randomly selected 30% training and 70% testing data sets for Calcium (Ca).



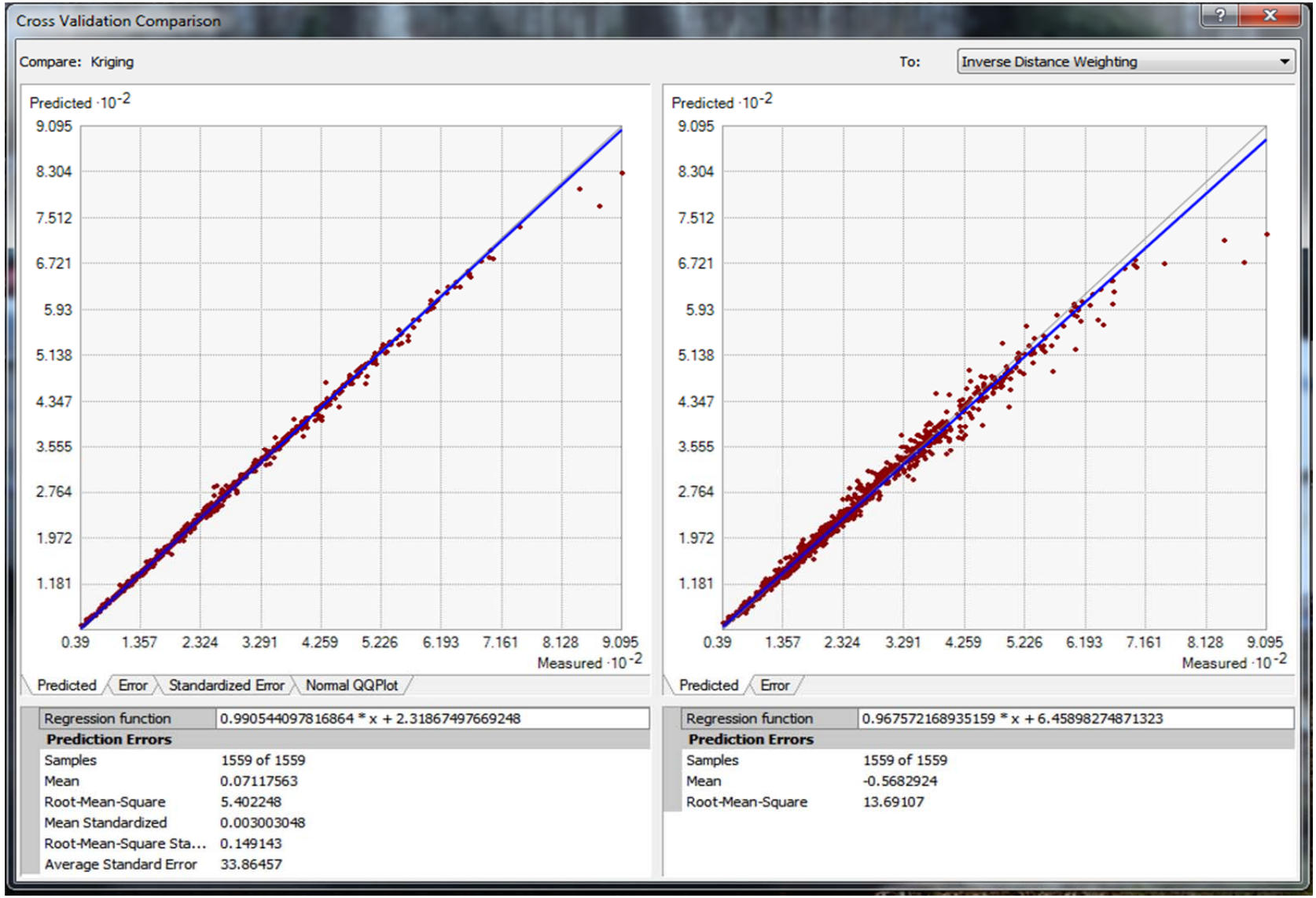
Graph 8: Cross-validation comparison of predicted error for the randomly selected 30% Ca training data set.



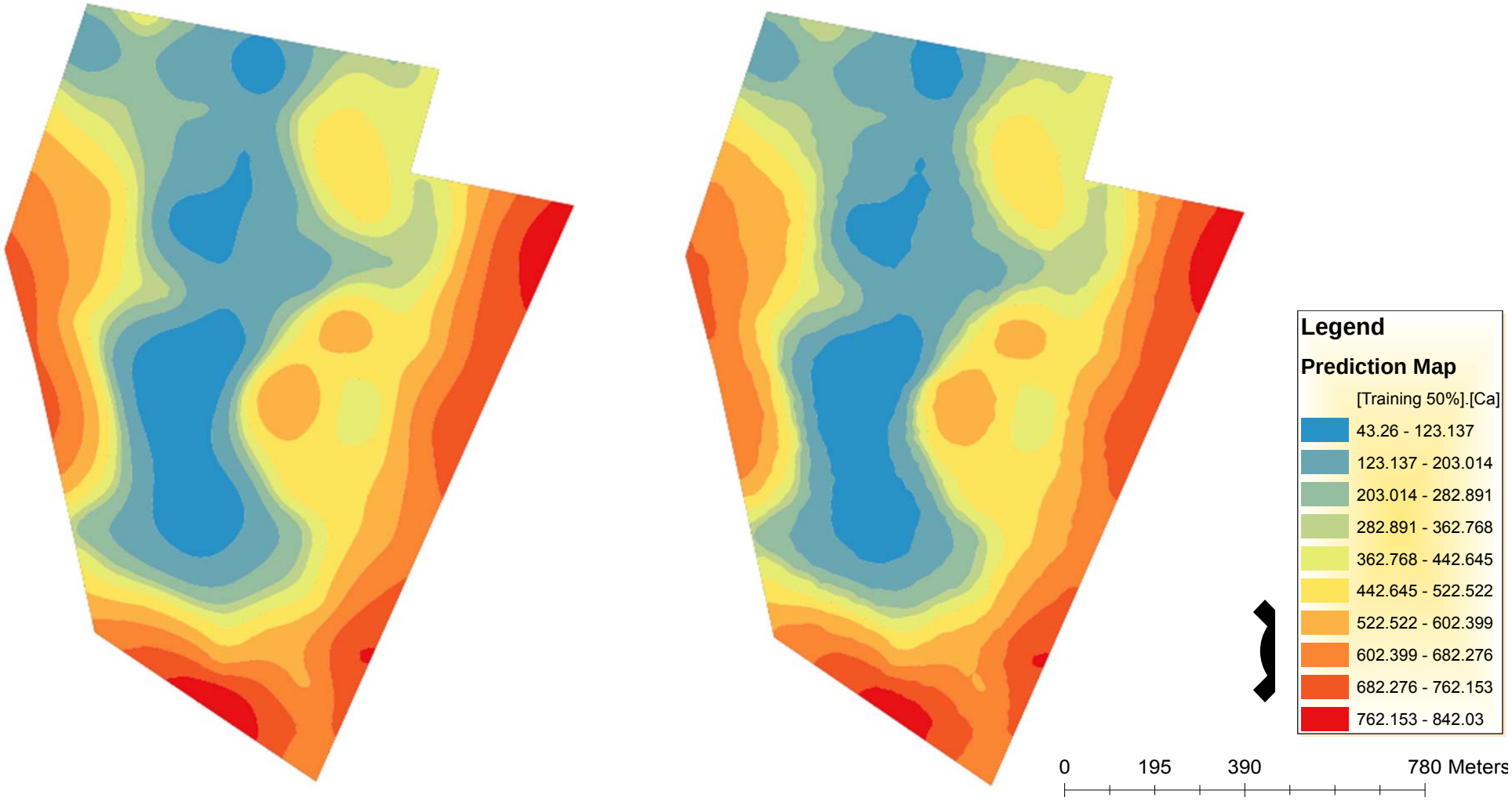
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 24: Prediction maps of randomly selected 40% training and 60% testing data sets for Calcium (Ca).



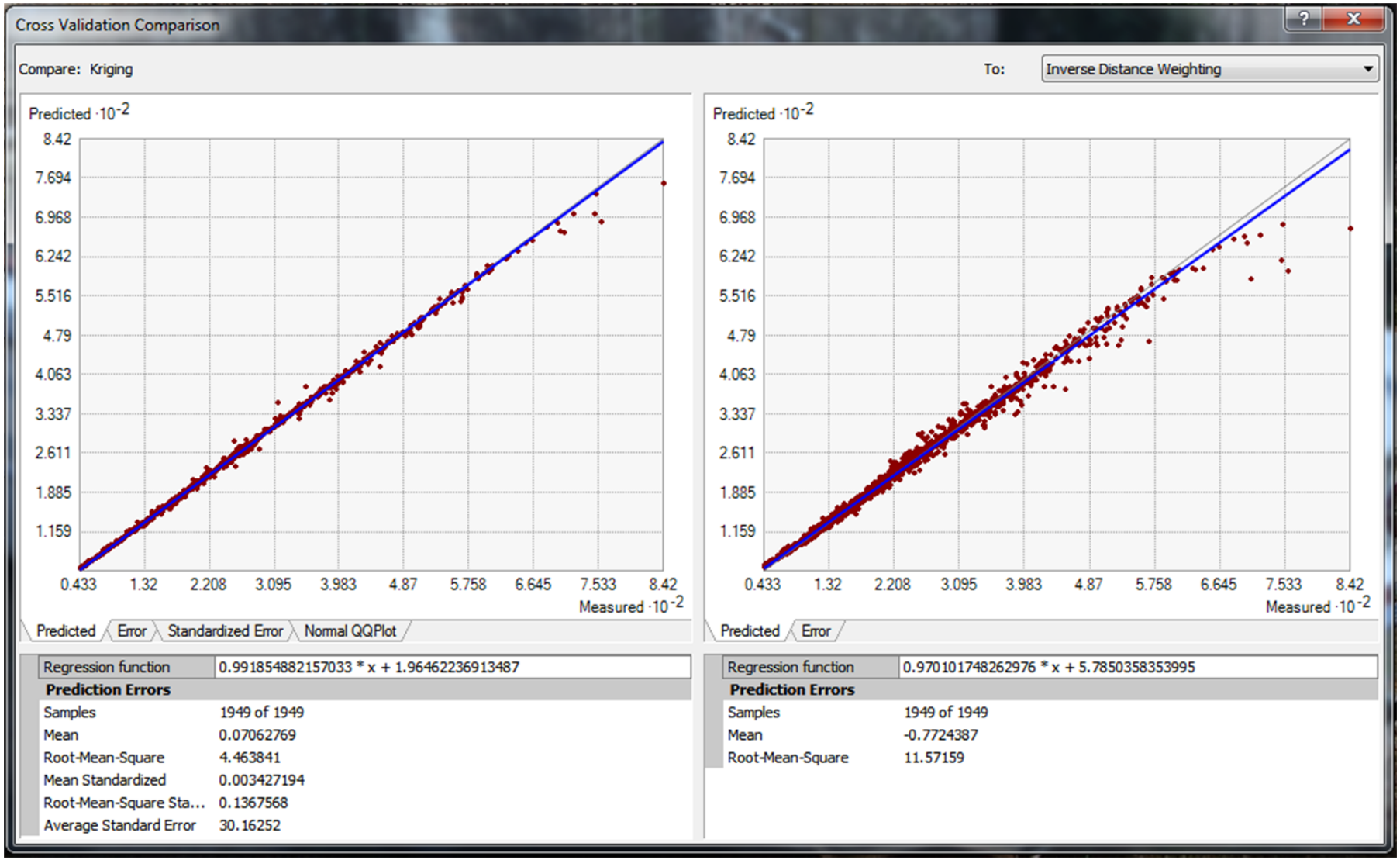
Graph 9: Cross-validation comparison of predicted error for the randomly selected 40% Ca training data set.



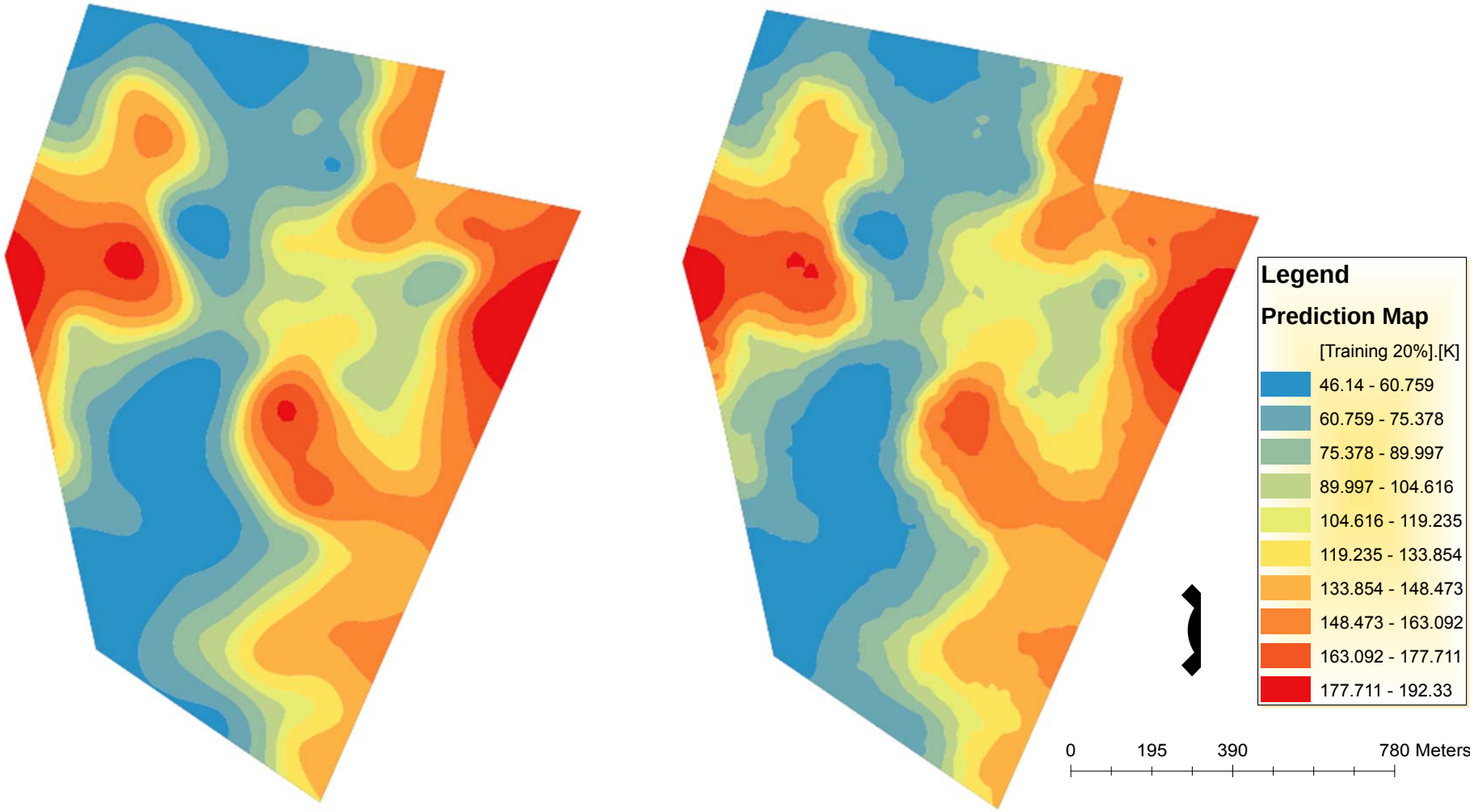
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 25: Prediction maps of randomly selected 50% training and 50% testing data sets for Calcium (Ca).



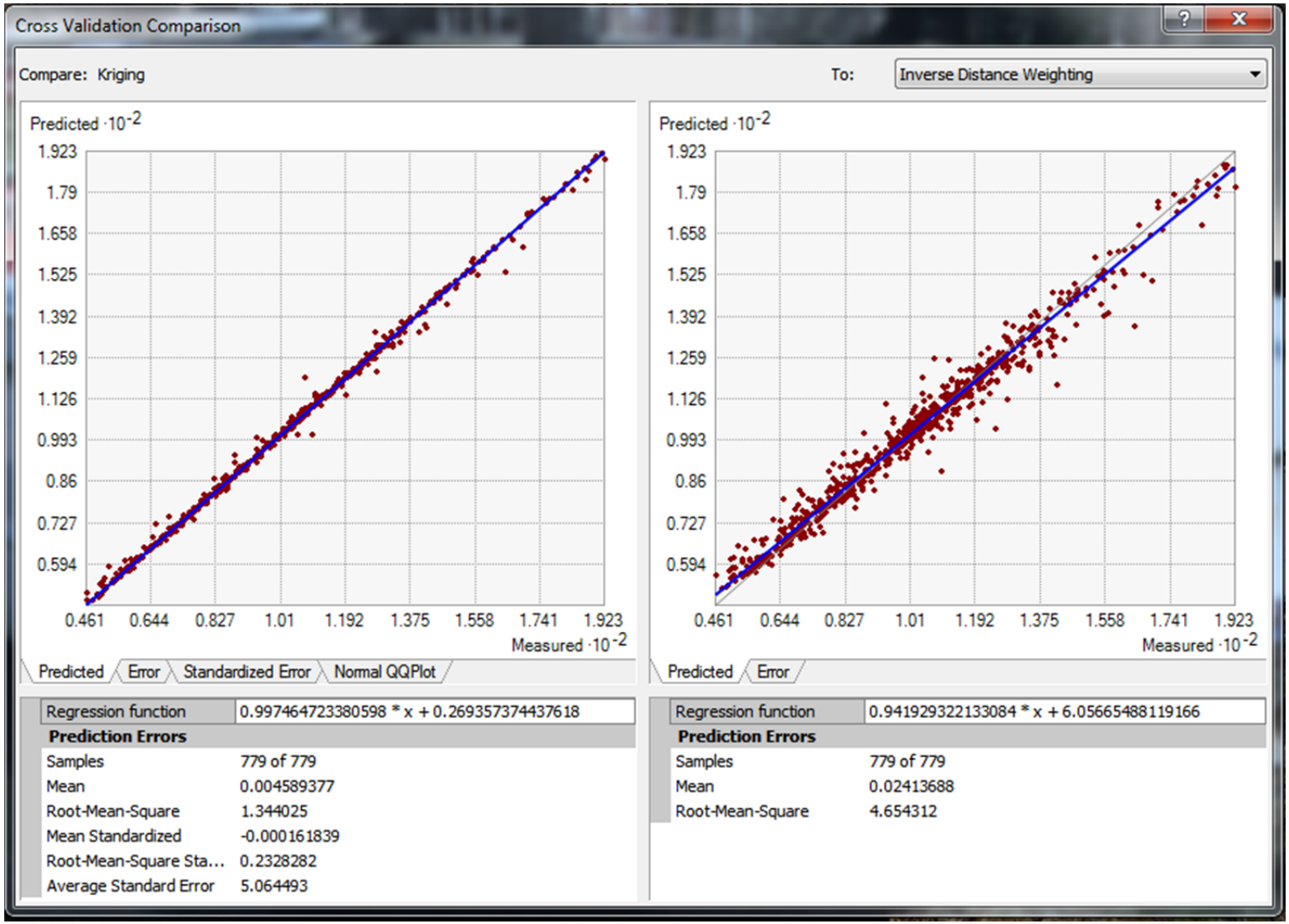
Graph 10: Cross-validation comparison of predicted error for the randomly selected 50% Ca training data set.



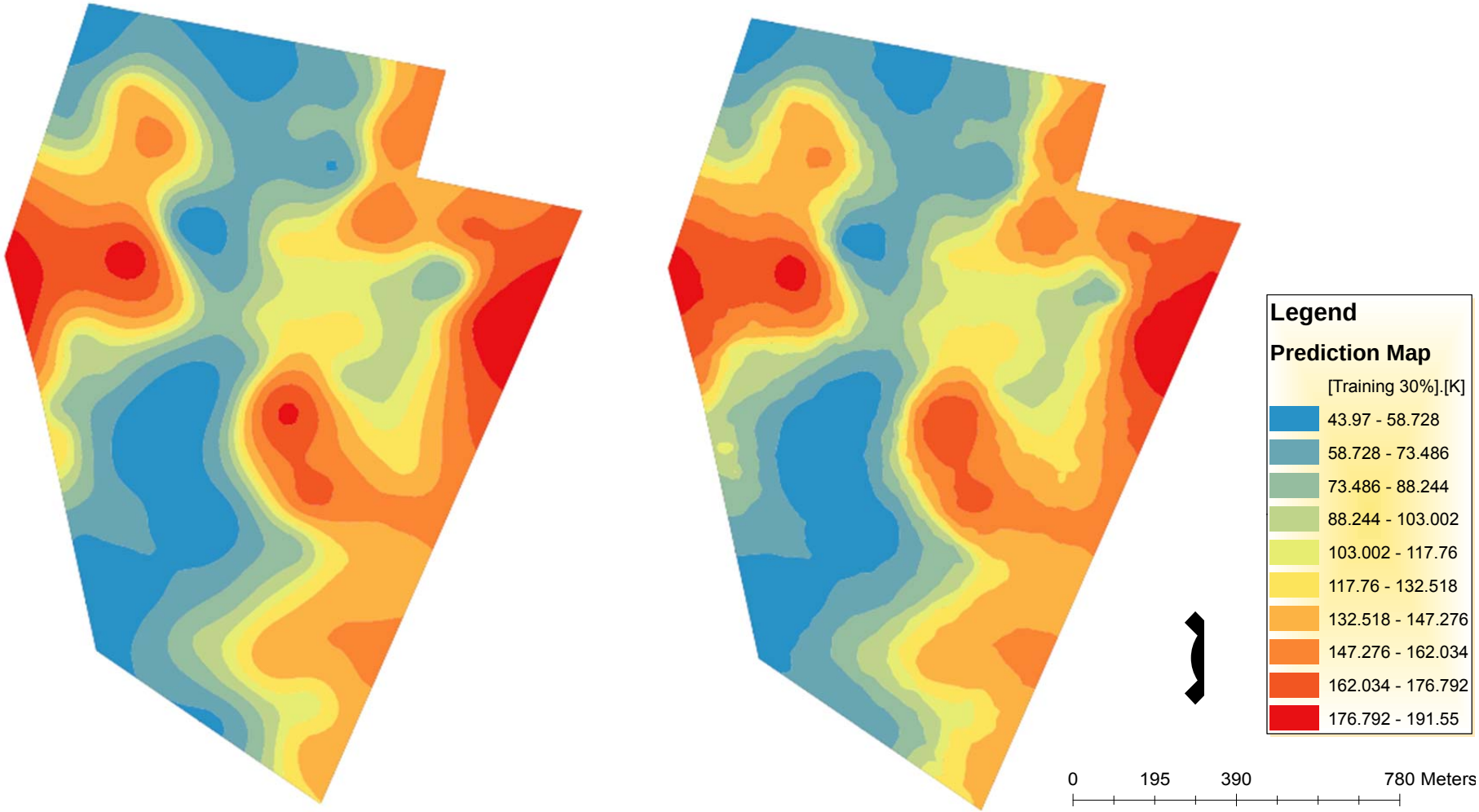
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 26: Prediction maps of randomly selected 20% training and 80% testing data sets for Potassium (K).



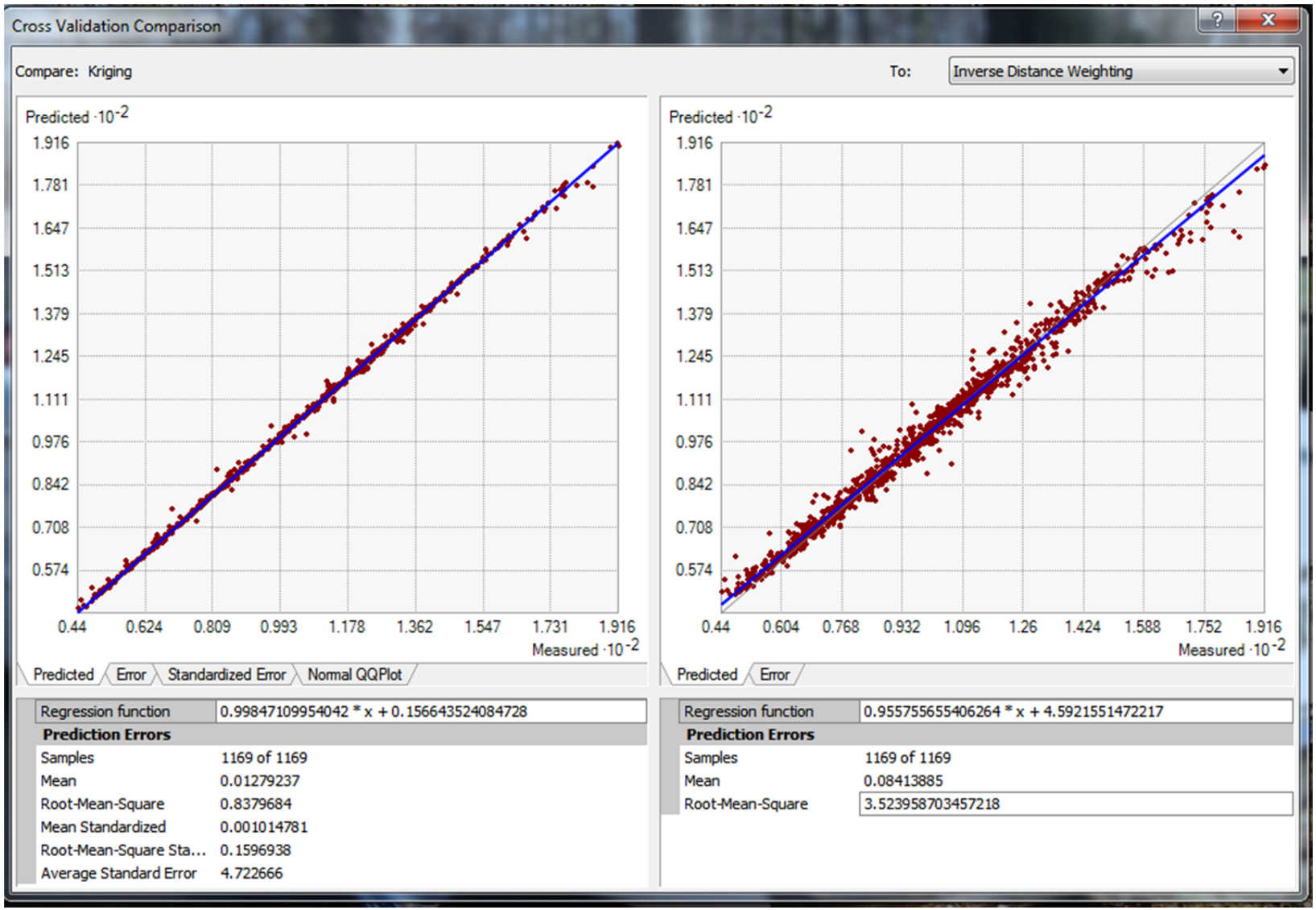
Graph 11: Cross-validation comparison of predicted error for the randomly selected 20% K training data set.



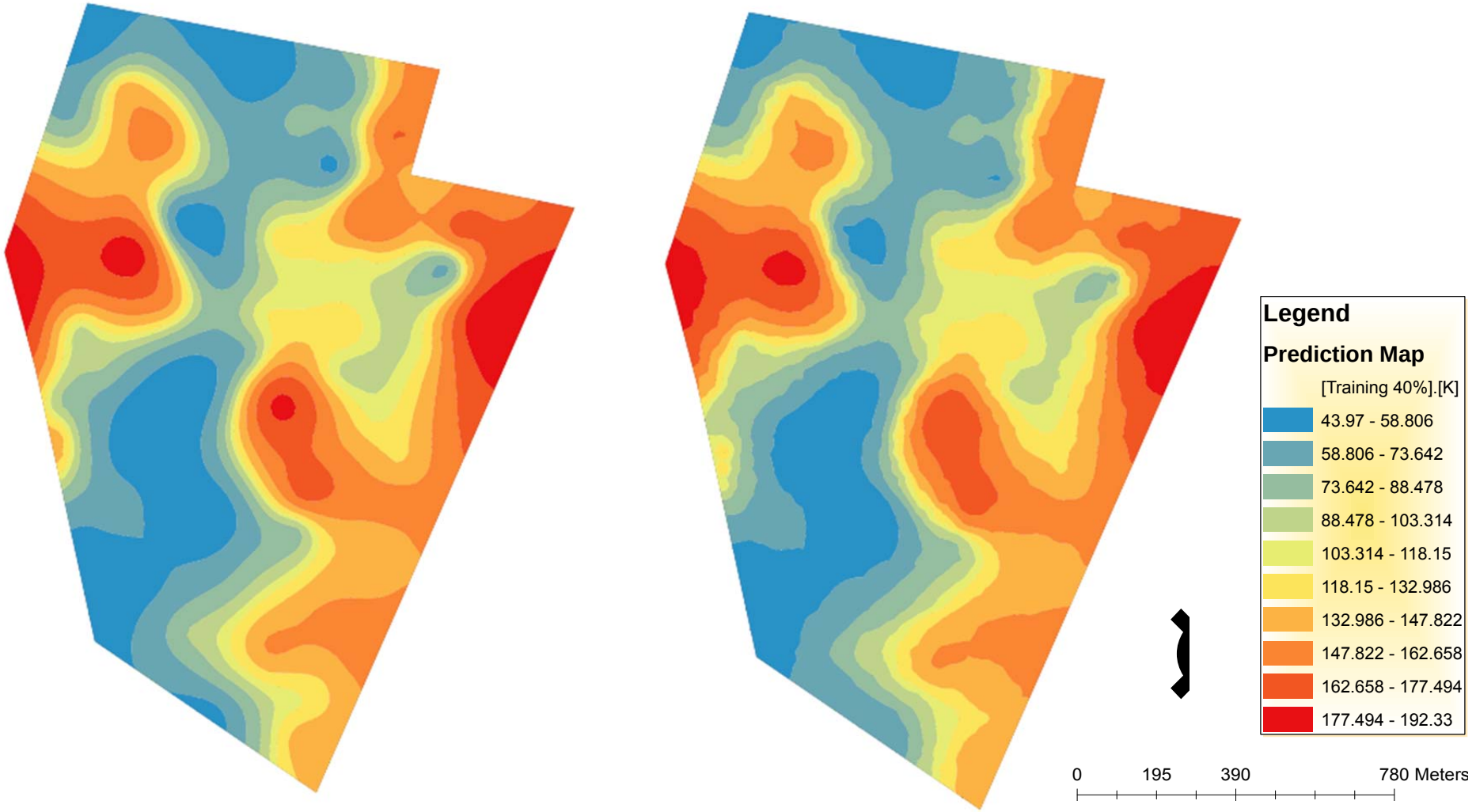
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 27: Prediction maps of randomly selected 30% training and 70% testing data sets for Potassium (K).



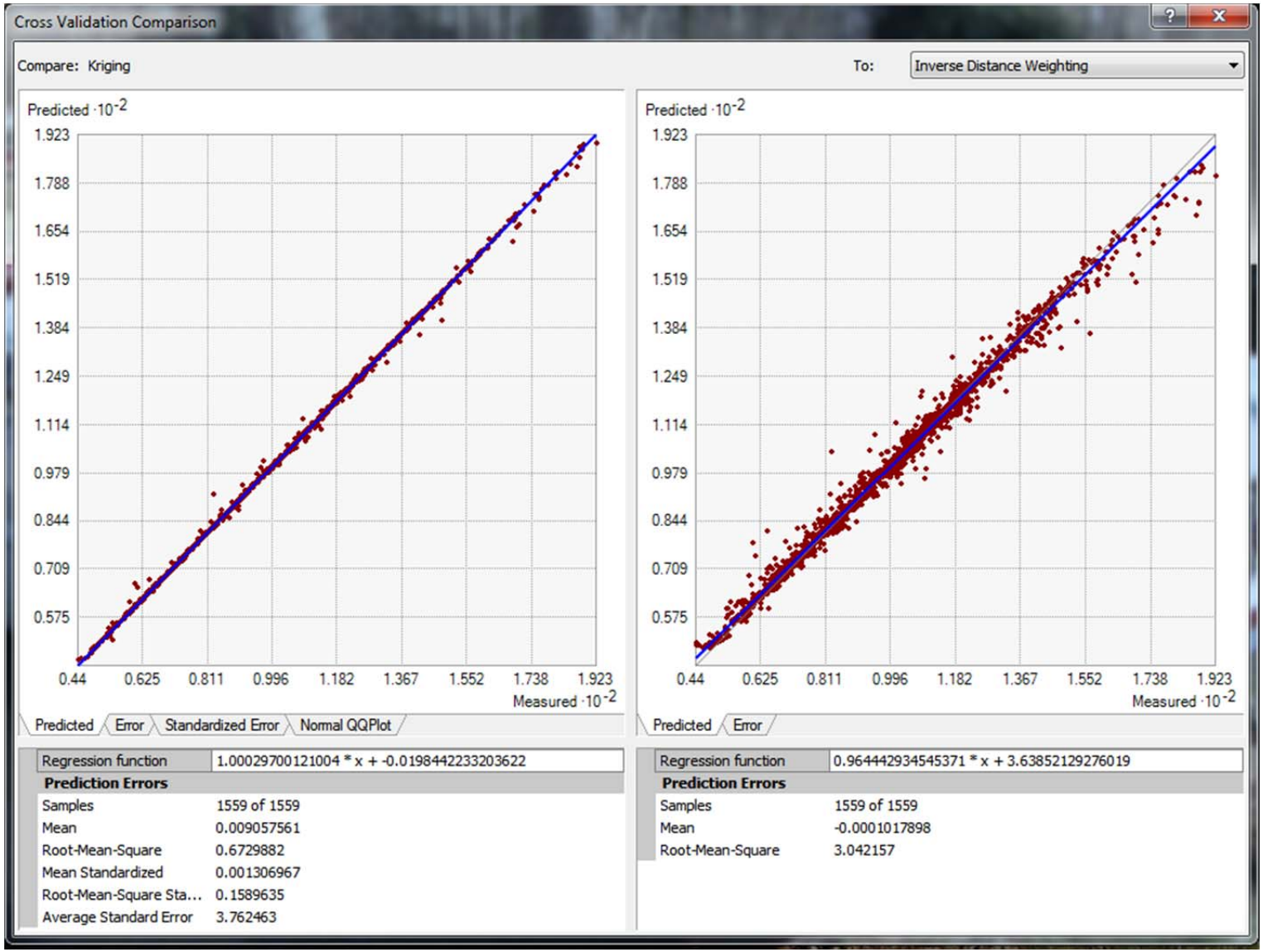
Graph 12: Cross-validation comparison of predicted error for the randomly selected 30% K training data set.



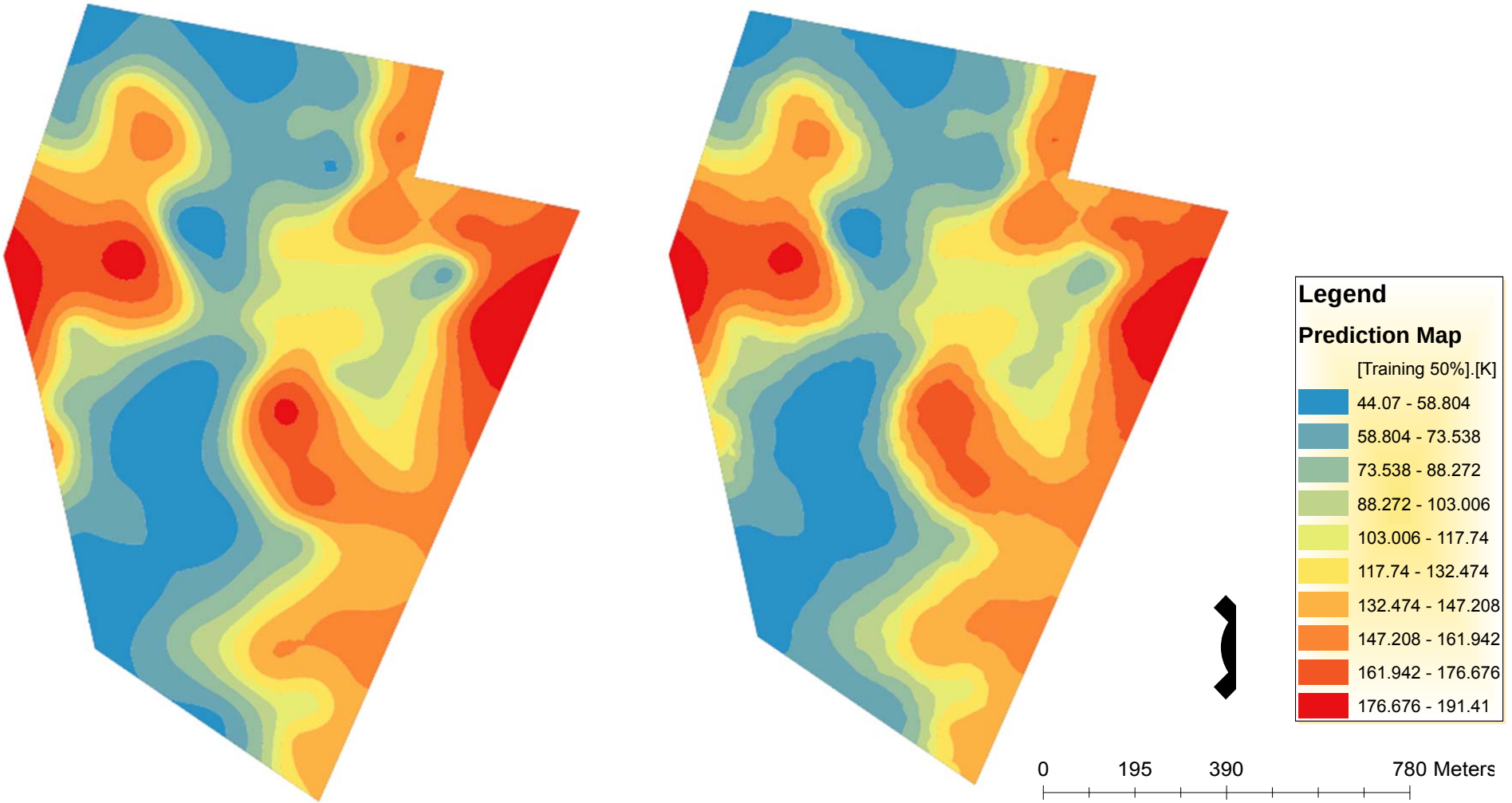
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 28: Prediction maps of randomly selected 40% training and 60% testing data sets for Potassium (K).



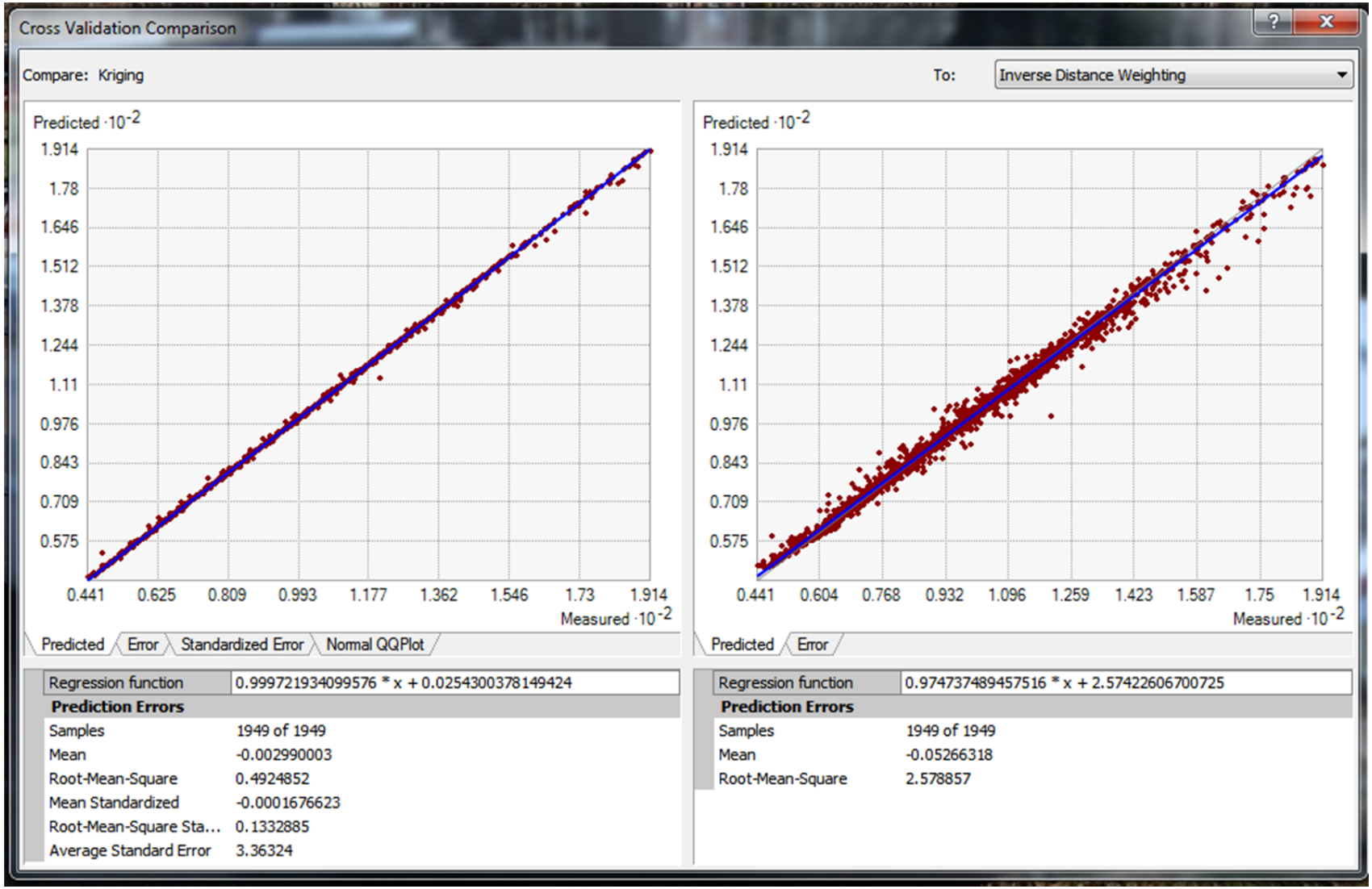
Graph 13: Cross-validation comparison of predicted error for the randomly selected 40% K training data set.



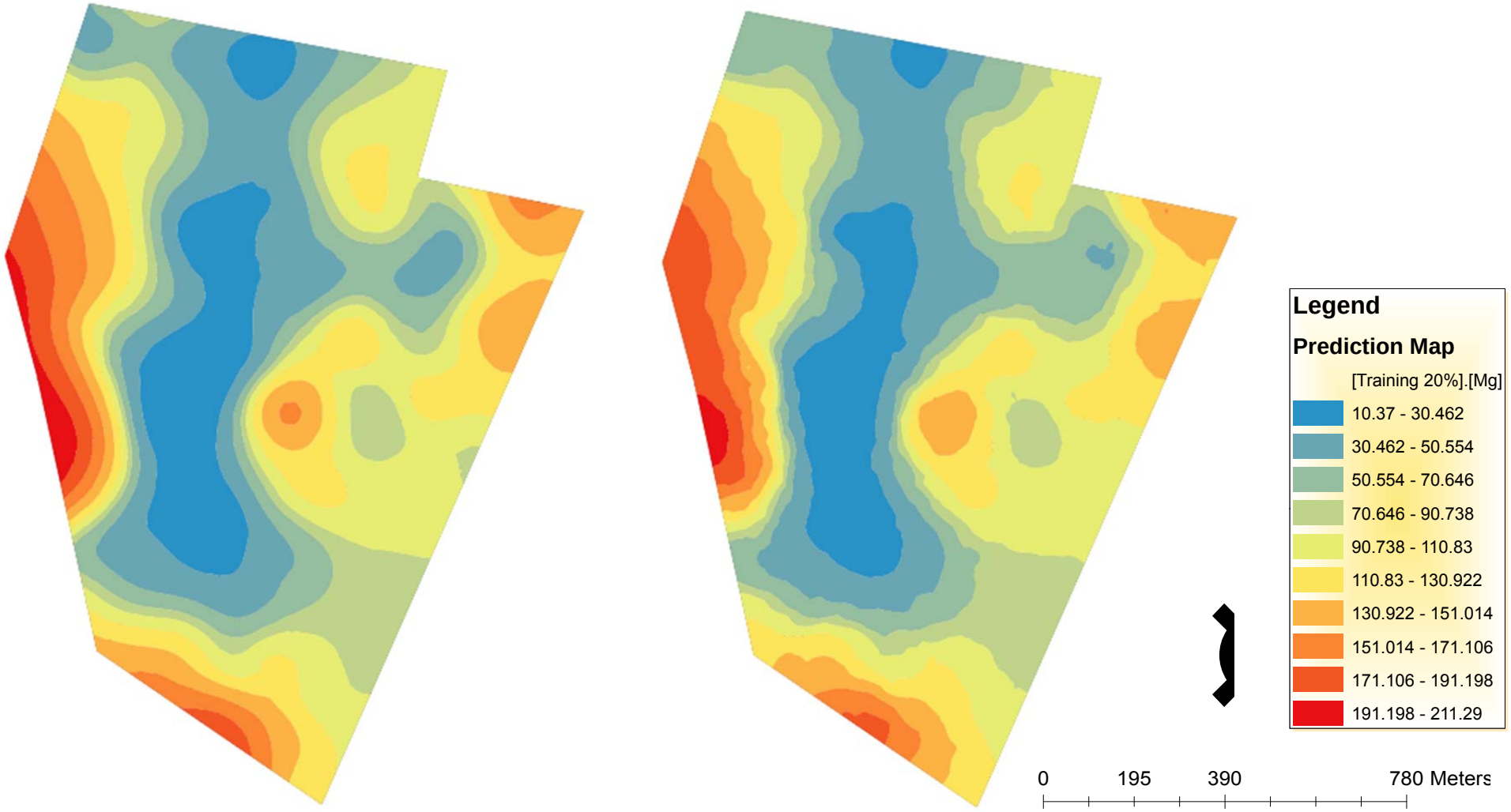
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 29: Prediction maps of randomly selected 50% training and 50% testing data sets for Potassium (K).



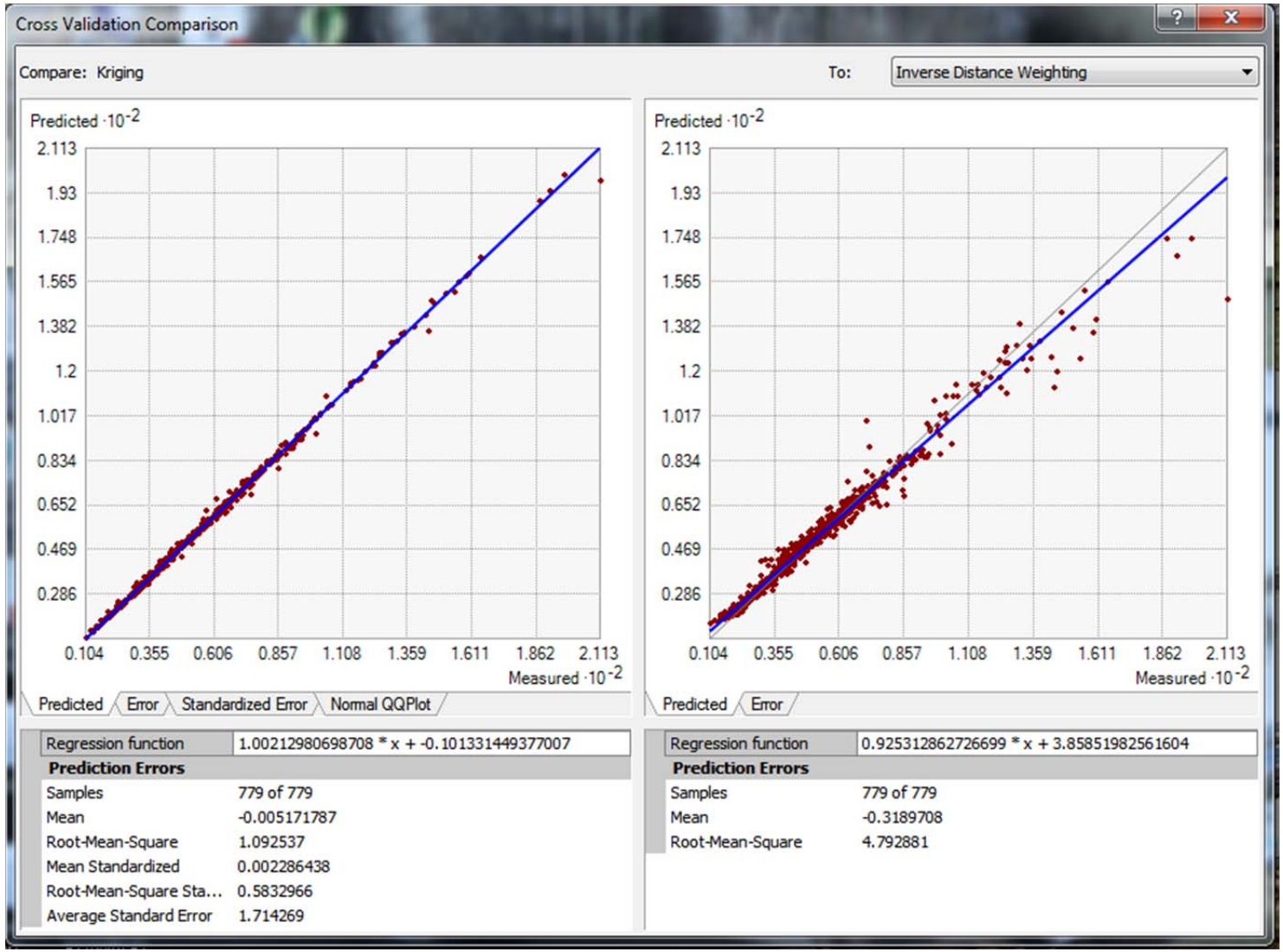
Graph 14: Cross-validation comparison of predicted error for the randomly selected 50% K training data set.



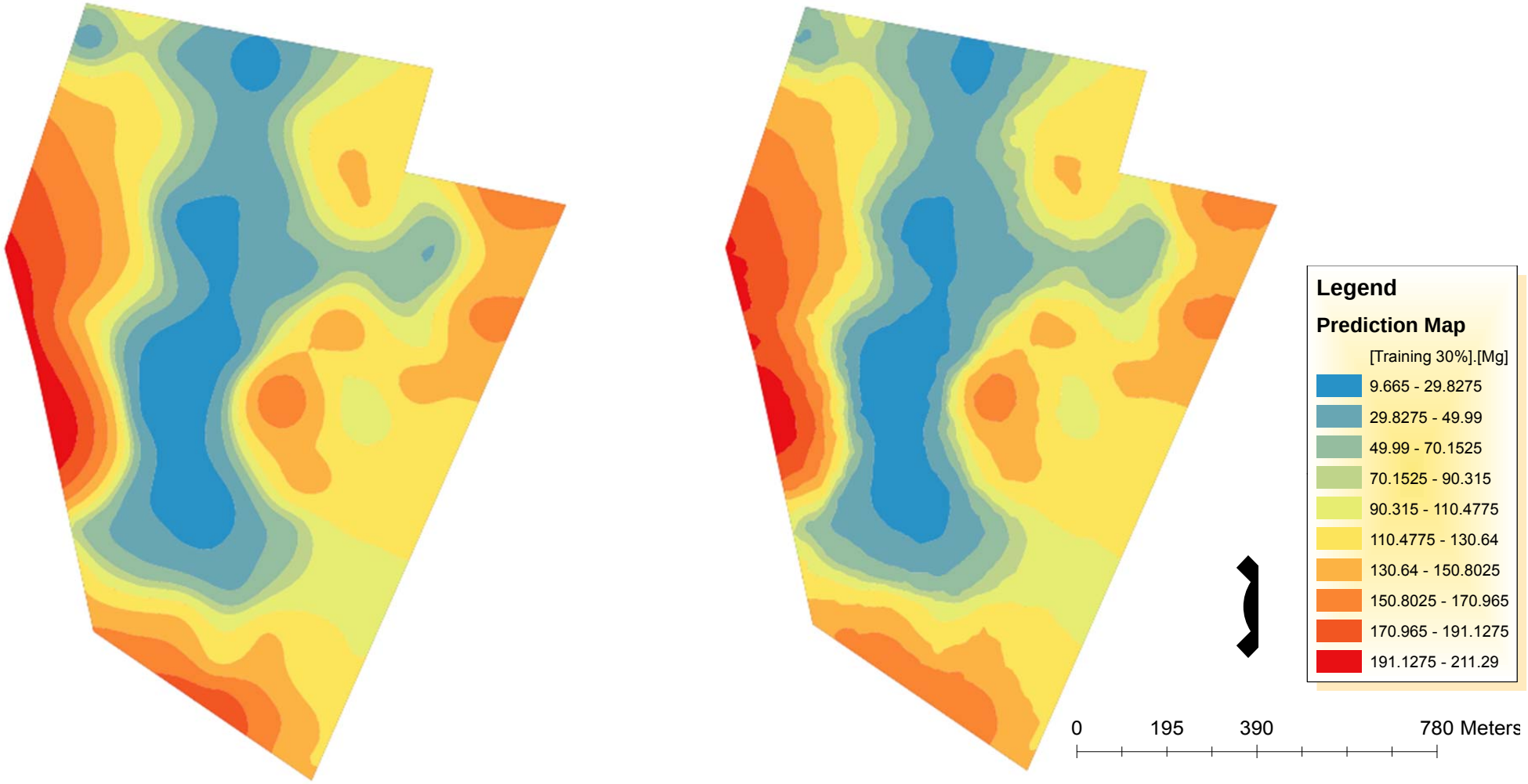
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 30: Prediction maps of randomly selected 20% training and 80% testing data sets for Magnesium (Mg).



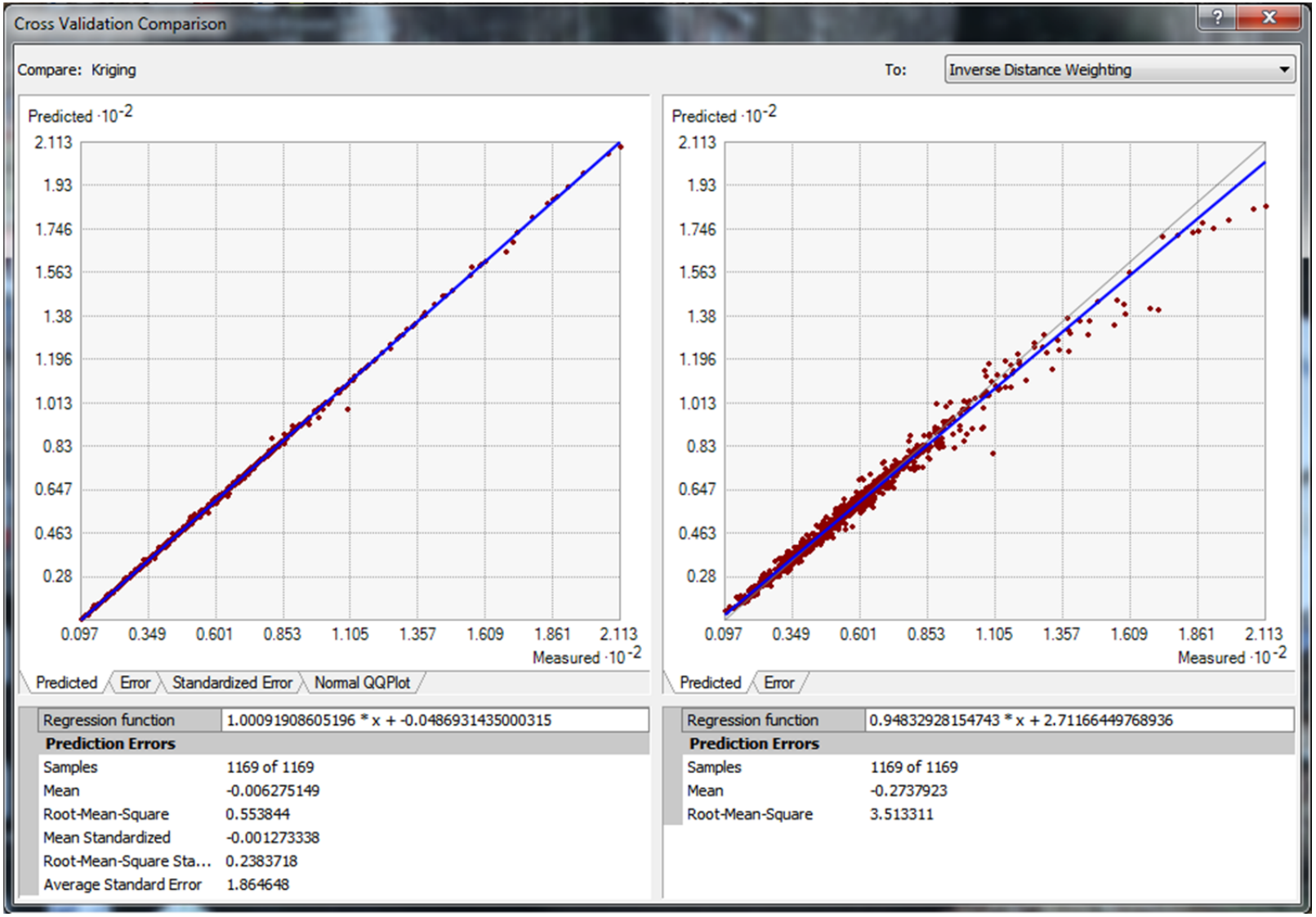
Graph 15: Cross-validation comparison of predicted error for the randomly selected 20% Mg training data set



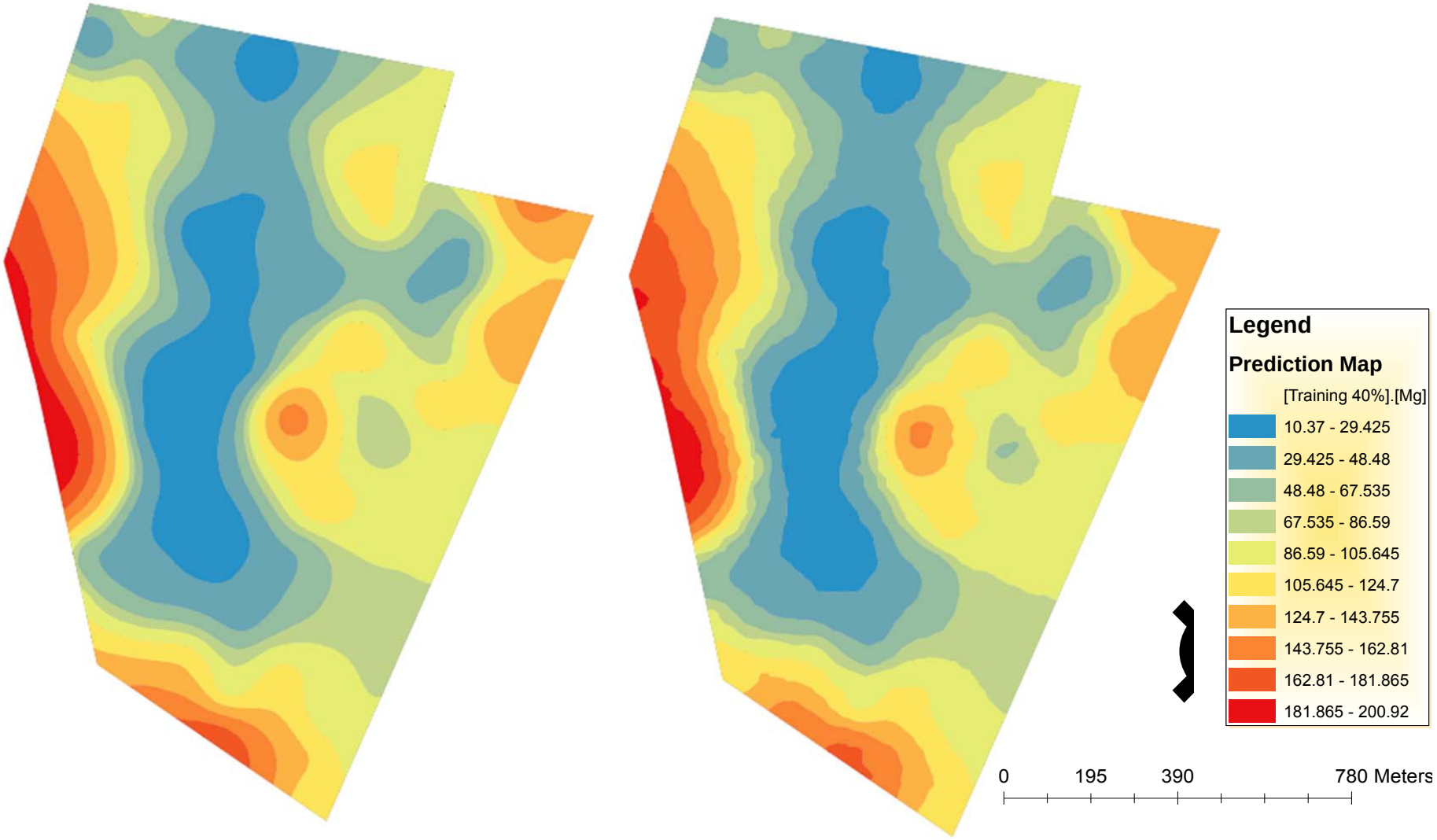
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 31: Prediction maps of randomly selected 30% training and 70% testing data sets for Magnesium (Mg).



Graph 16: Cross-validation comparison of predicted error for the randomly selected 30% Mg training data set.

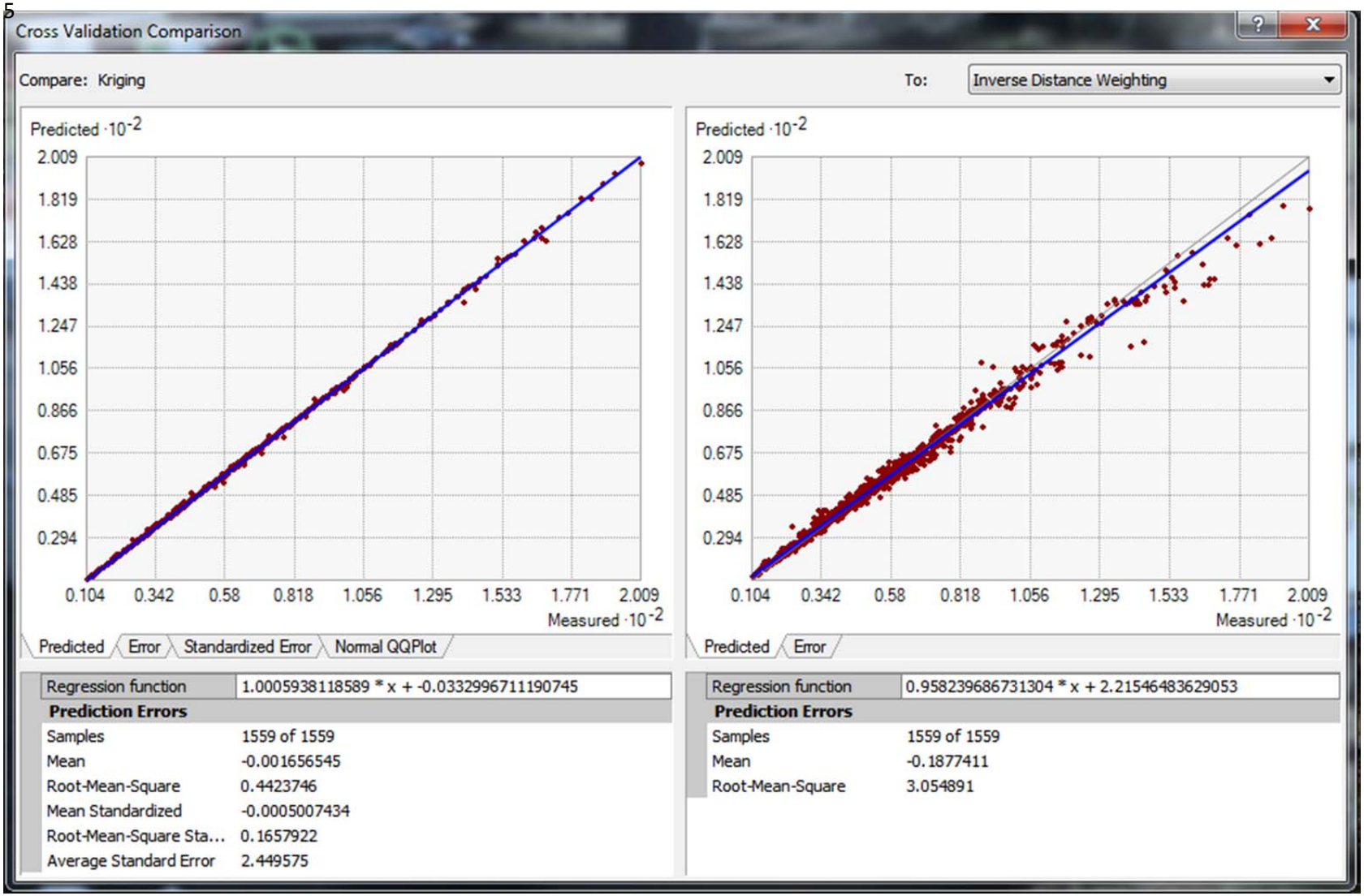


(a) Ordinary Kriging

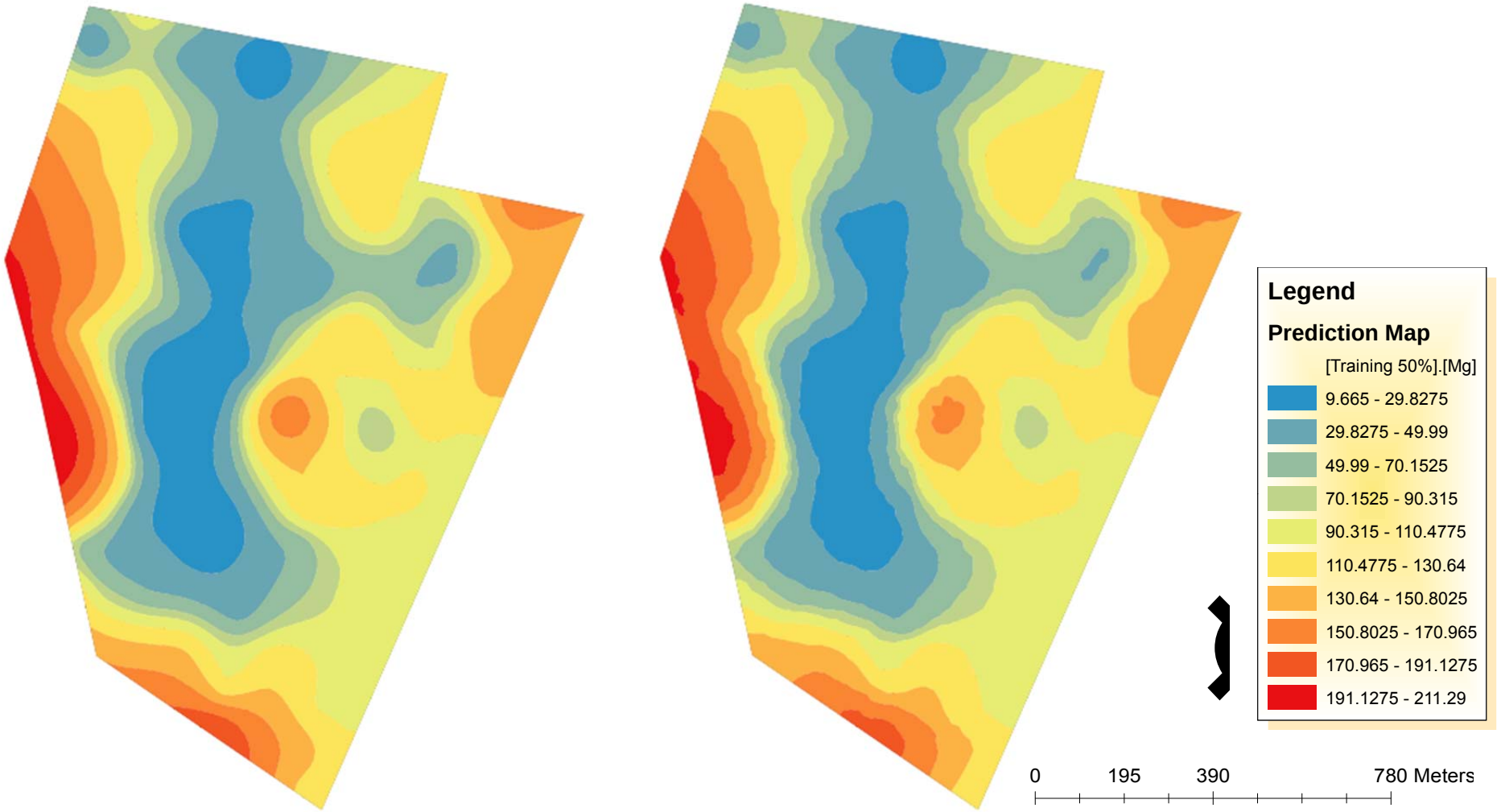
(b) Inverse Distance Weighting

Figure 32: Prediction maps of randomly selected 40% training and 60% testing data sets for Magnesium (Mg).

5



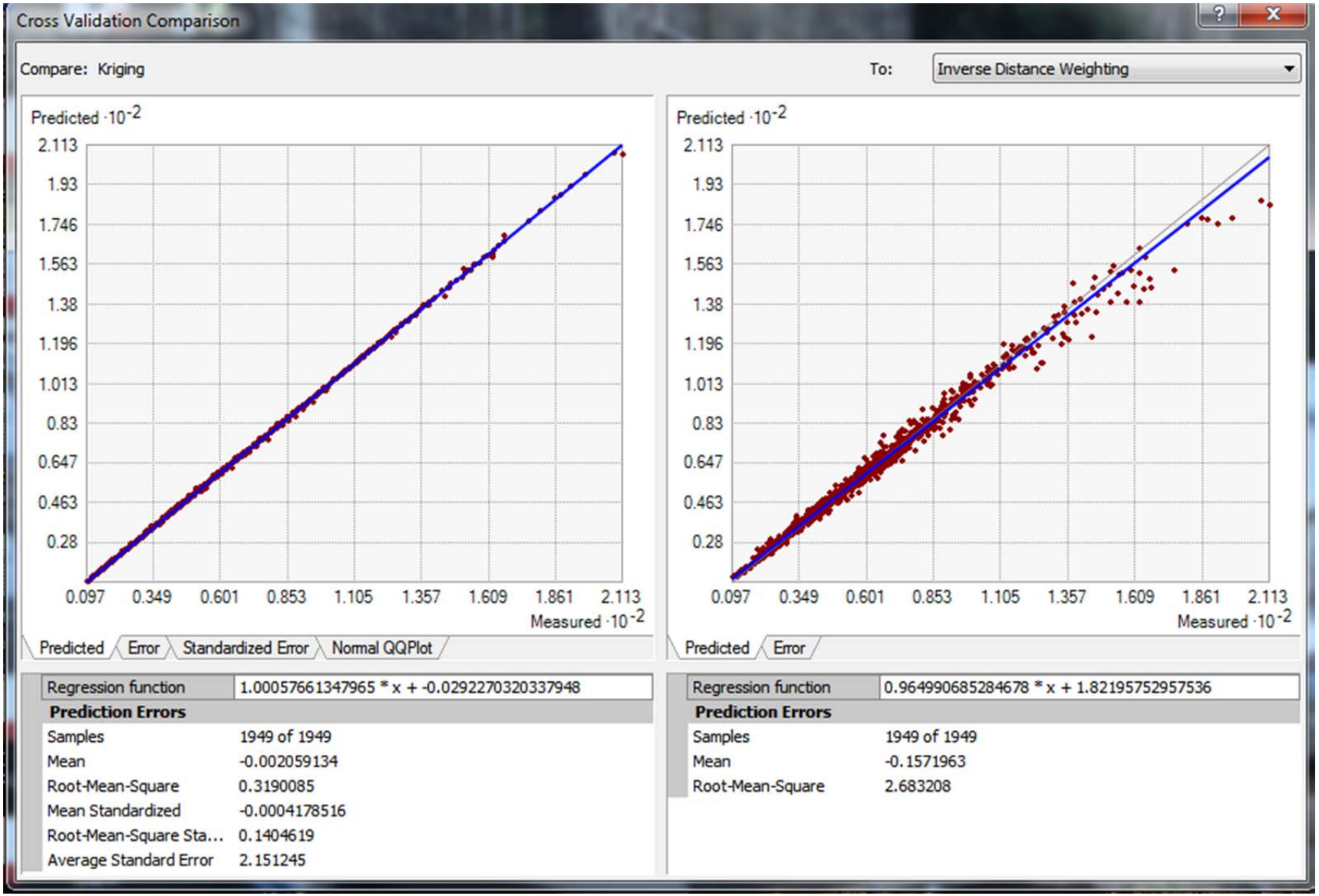
Graph 17: Cross-validation comparison of predicted error for the randomly selected 40% Mg training data set.



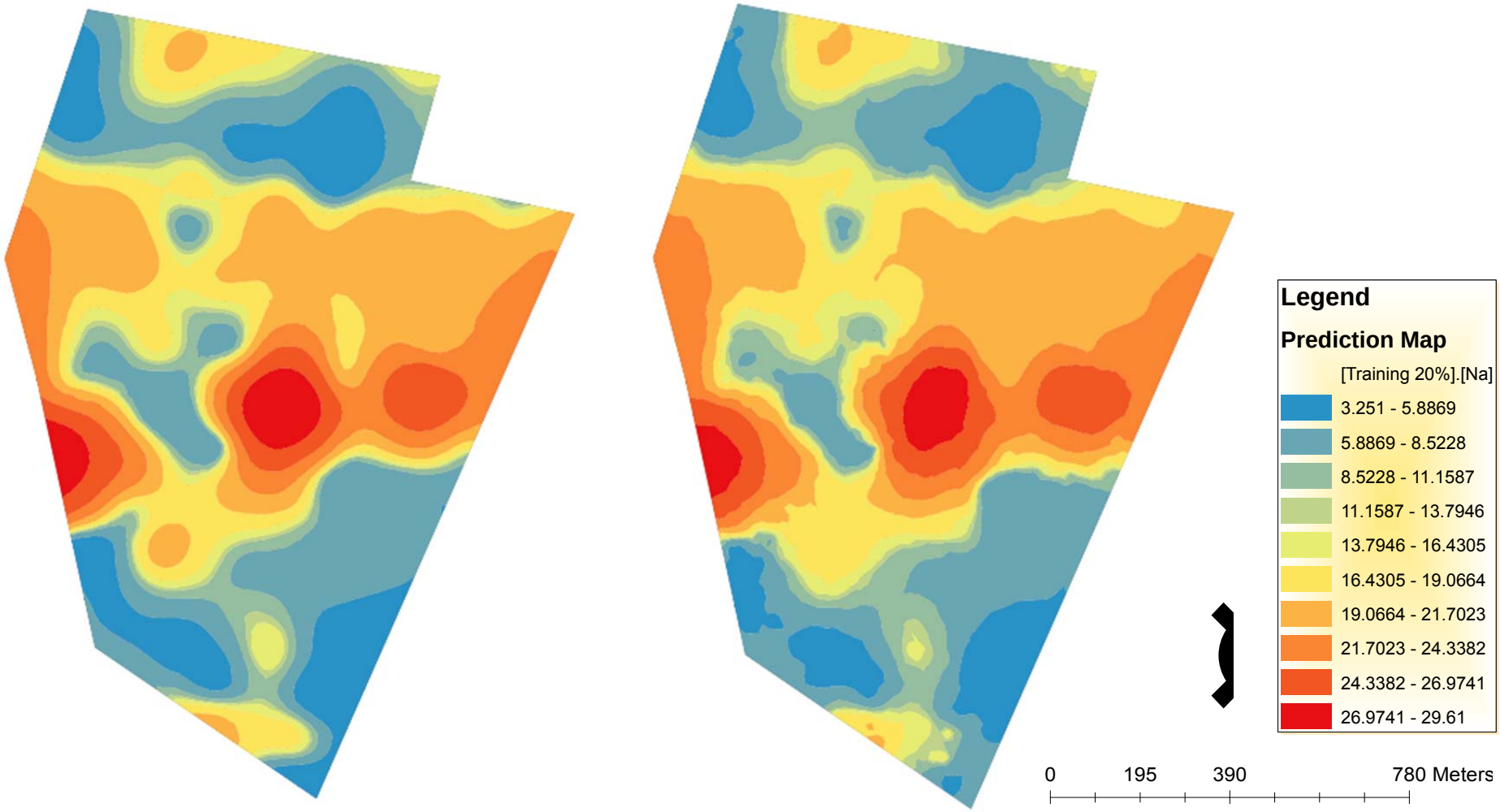
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 33: Prediction maps of randomly selected 50% training and 50% testing data sets for Magnesium (Mg).

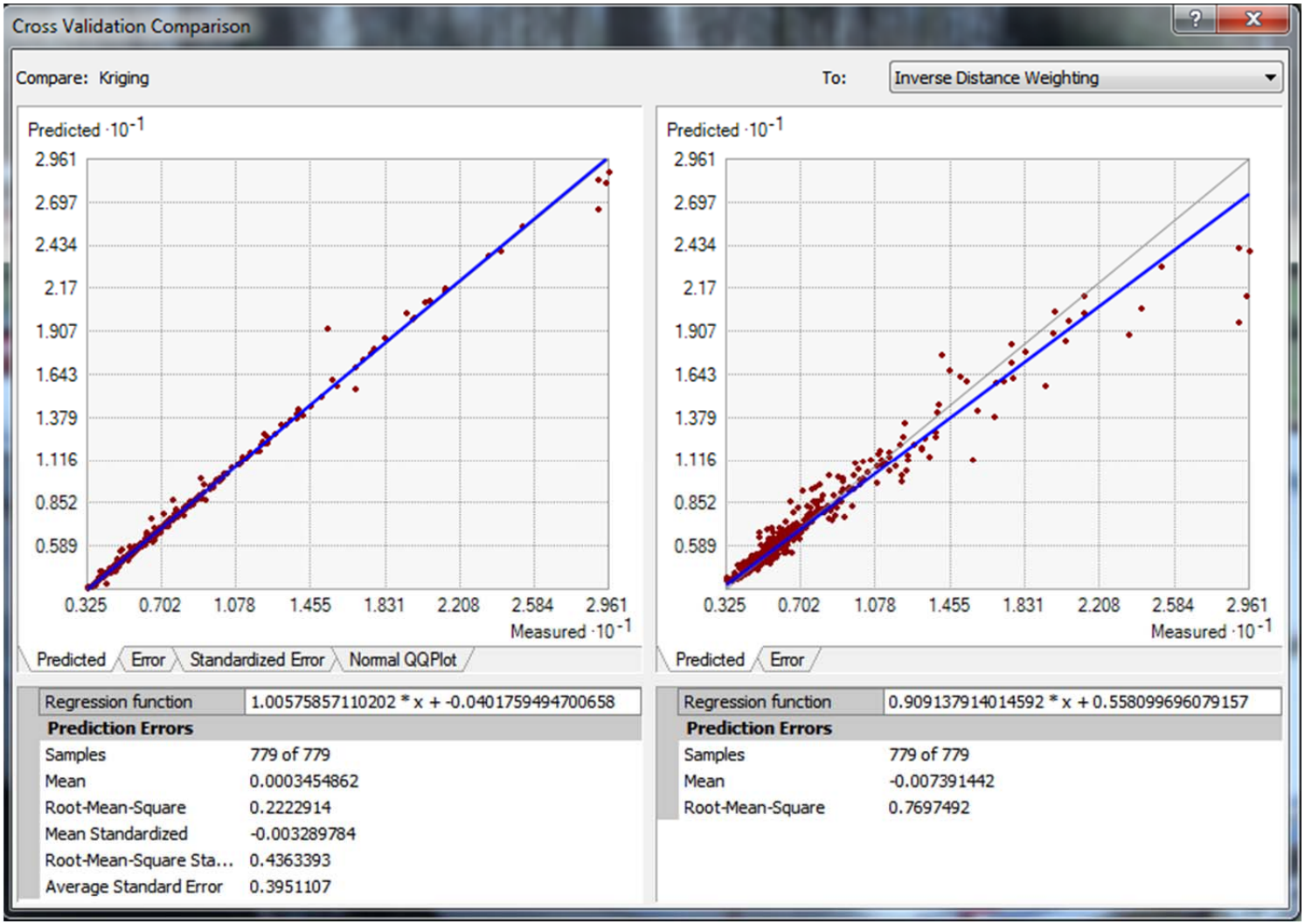


Graph 18: Cross-validation comparison of predicted error for the randomly selected 50% Mg training data set.

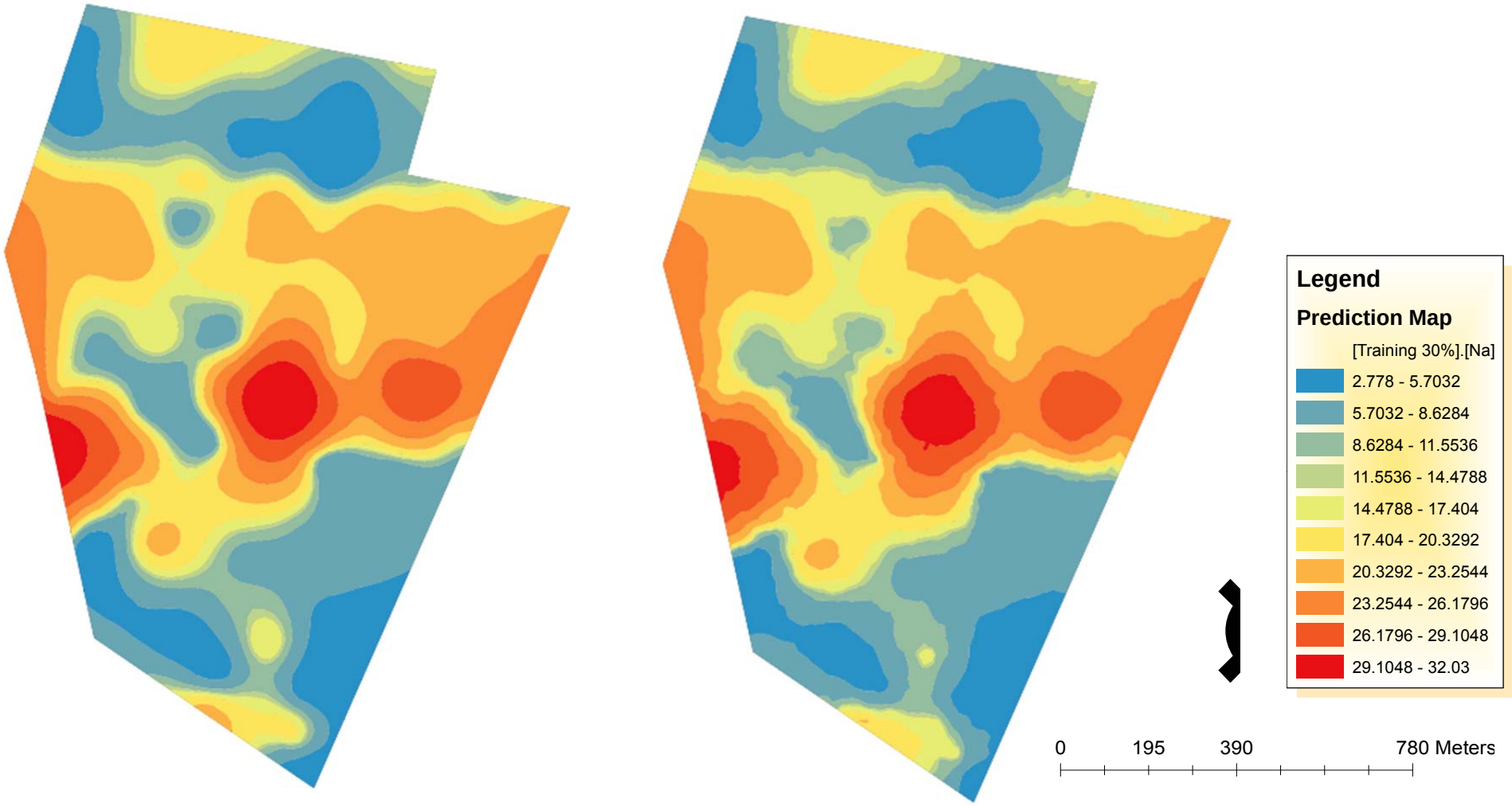


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 34: Prediction maps of randomly selected 20% training and 80% testing data sets for Sodium (Na).

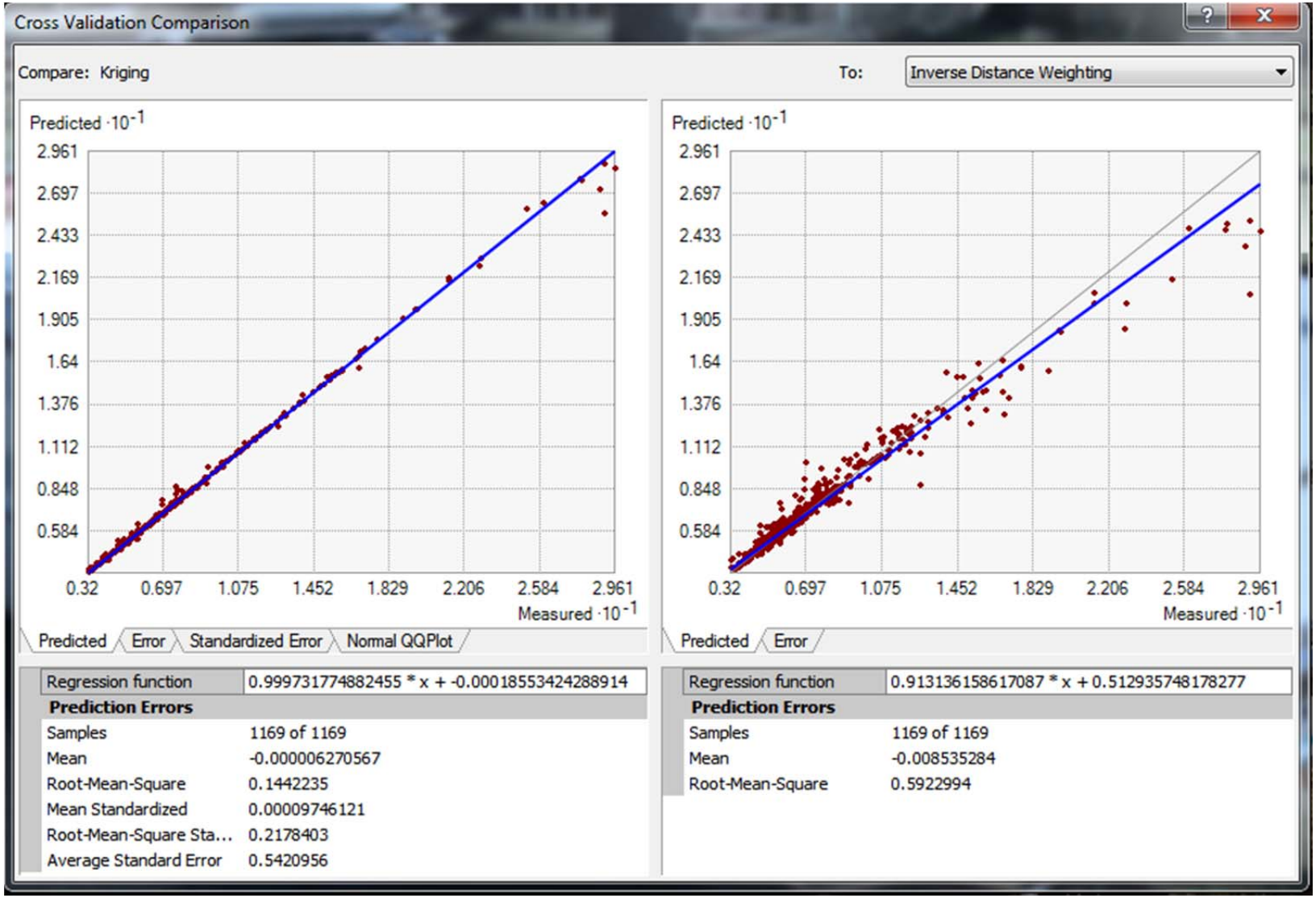


Graph 19: Cross-validation comparison of predicted error for the randomly selected 20% Na training data set.

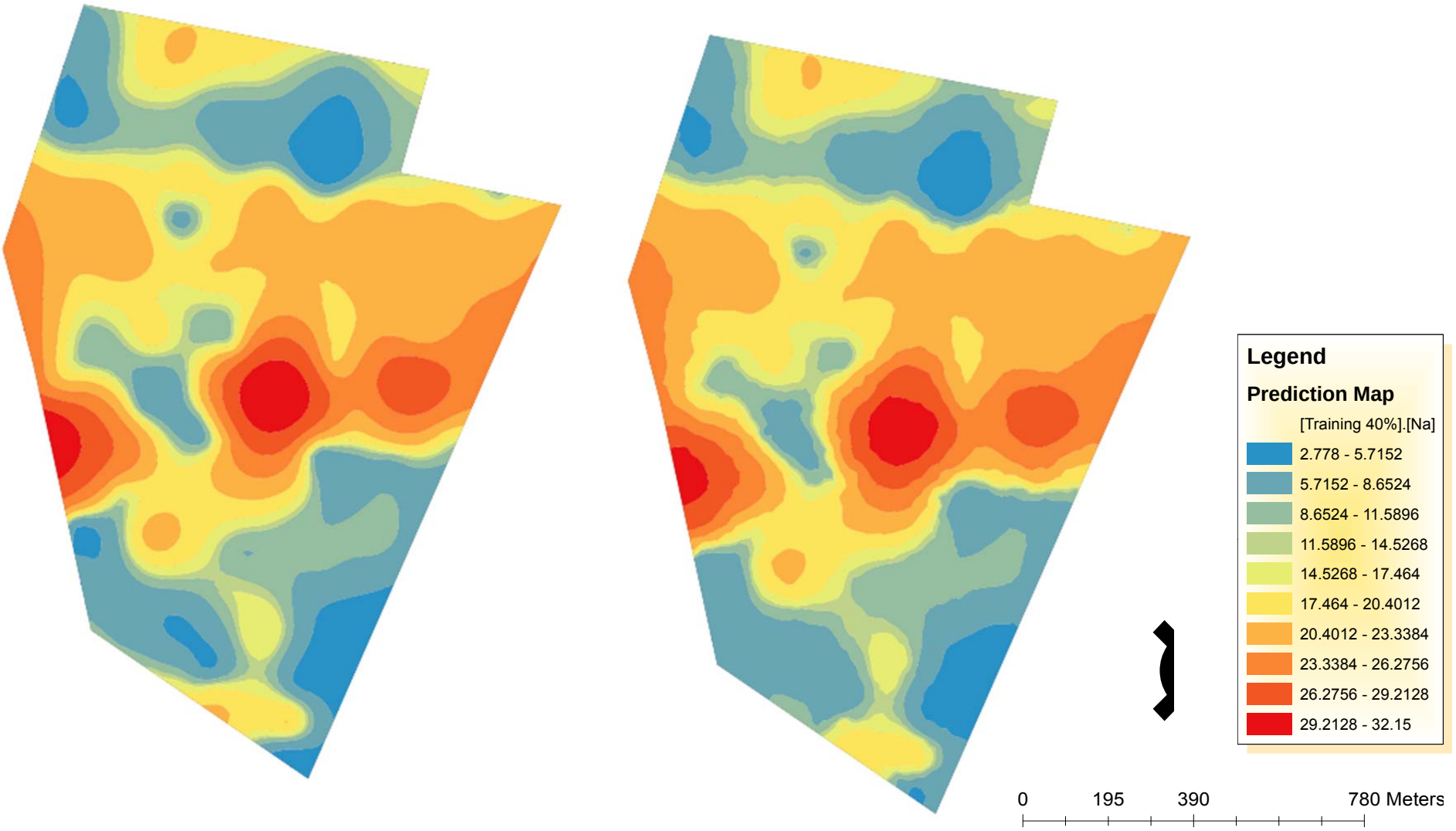


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 35: Prediction maps of randomly selected 30% training and 70% testing data sets for Sodium (Na).

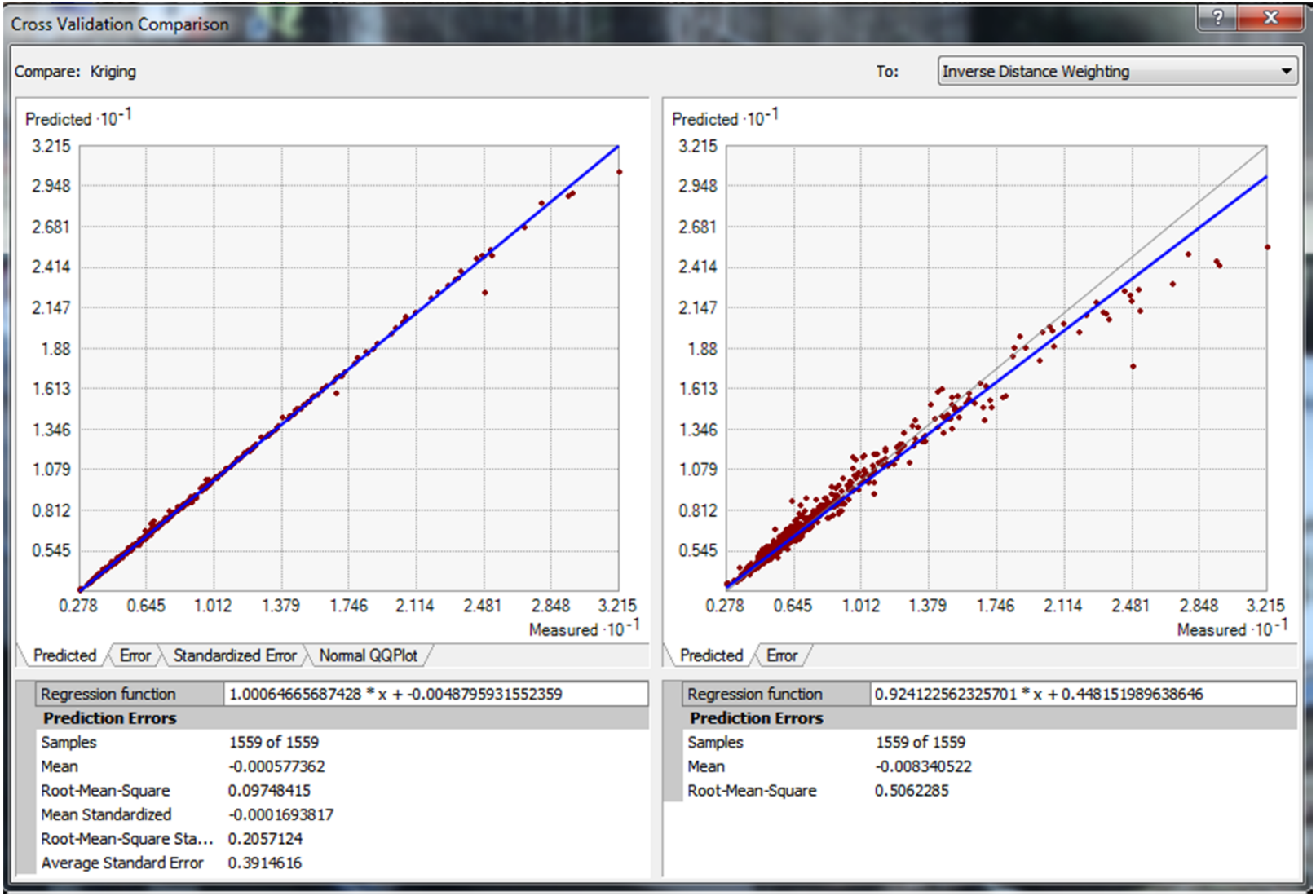


Graph 20: Cross-validation comparison of predicted error for the randomly selected 30% Na training data set.

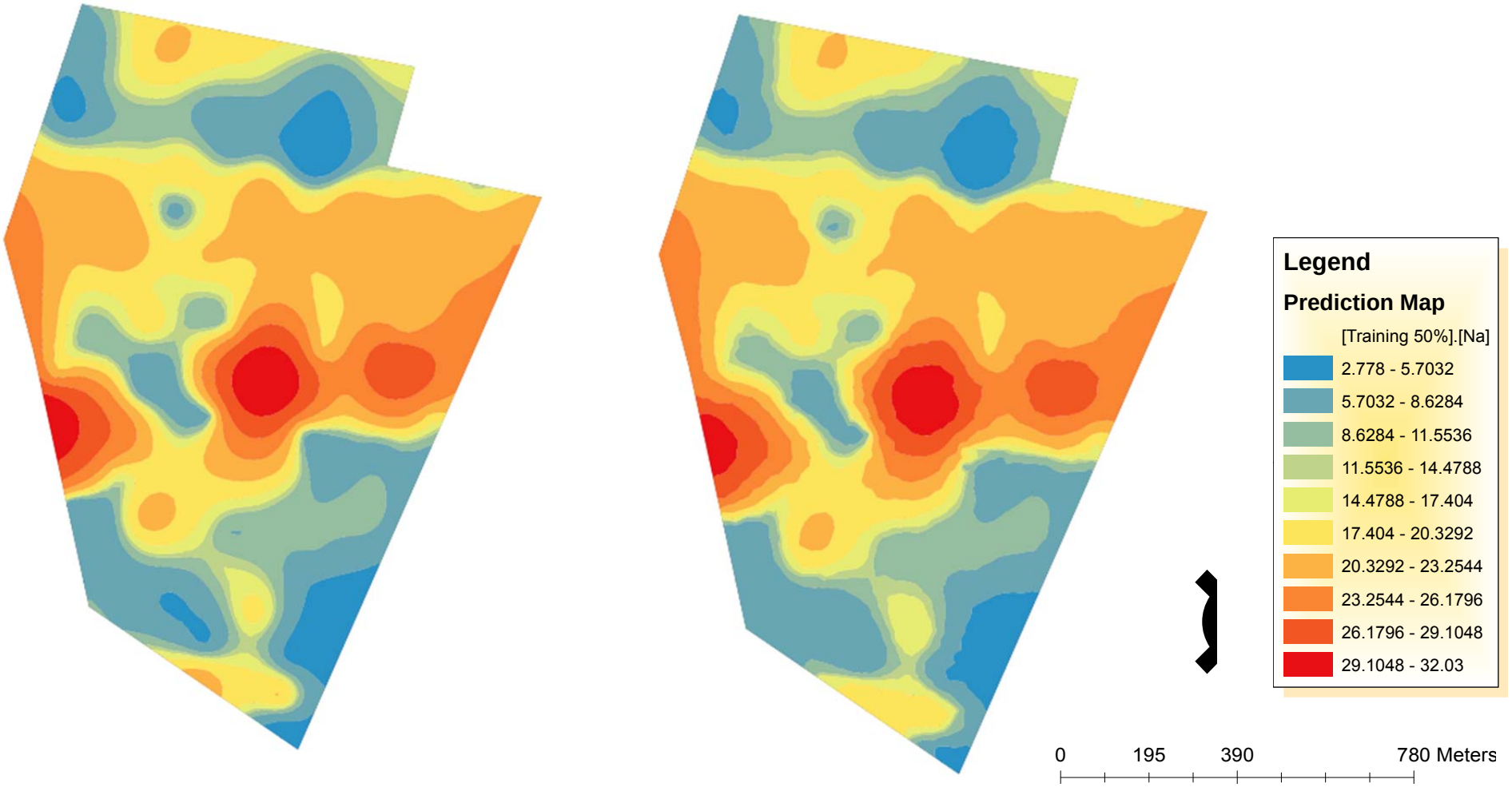


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 36: Prediction maps of randomly selected 40% training and 60% testing data sets for Sodium (Na).

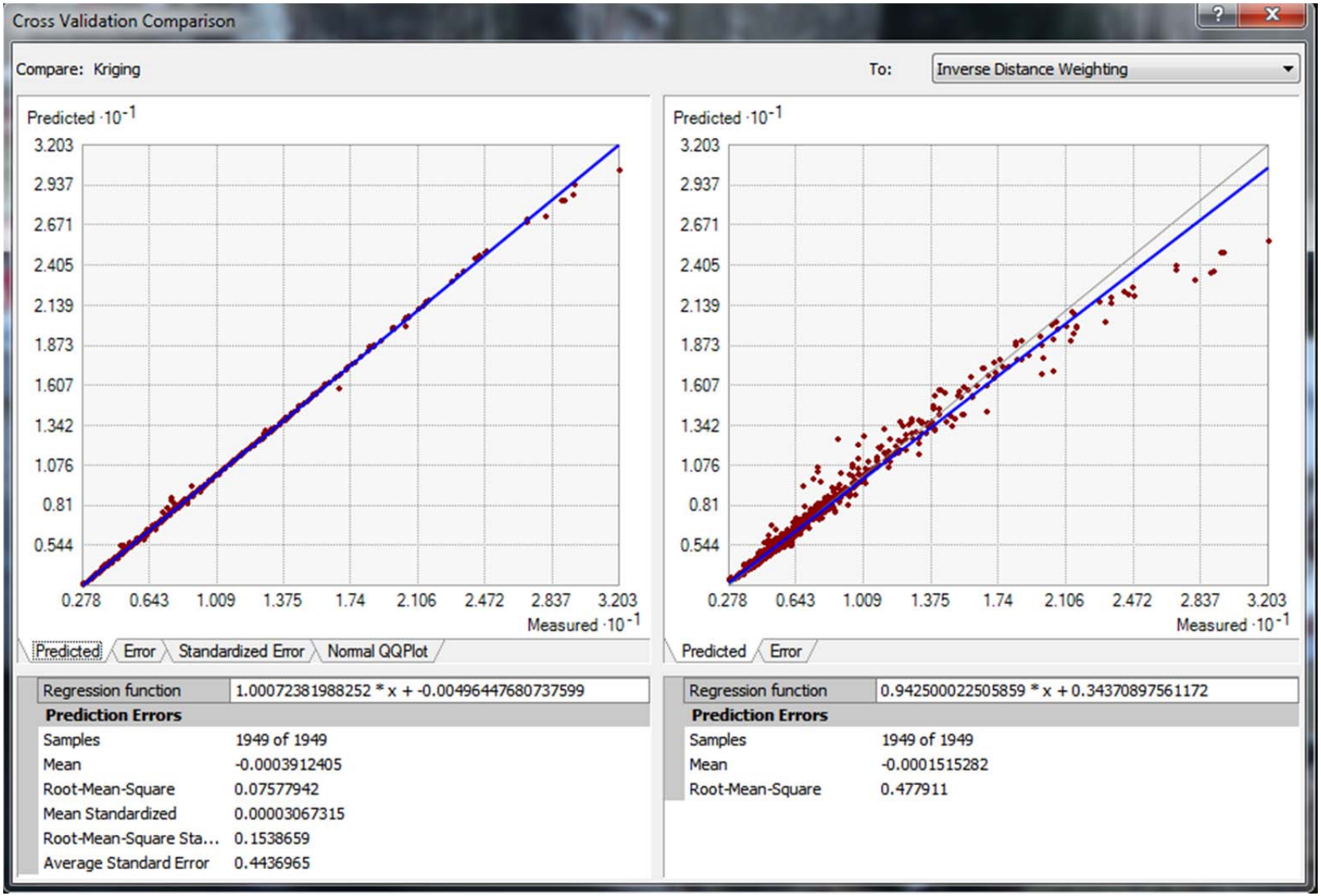


Graph 21: Cross-validation comparison of predicted error for the randomly selected 40% Na training data set.

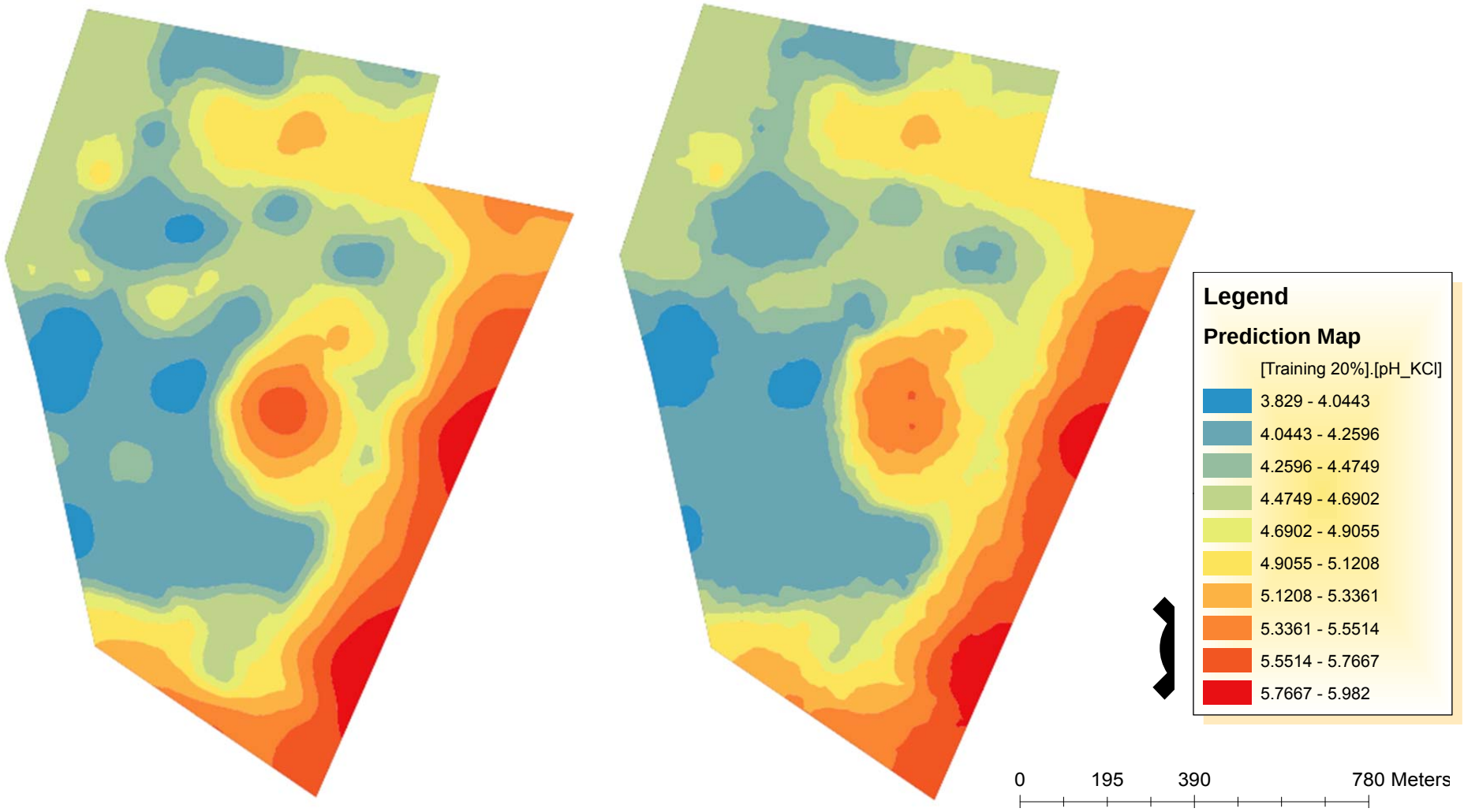


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 37: Prediction maps of randomly selected 50% training and 50% testing data sets for Sodium (Na).

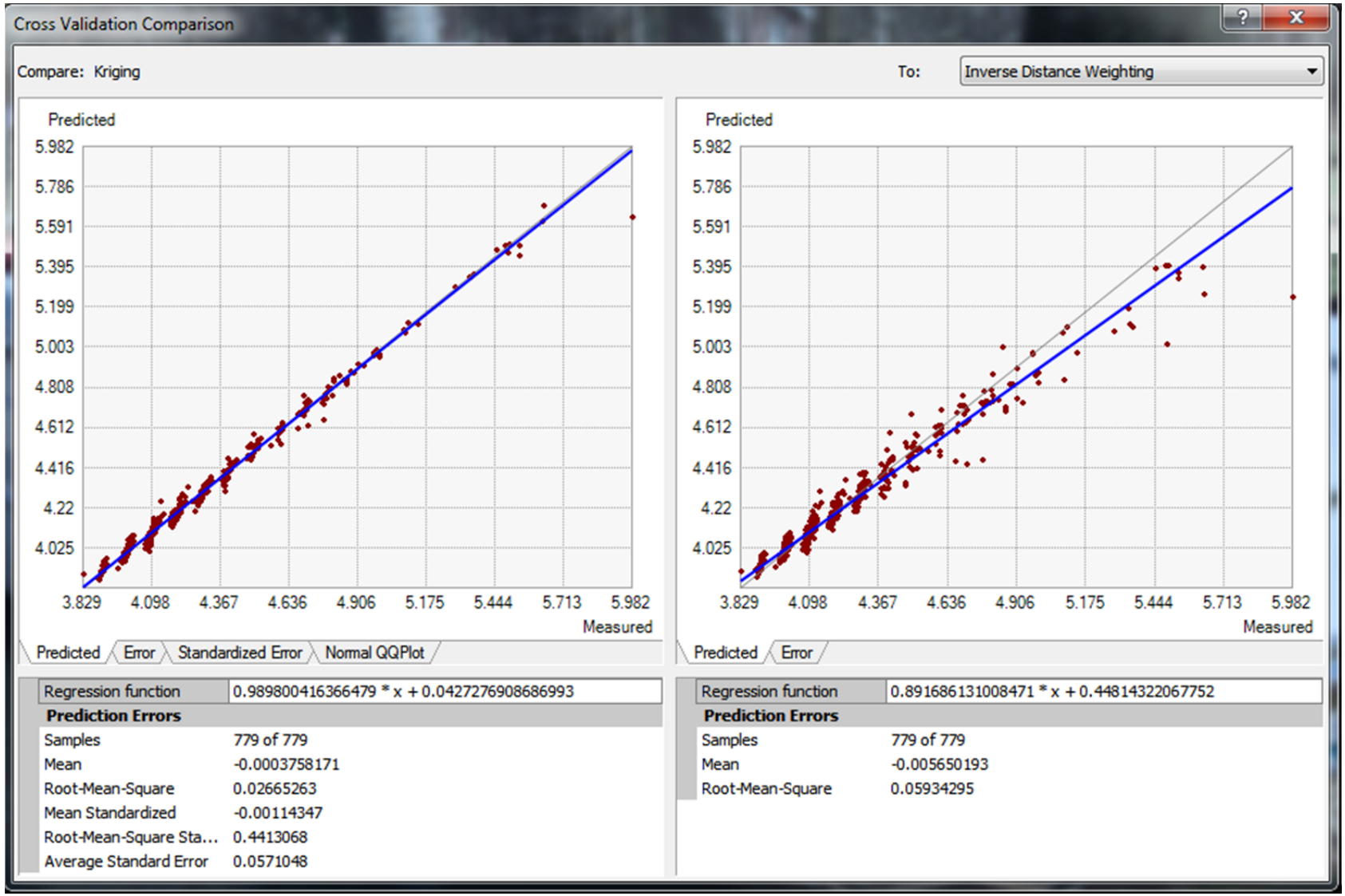


Graph 22: Cross-validation comparison of predicted error for the randomly selected 50% Na training data set.

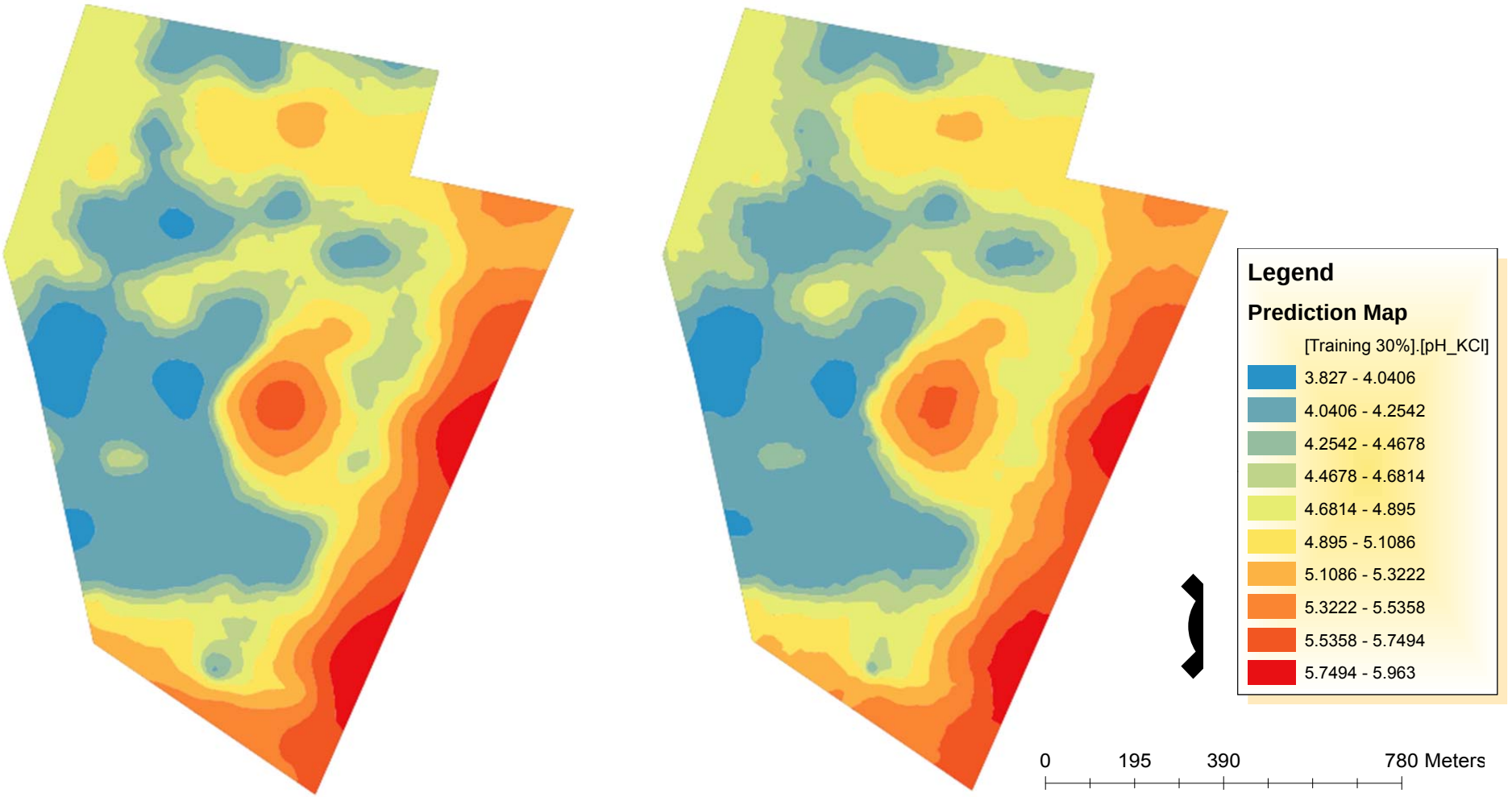


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 38: Prediction maps of randomly selected 20% training and 80% testing data sets for the pH of Potassium Chloride (pH-KCl).

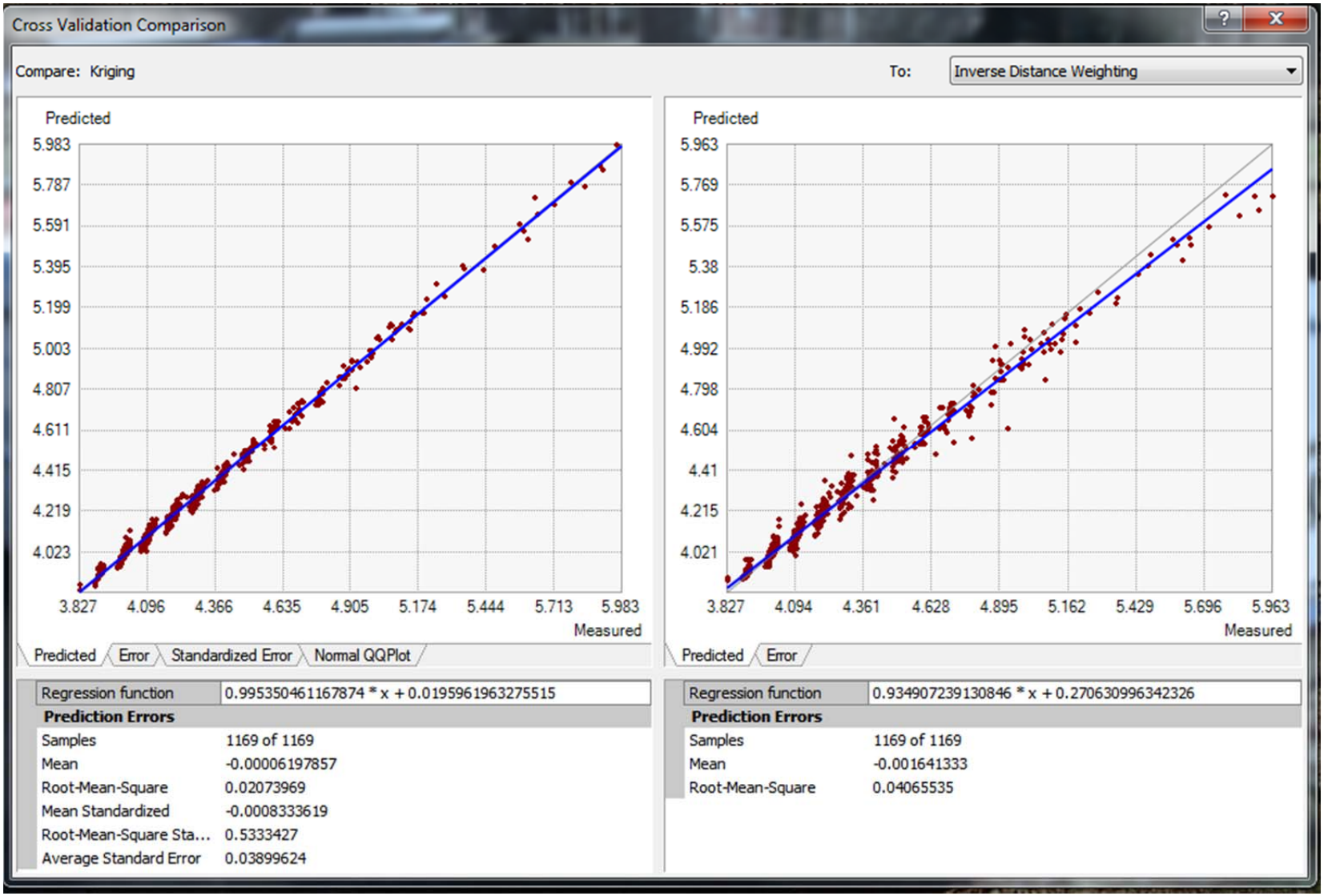


Graph 23: Cross-validation comparison of predicted error for the randomly selected 20% pH-KCl training data set.

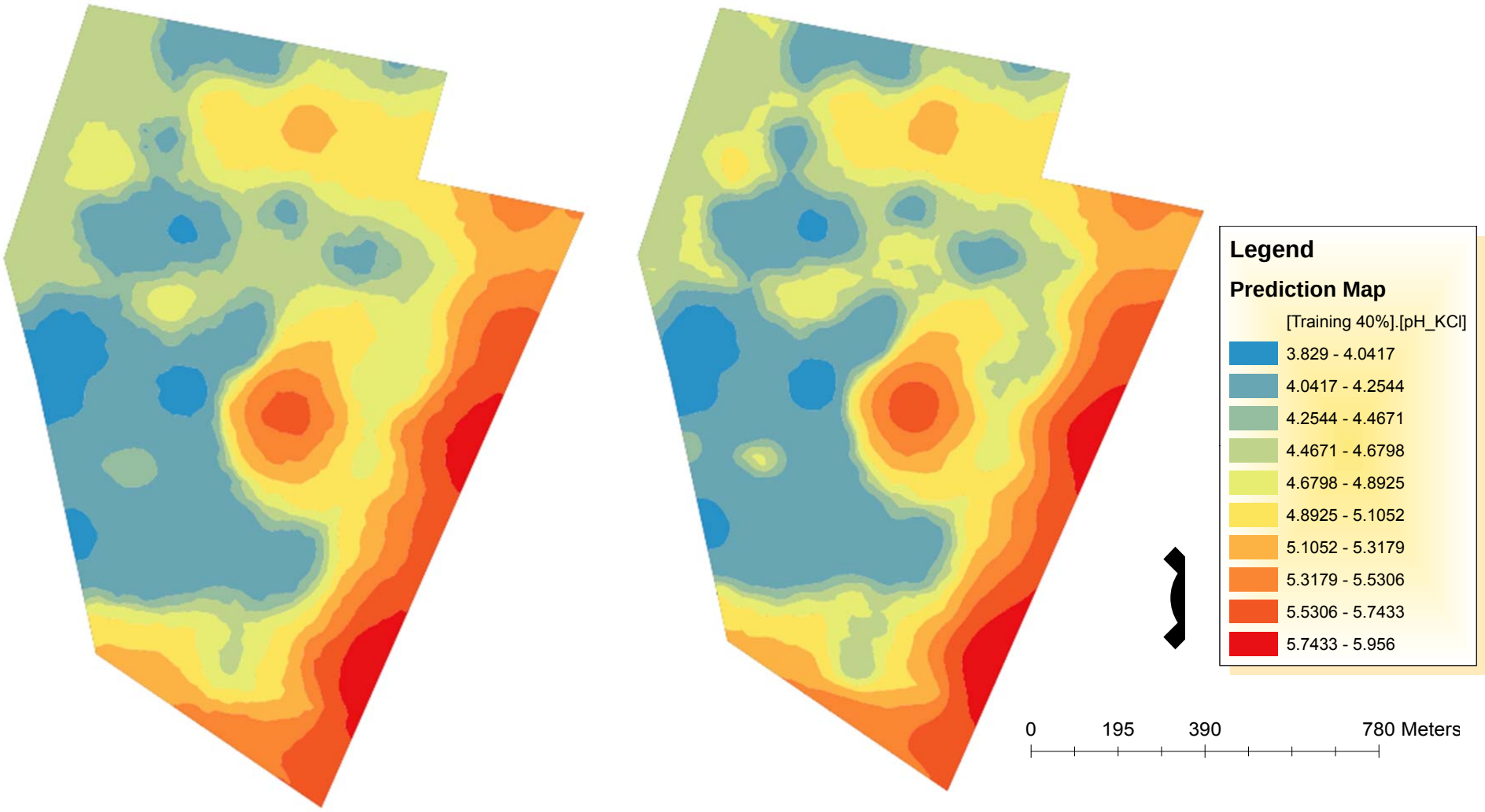


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 39: Prediction maps of randomly selected 30% training and 70% testing data sets for the pH of Potassium Chloride (pH-KCl).

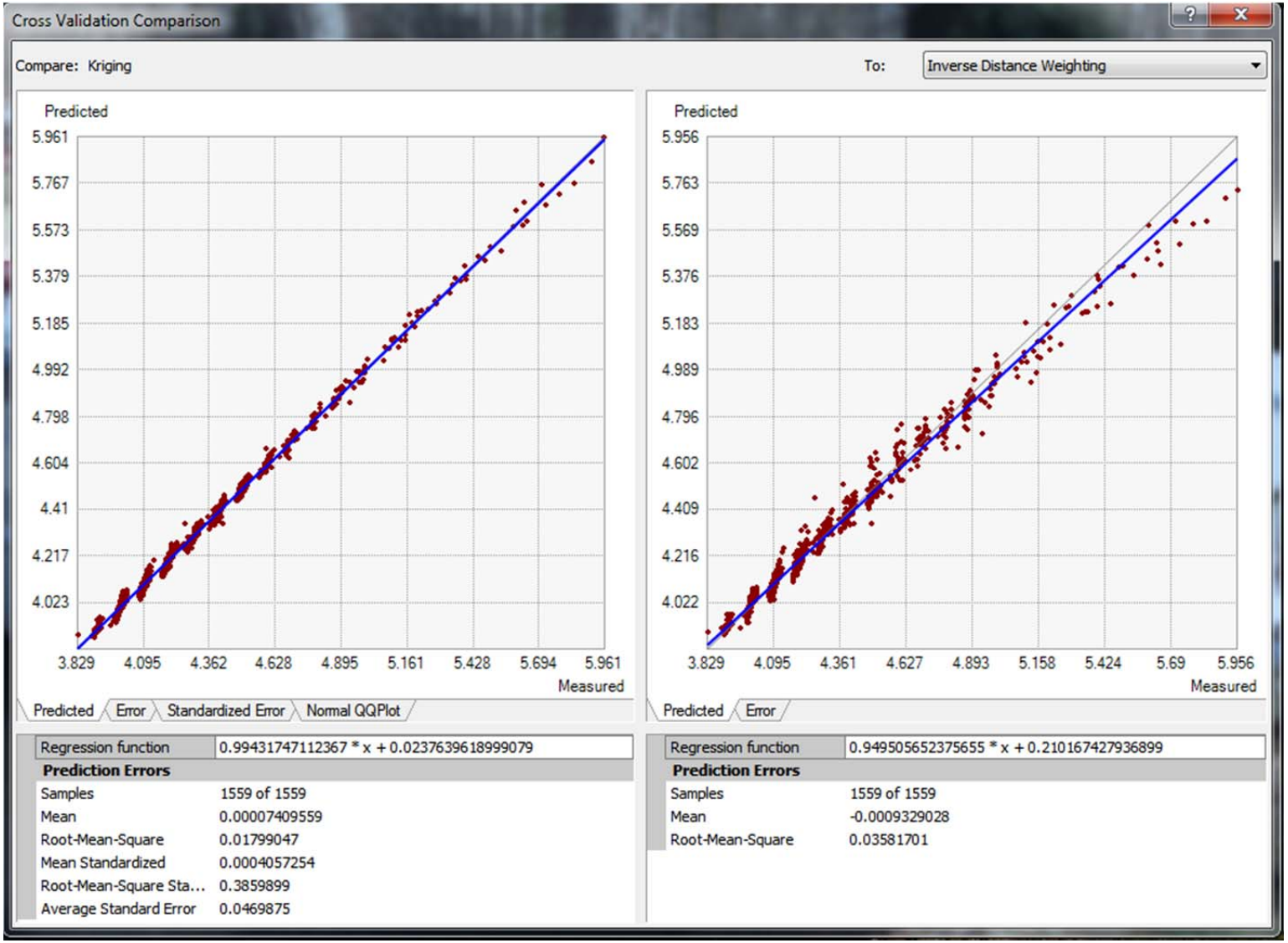


Graph 24: Cross-validation comparison of predicted error for the randomly selected 30% pH-KCl training data set.

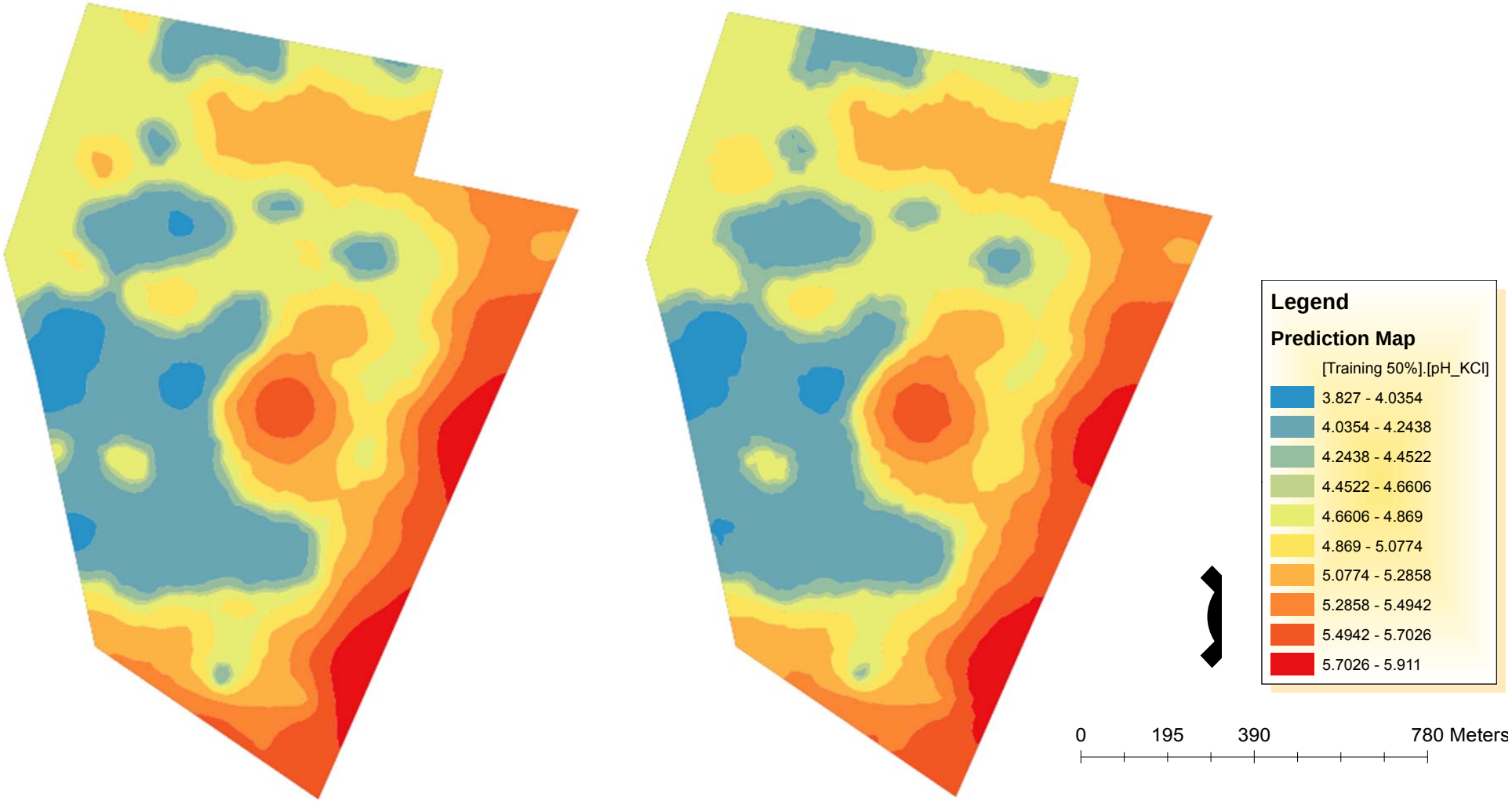


(a) Ordinary Kriging (b) Inverse Distance Weighting

Figure 40: Prediction maps of randomly selected 40% training and 60% testing data sets for the pH of Potassium Chloride (pH-KCl).



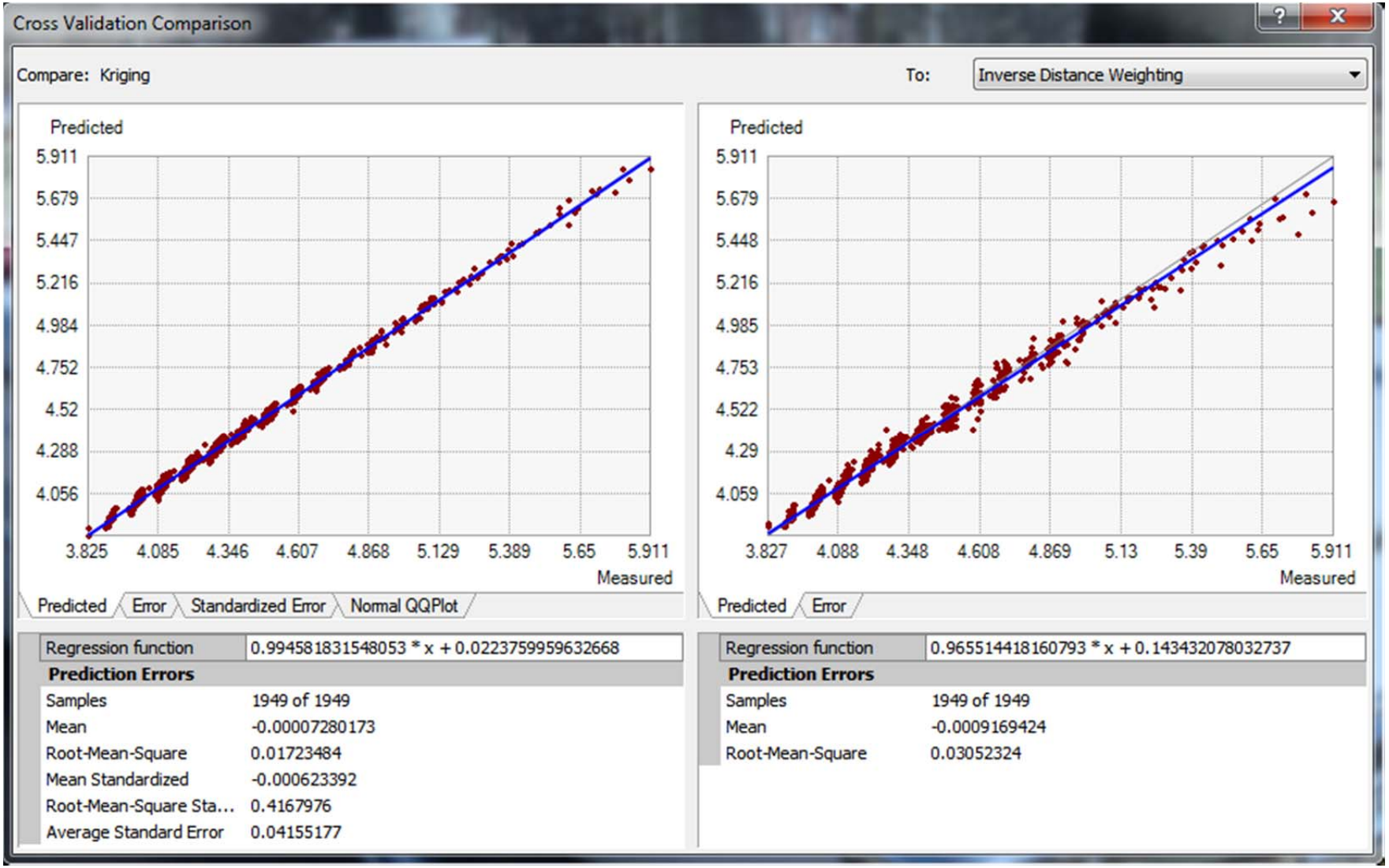
Graph 25: Cross-validation comparison of predicted error for the randomly selected 40% pH-KCl training data set.



(a) Ordinary Kriging

(b) Inverse Distance Weighting

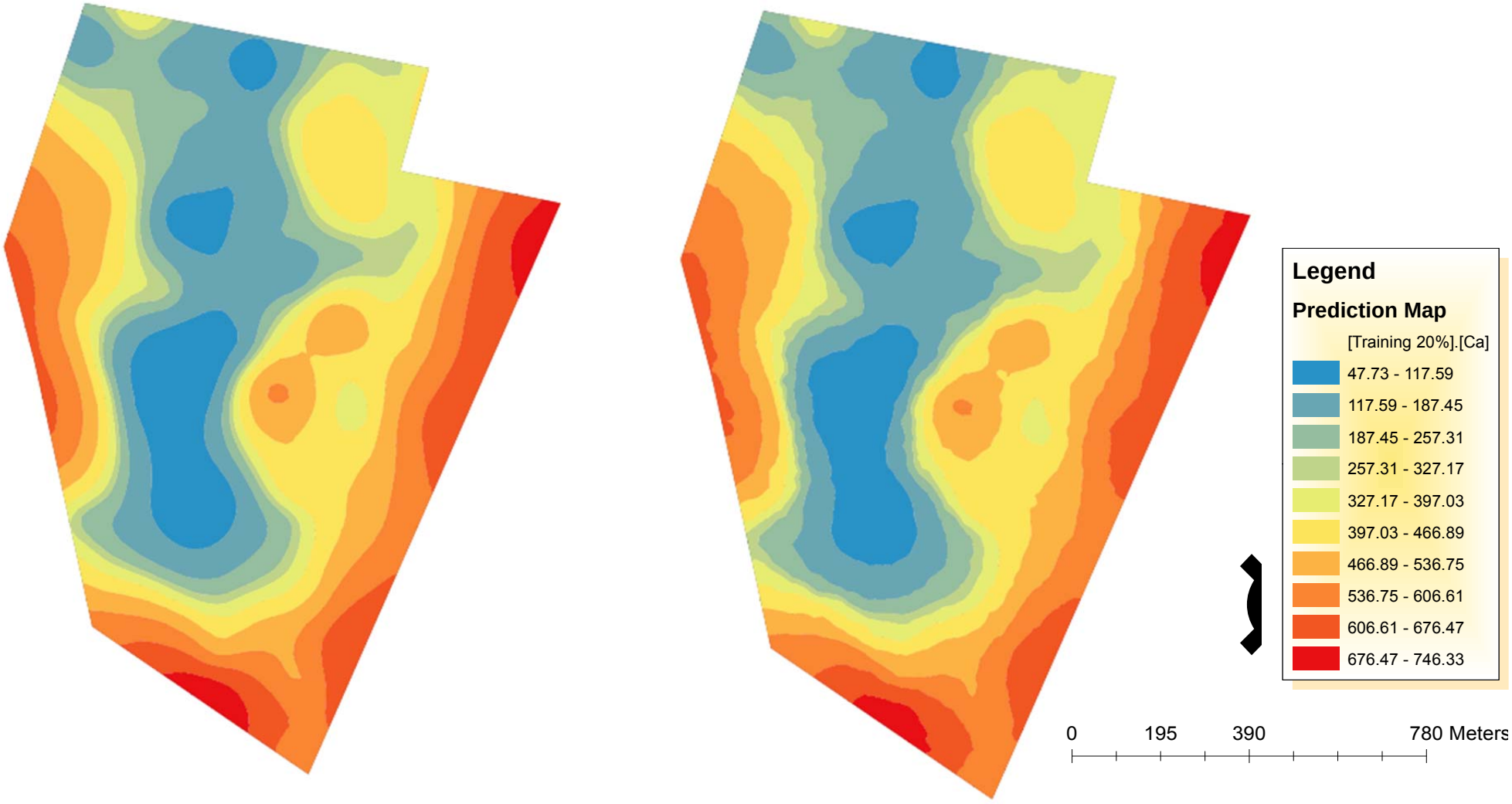
Figure 41: Prediction maps of randomly selected 50% training and 50% testing data sets for the pH of Potassium Chloride (pH-KCl).



Graph 26: Cross-validation comparison of predicted error for the randomly selected 50% pH-KCl training data set.

Appendix B

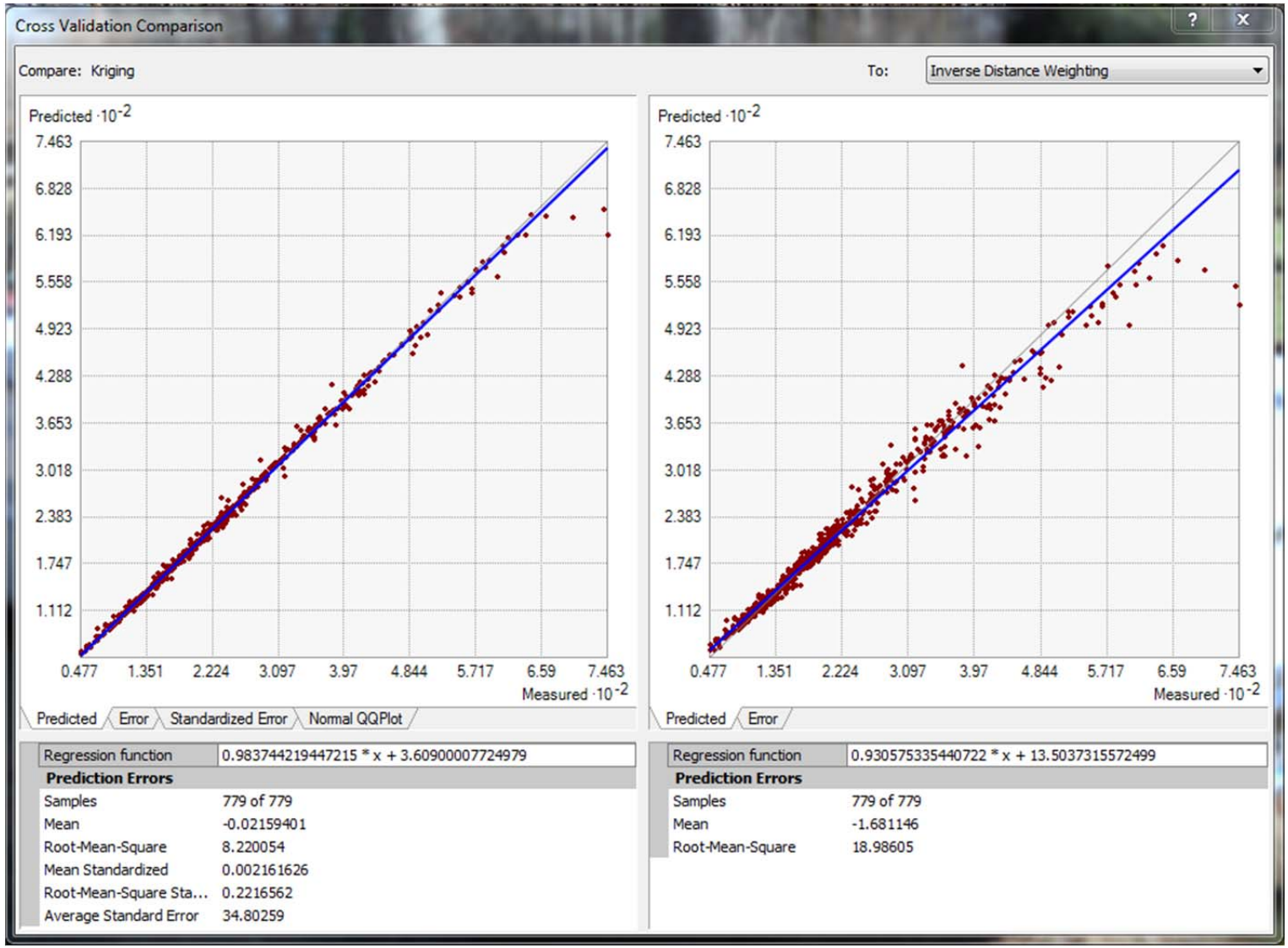
Appendix B contains: Prediction maps of systematically selected training and testing data sets for the soil chemical properties (Ca, K, Mg, Na and pH) as well as a cross-validation scatter plot graphs comparison of ordinary kriging and IDW predicted error for the systematic data set.



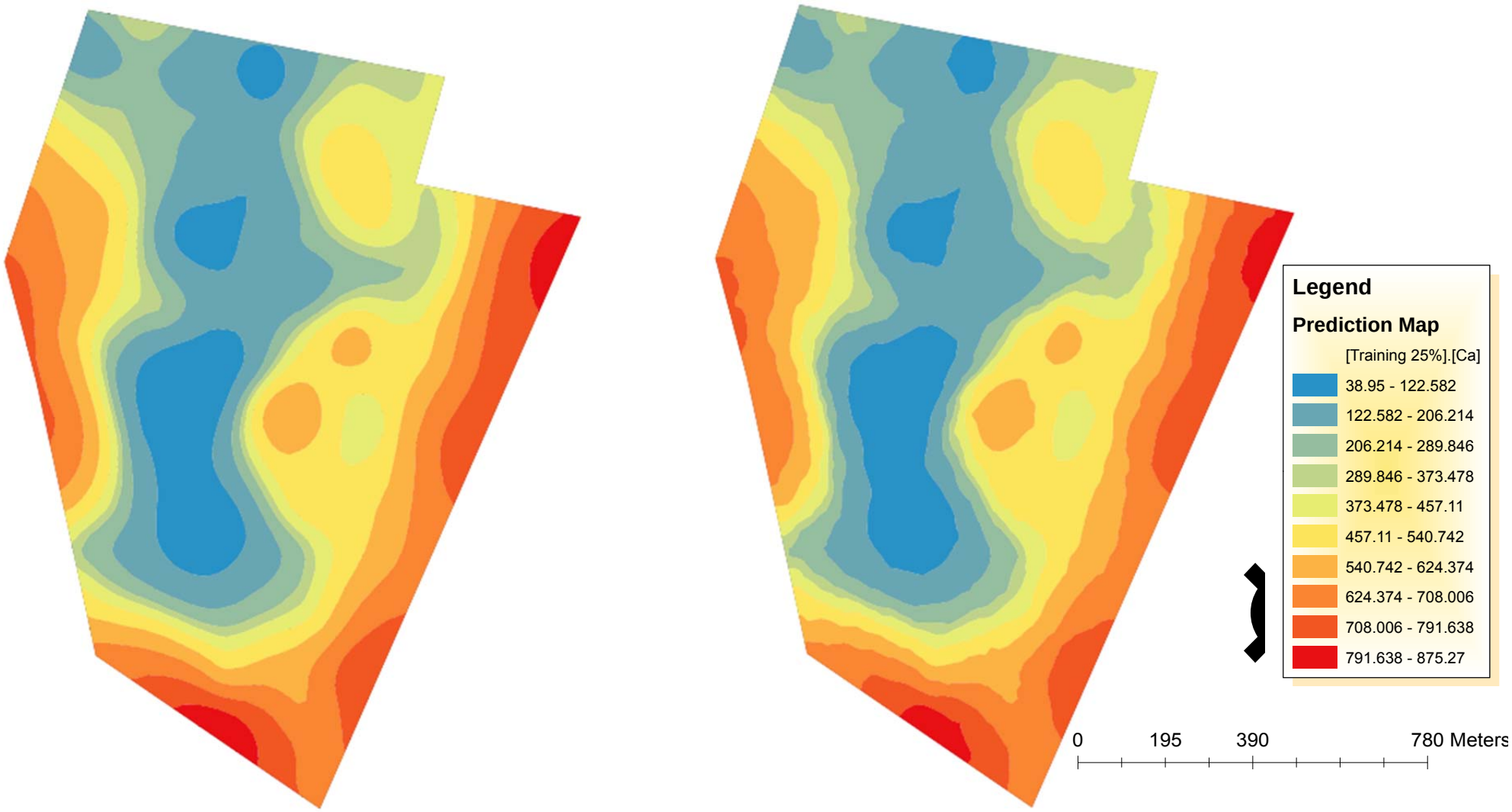
(b) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 42: Prediction maps of systematically selected 20% training and 80% testing (removal of every 5th data point) data sets for Calcium (Ca).



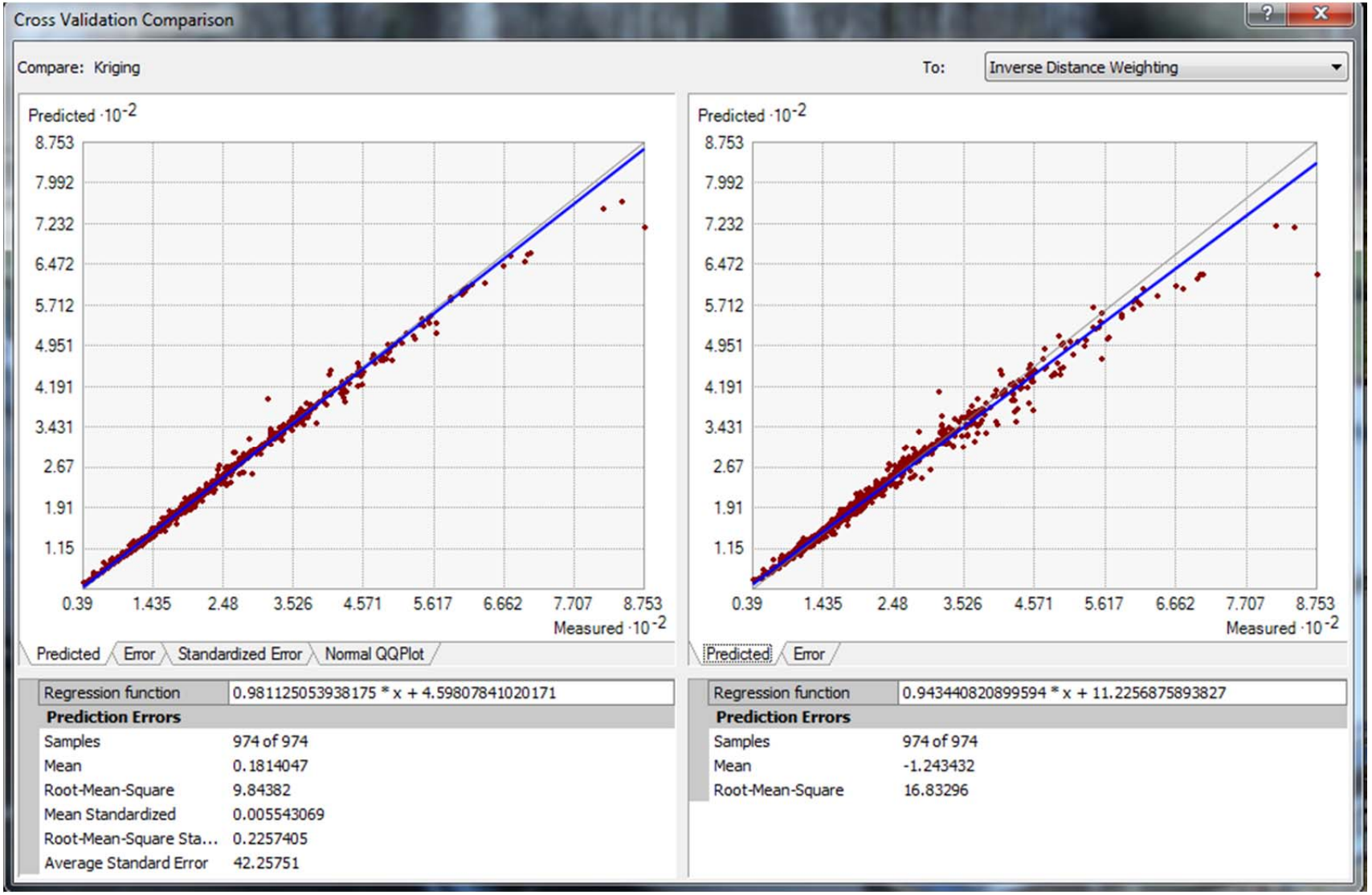
Graph 27: Cross-validation comparison of predicted error for the systematically selected 20% Ca training data set.



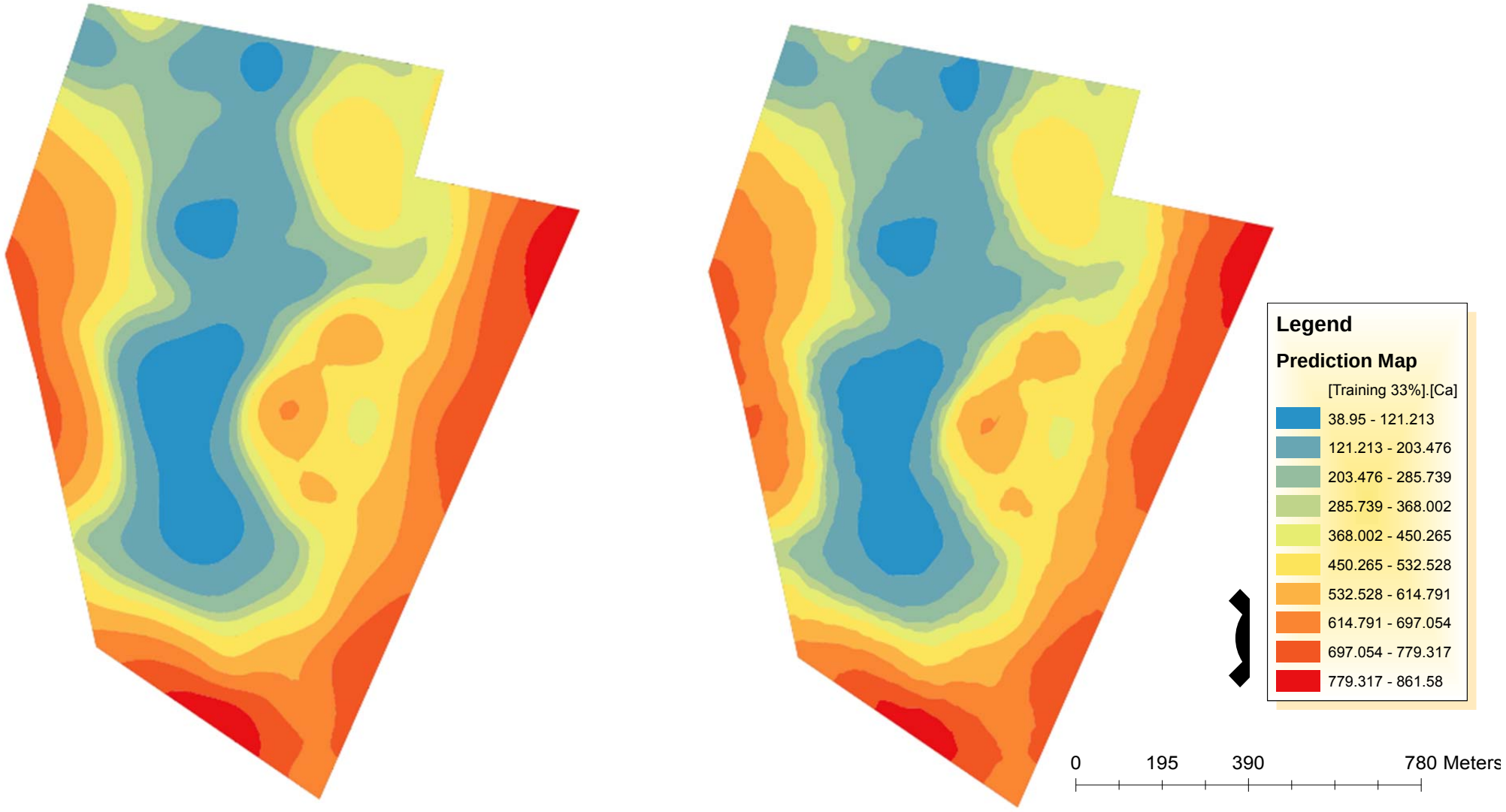
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 43: Prediction maps of systematically selected 25% training and 75% testing (removal of every 4th data point) data sets for Calcium (Ca).



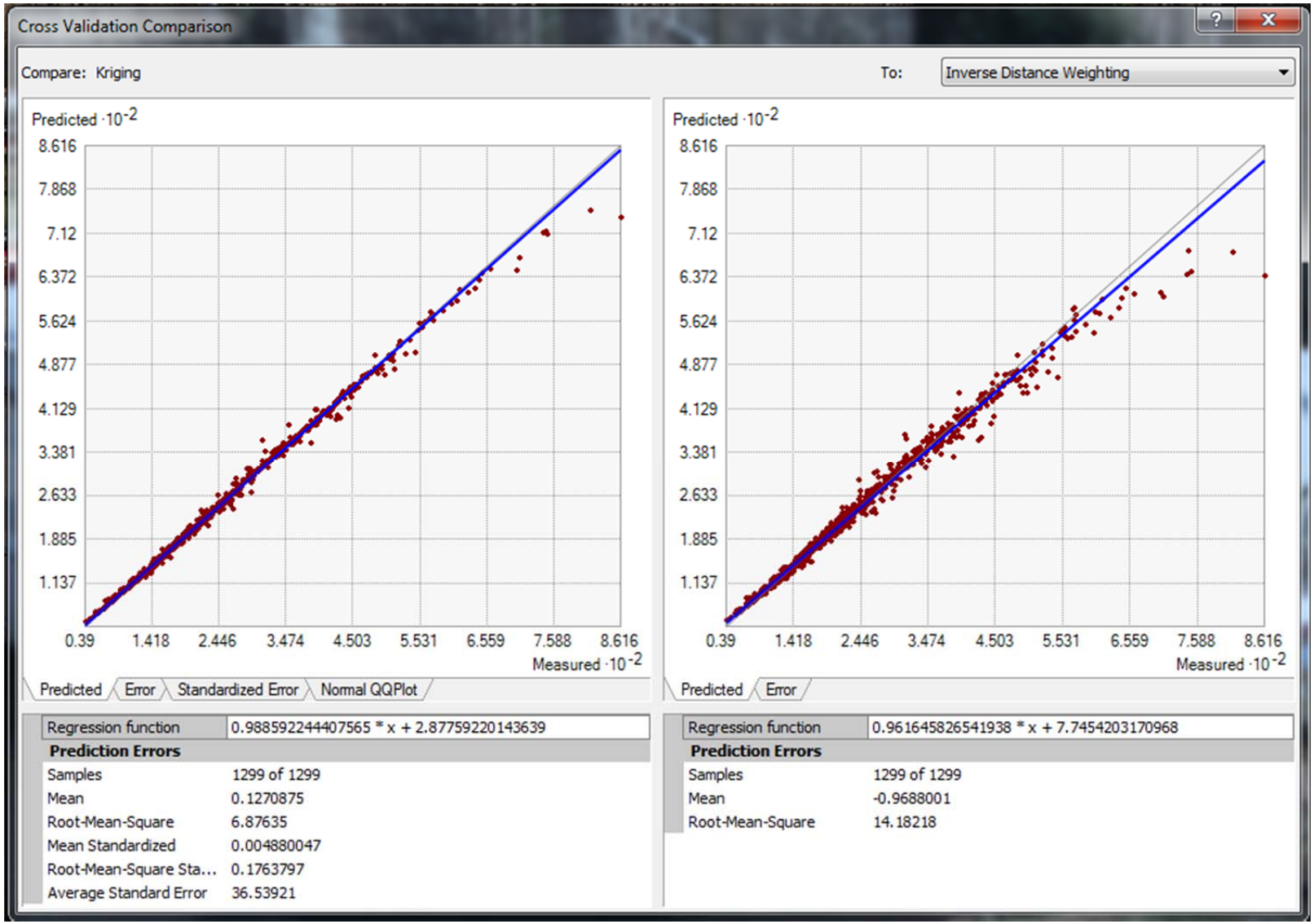
Graph 28: Cross-validation comparison of predicted error for the systematically selected 25% Ca training data set.



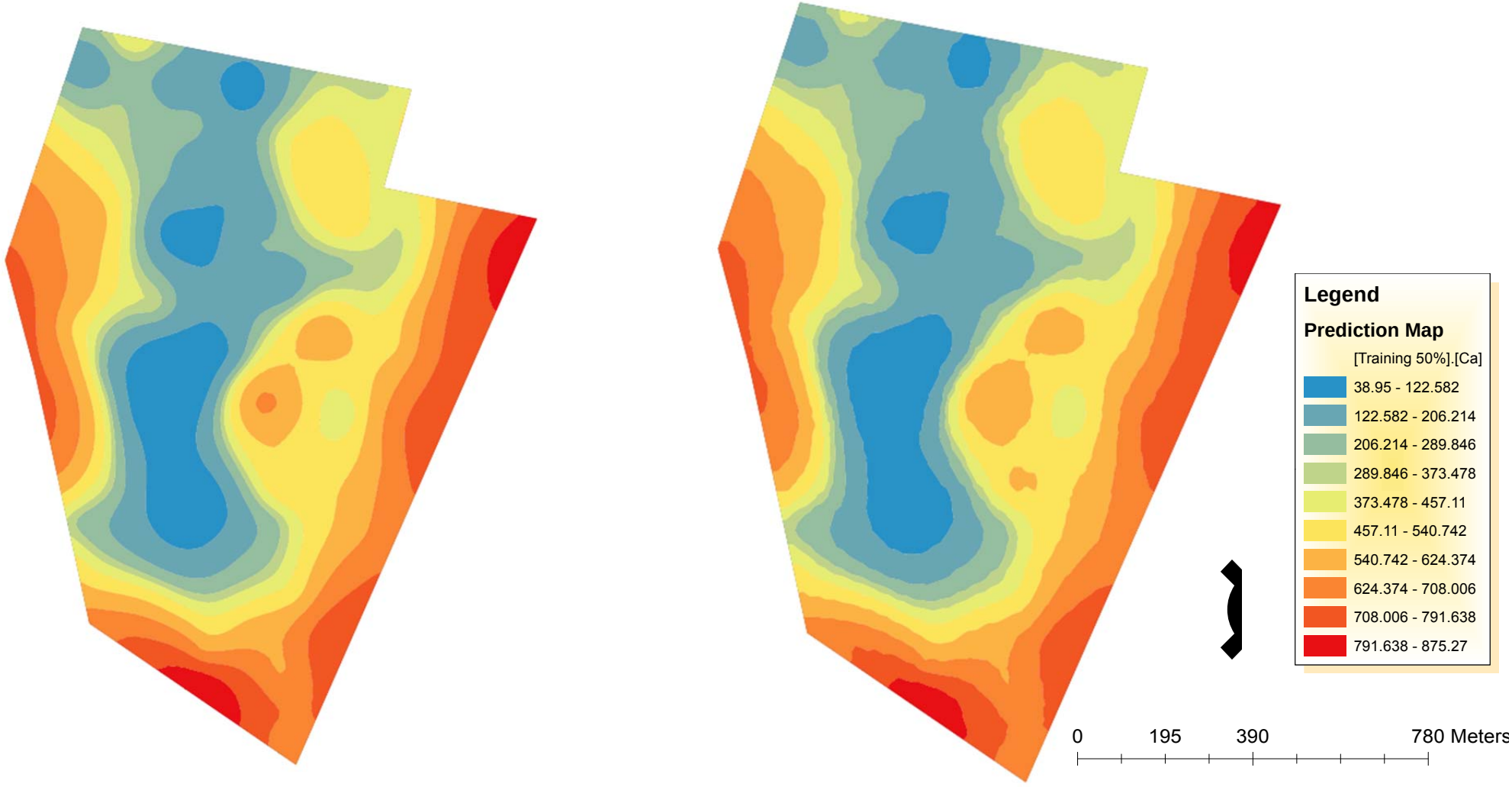
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 44: Prediction maps of systematically selected 33% training and 67% testing (removal of every 3rd data point) data sets for Calcium (Ca).



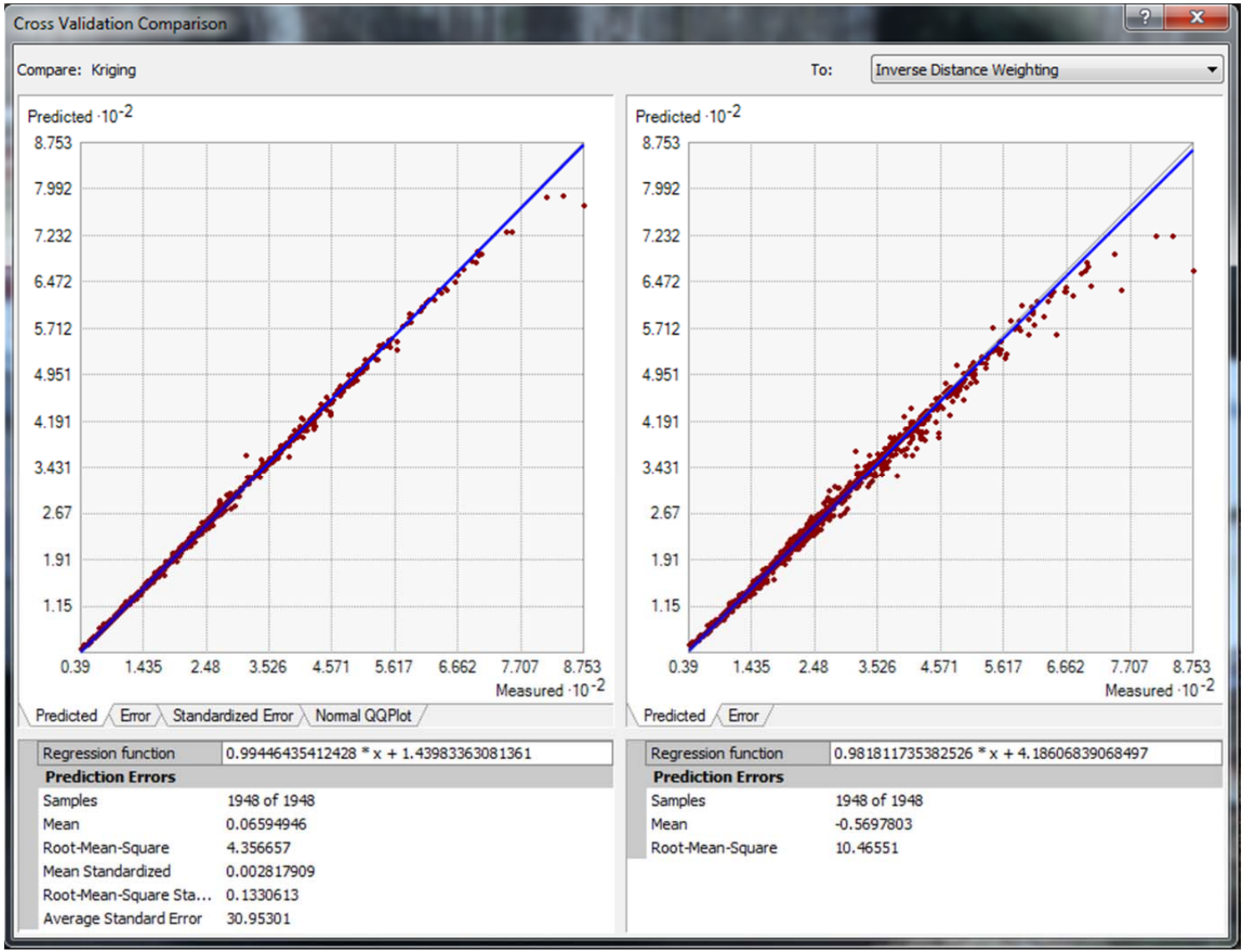
Graph 29: Cross-validation comparison of predicted error for the systematically selected 33% Ca training data set



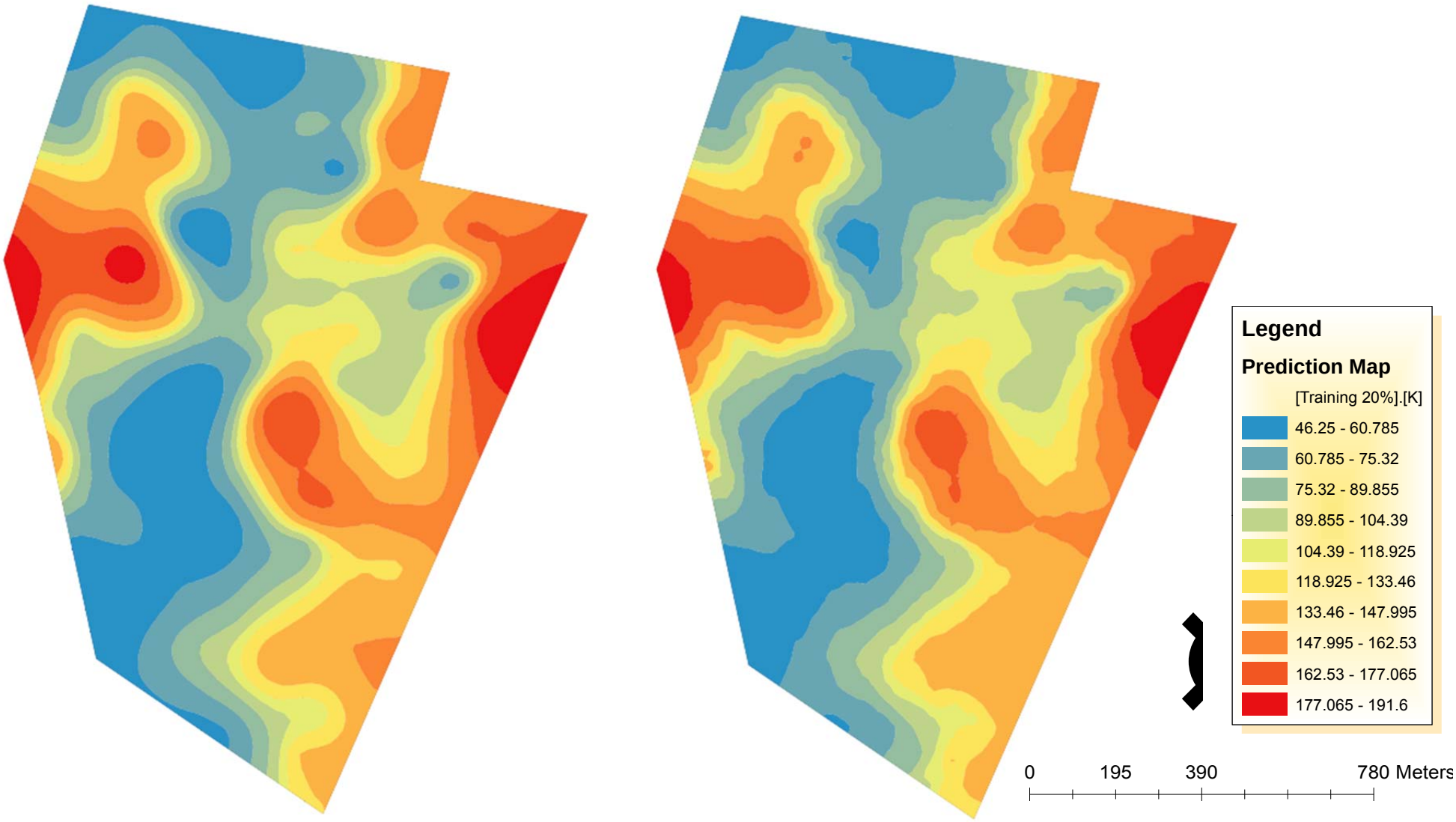
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 45: Prediction maps of systematically selected 50% training and 50% testing (removal of every 2nd data point) data sets for Calcium (Ca).



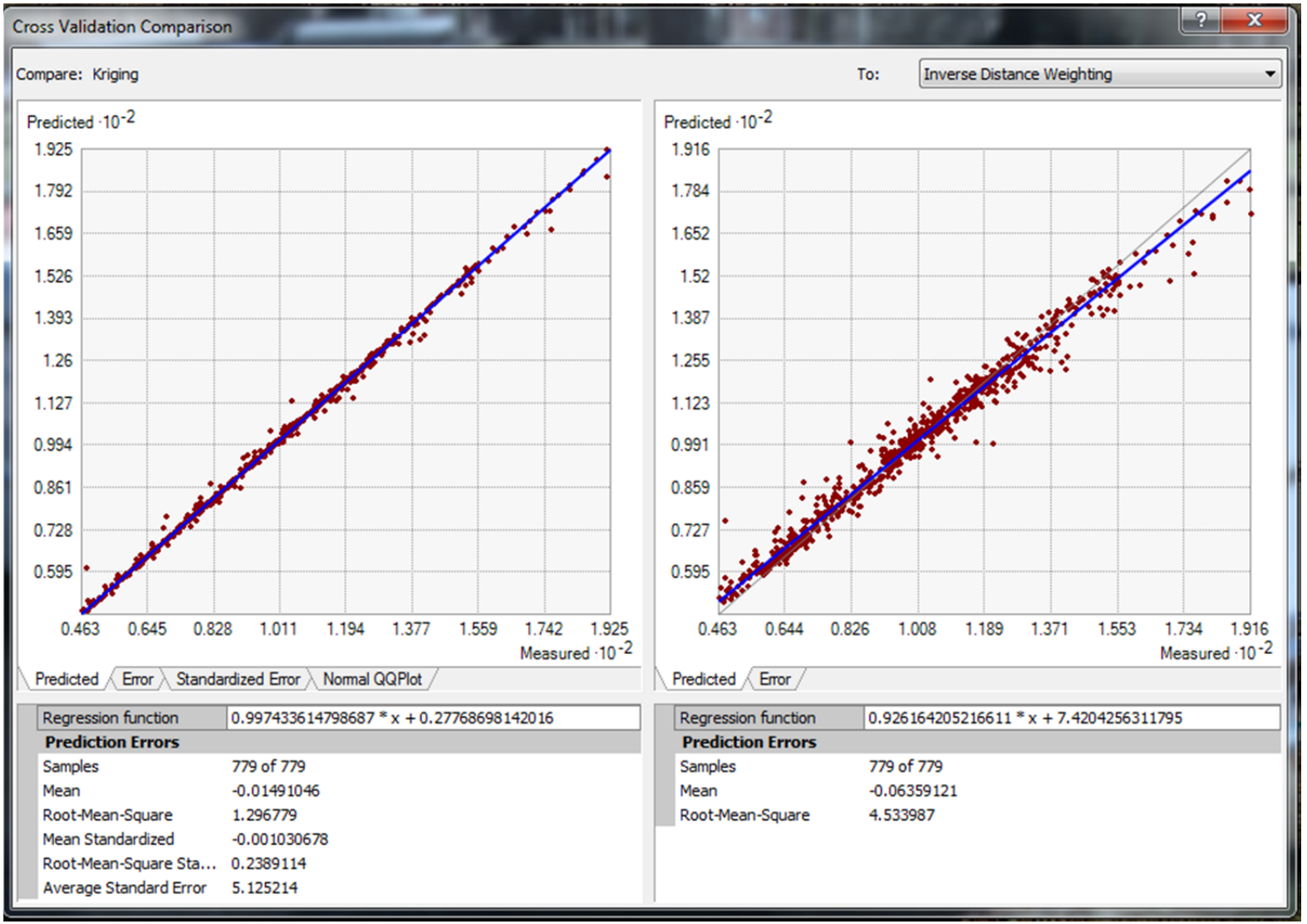
Graph 30: Cross-validation comparison of predicted error for the systematically selected 50% Ca training data set.



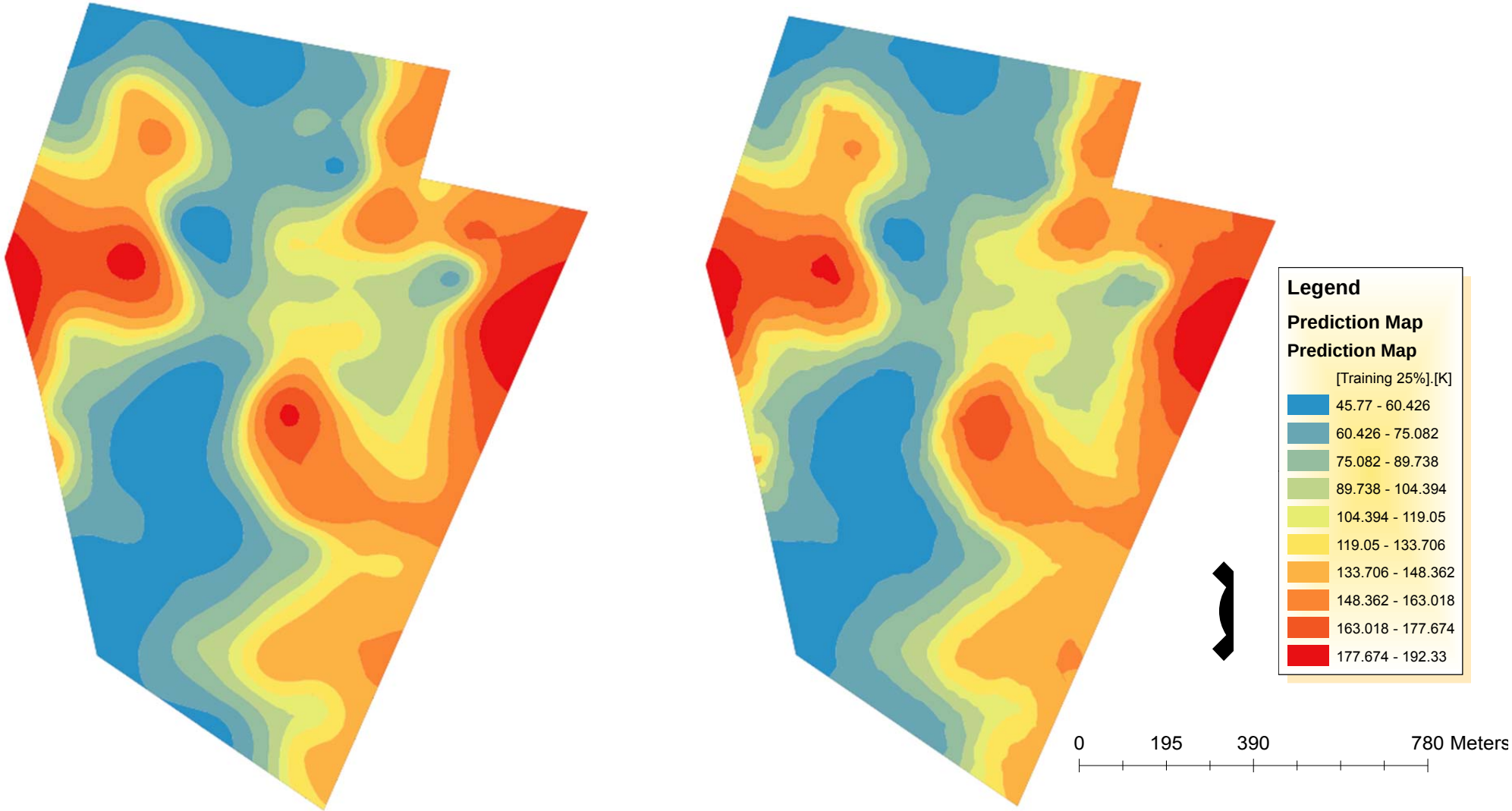
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 46: Prediction maps of systematically selected 20% training and 80% testing (removal of every 5th data point) data sets for Potassium (K).



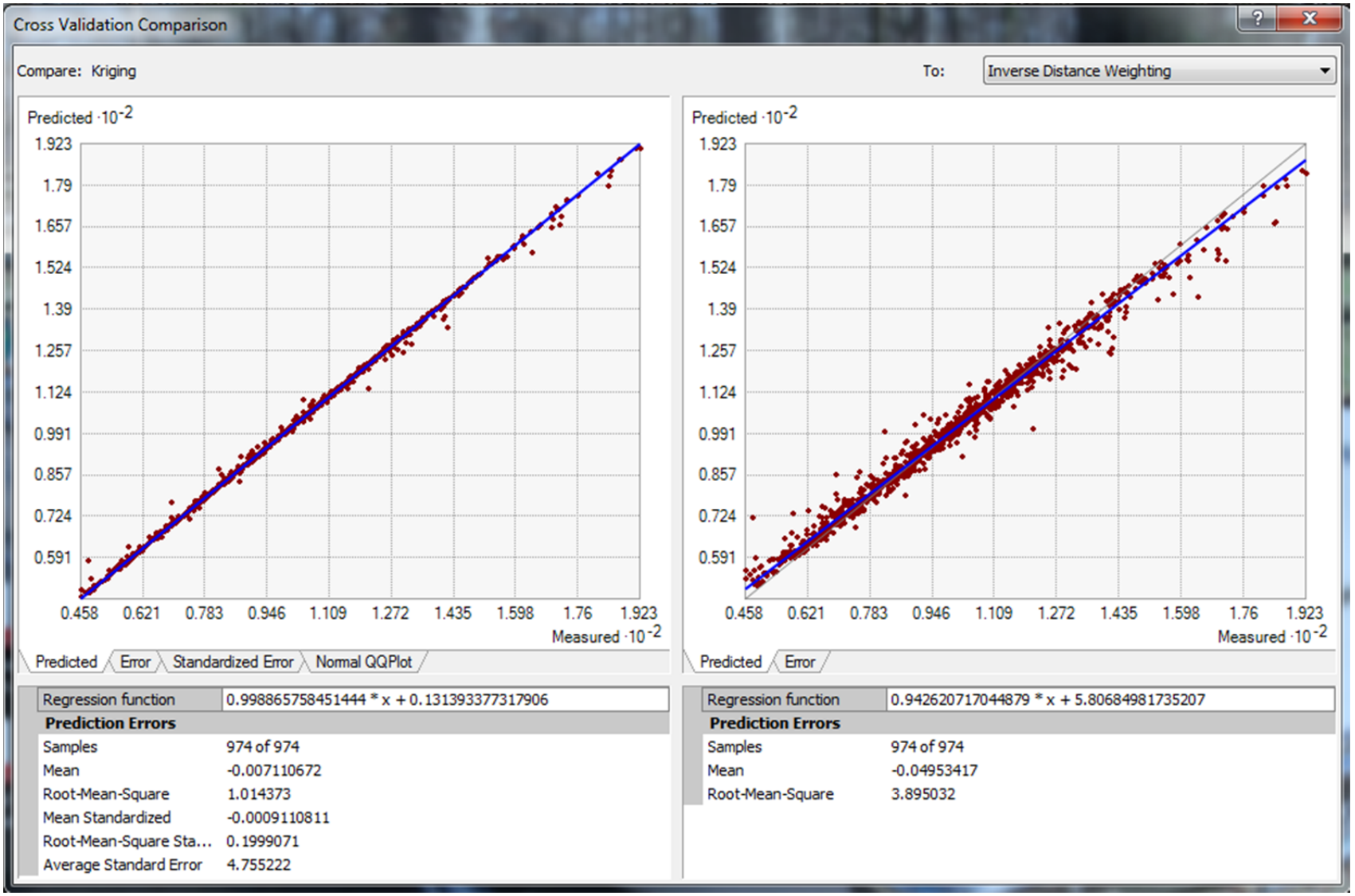
Graph 31: Cross-validation comparison of predicted error for the systematically selected 20% K training data set.



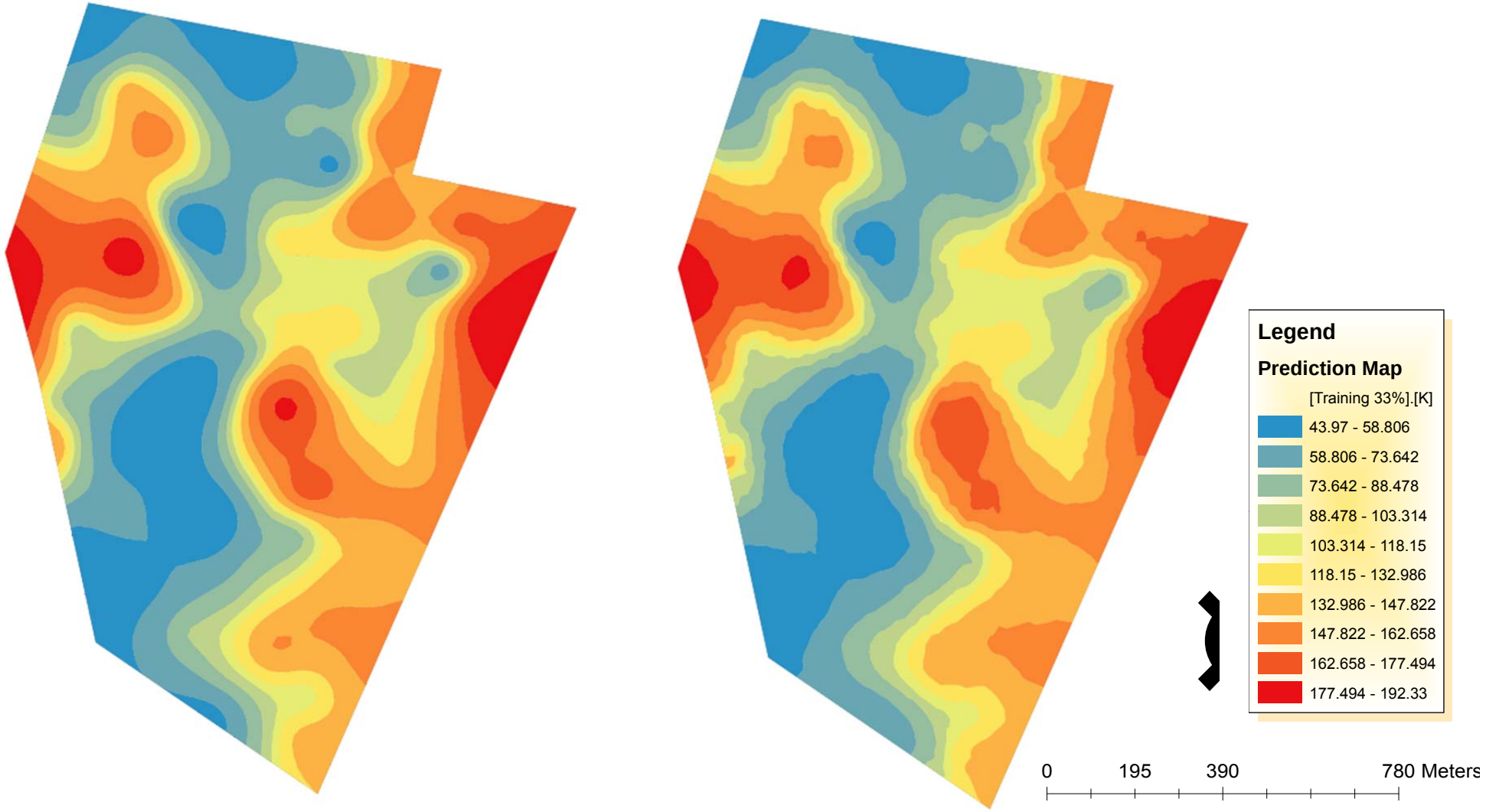
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 47: Prediction maps of systematically selected 25% training and 75% testing (removal of every 4th data point) data sets for Potassium (K).



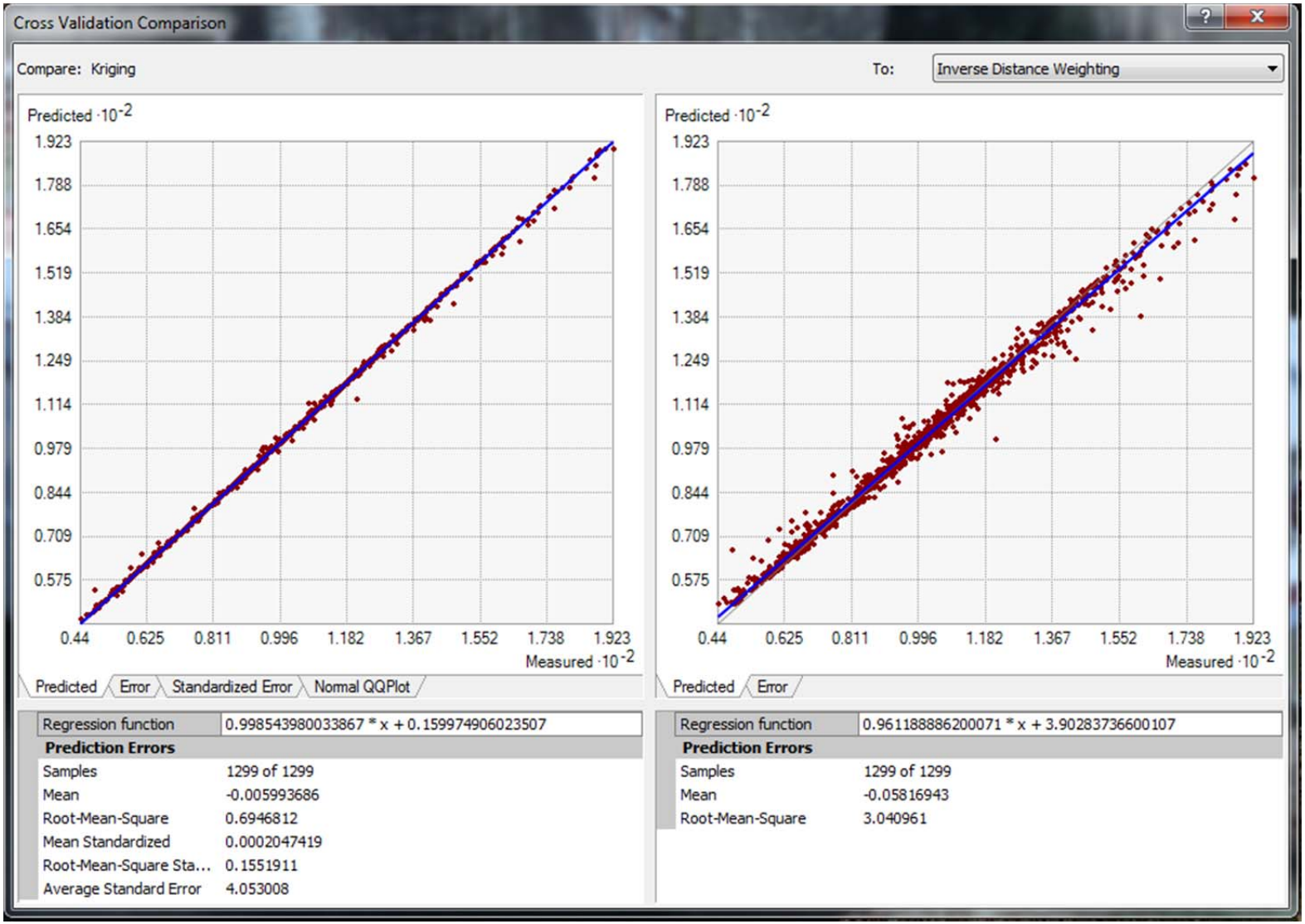
Graph 32: Cross-validation comparison of predicted error for the systematically selected 25% K training data set.



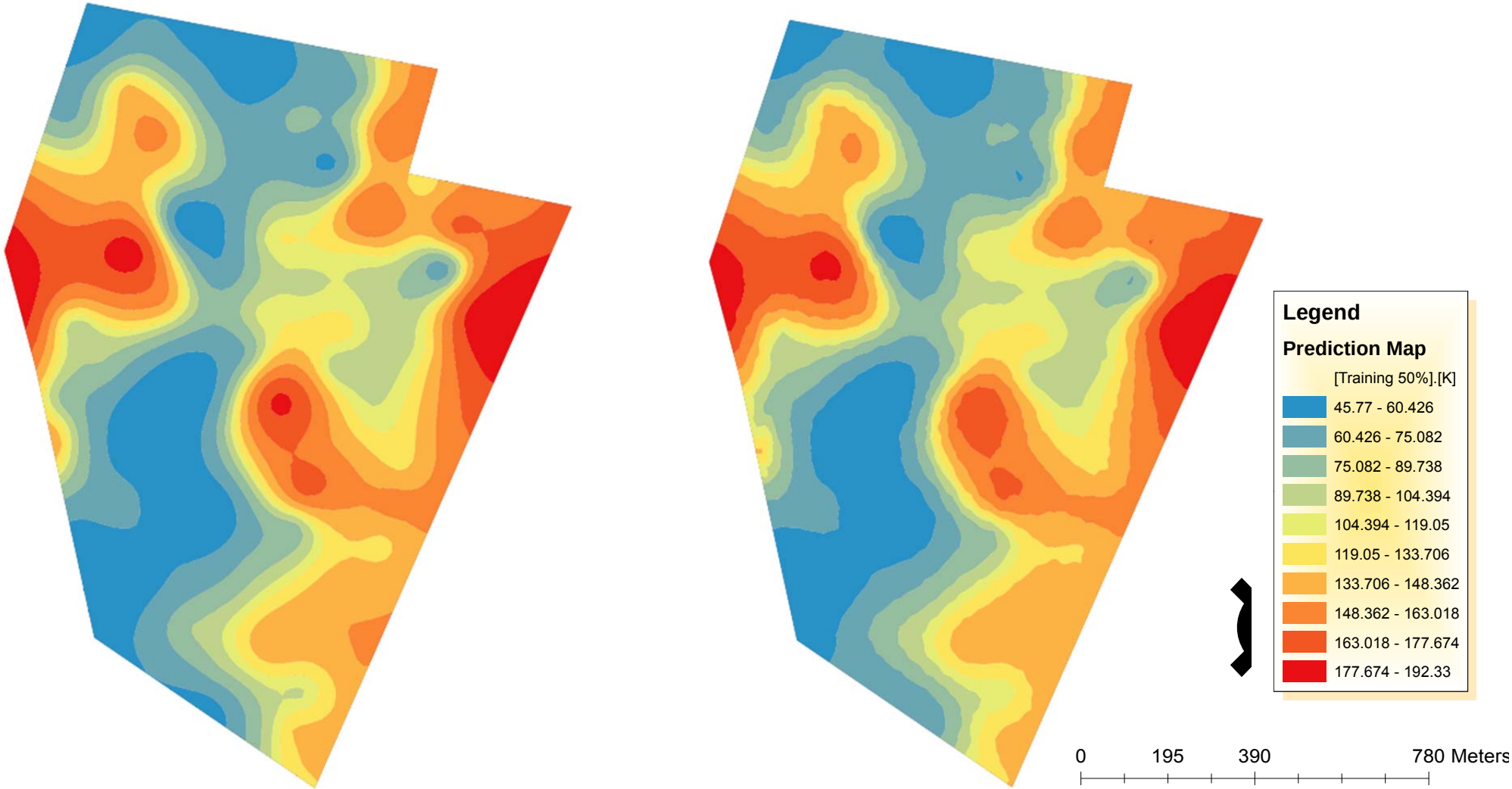
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 48: Prediction maps of systematically selected 33% training and 67 % testing (removal of every 3rd data point) data sets for Potassium (K).



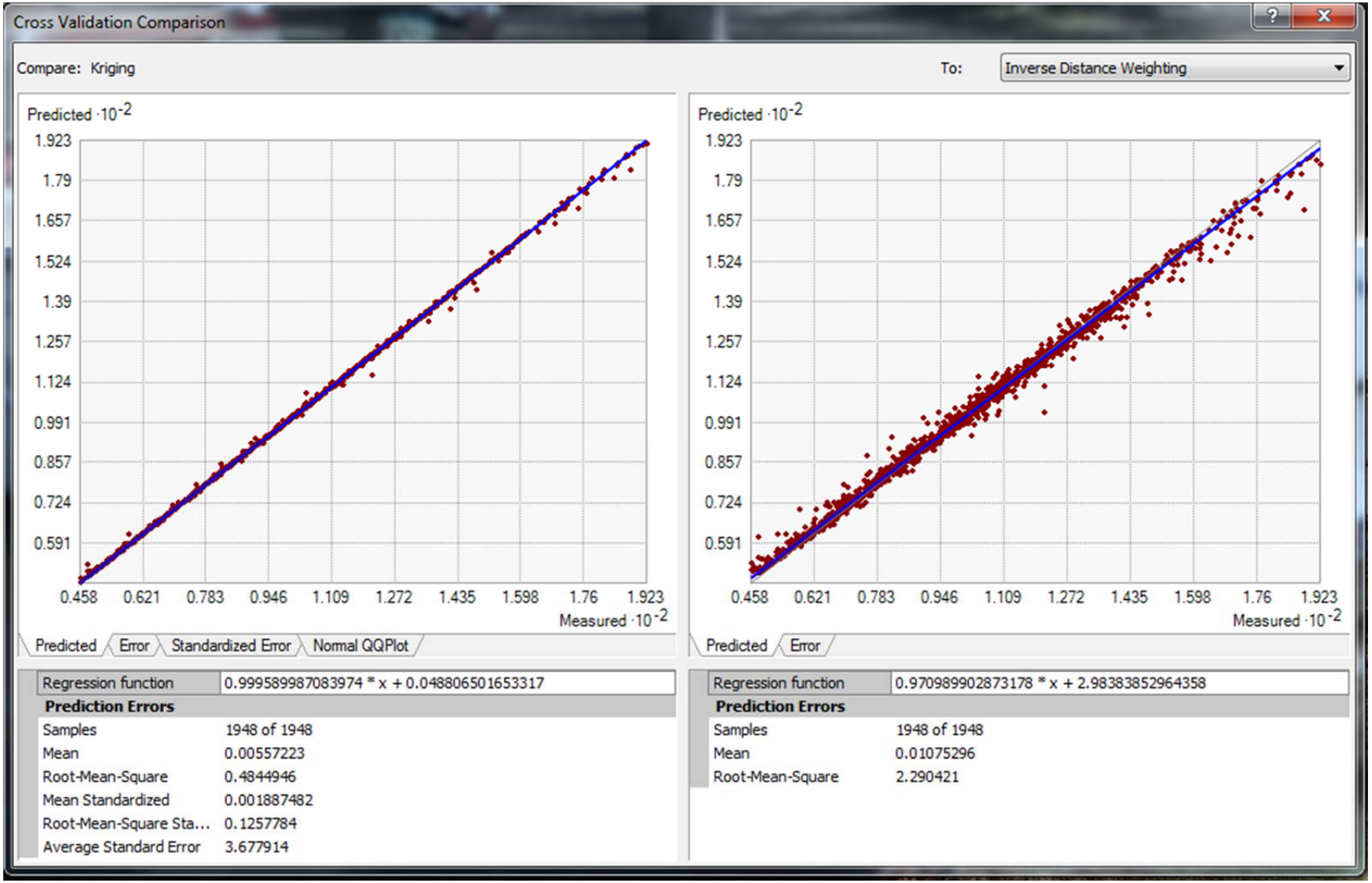
Graph 33: Cross-validation comparison of predicted error for the systematically selected 33% K training data set.



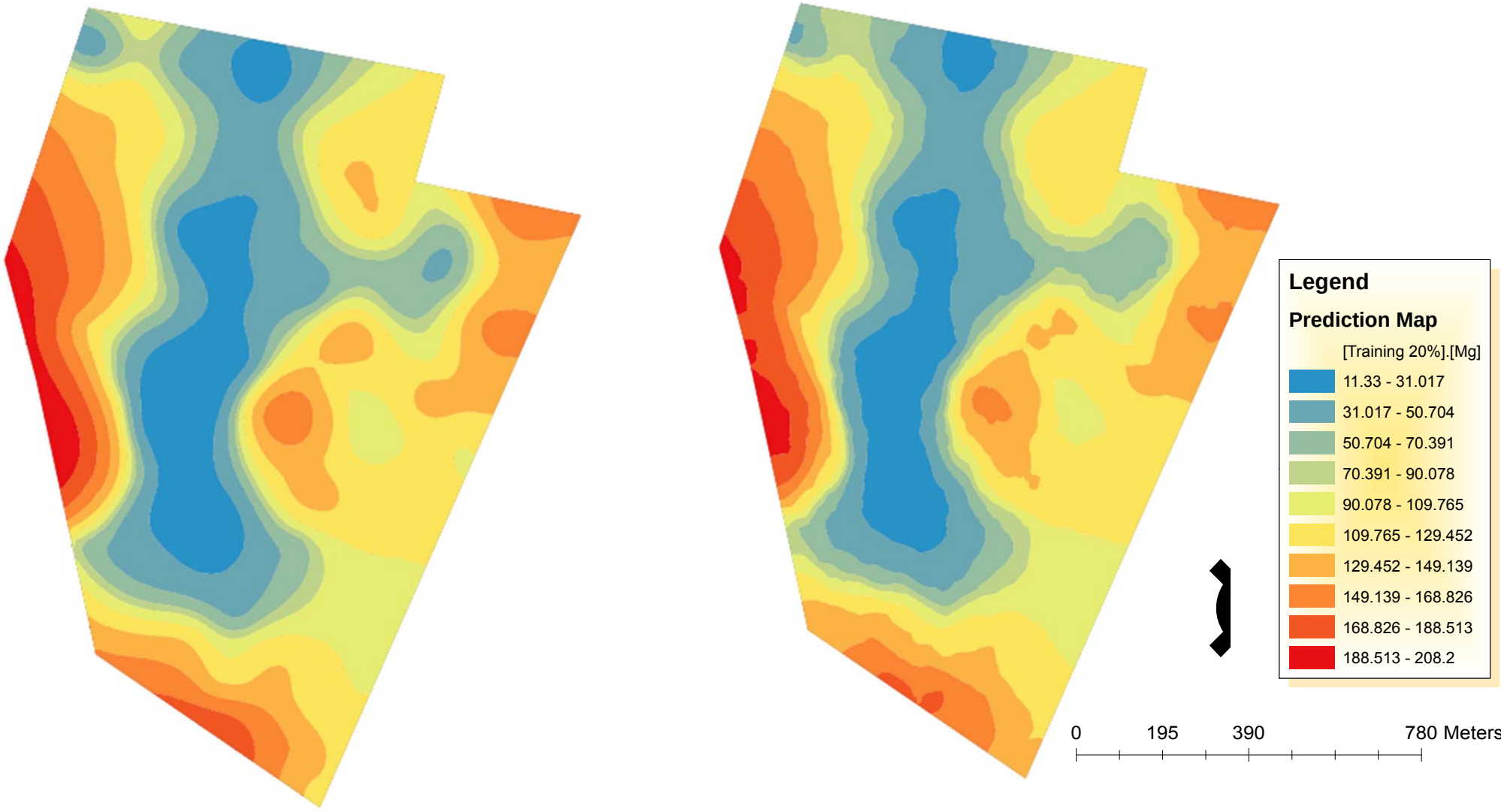
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 49: Prediction maps of systematically selected 50% training and 50% testing (removal of every 2nd data point) data sets for Potassium (K).



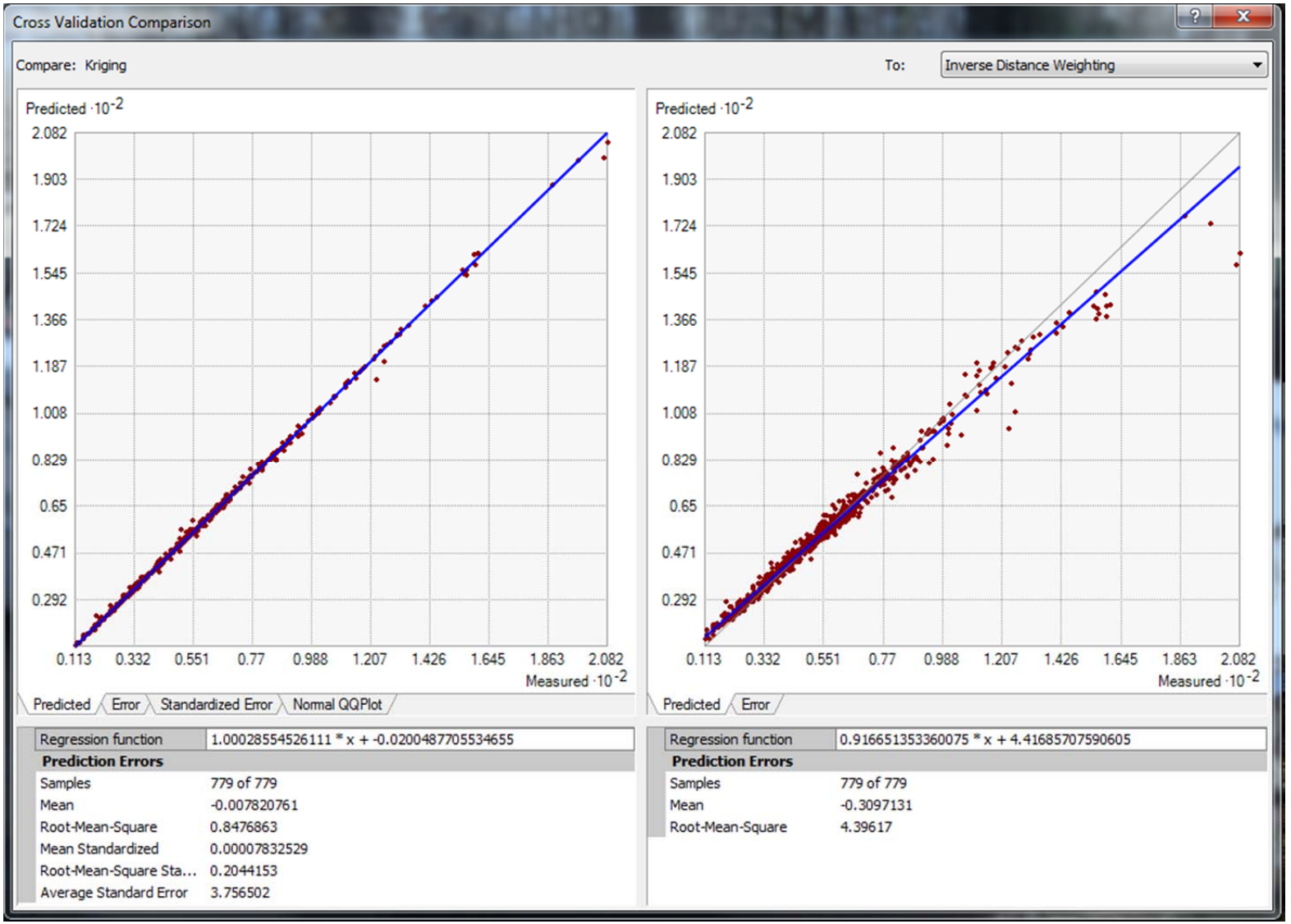
Graph 34: Cross-validation comparison of predicted error for the systematically selected 50% K training data set.



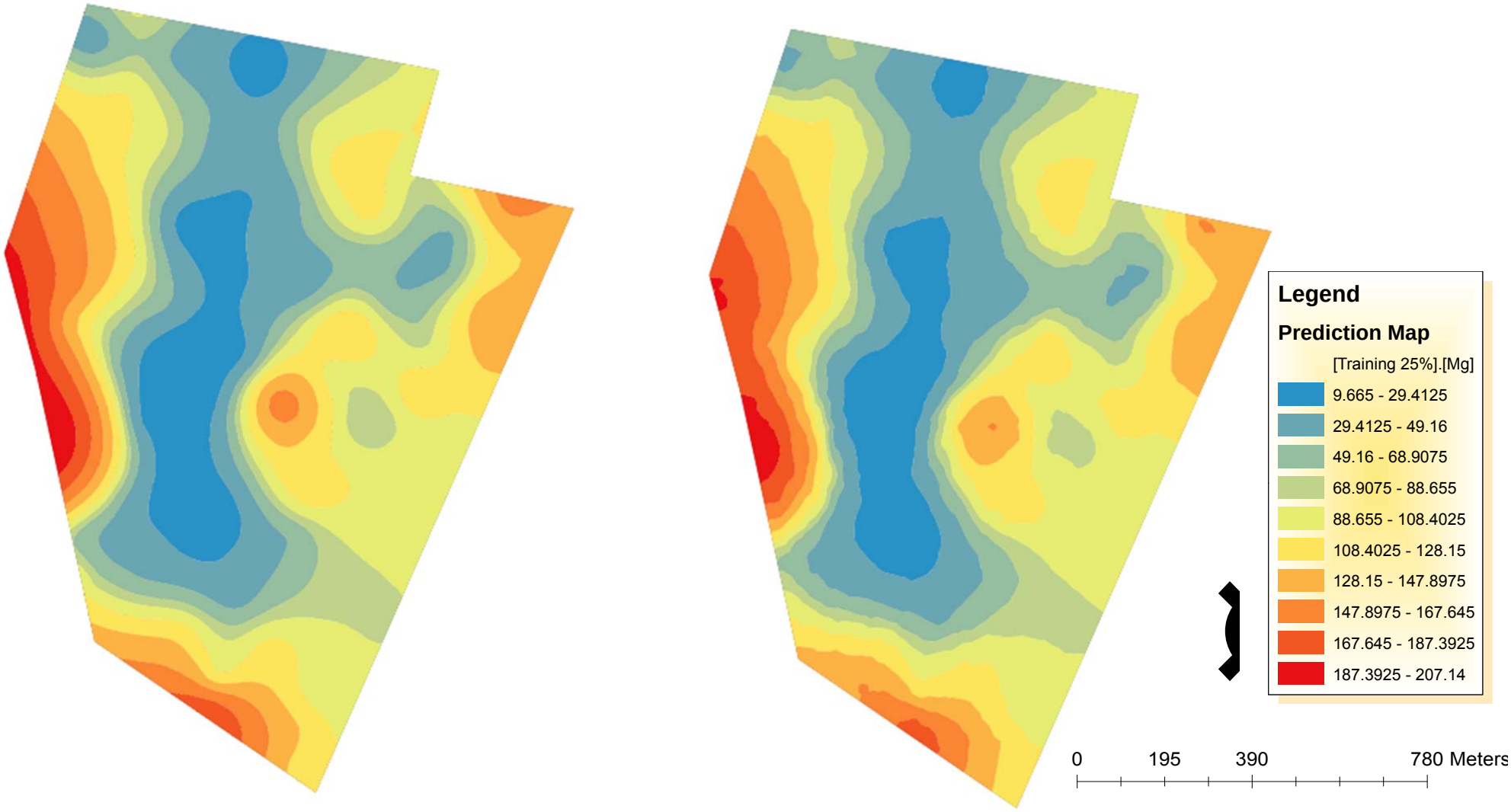
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 50: Prediction maps of systematically selected 20% training and 80% testing (removal of every 5th data point) data sets for Magnesium (Mg).



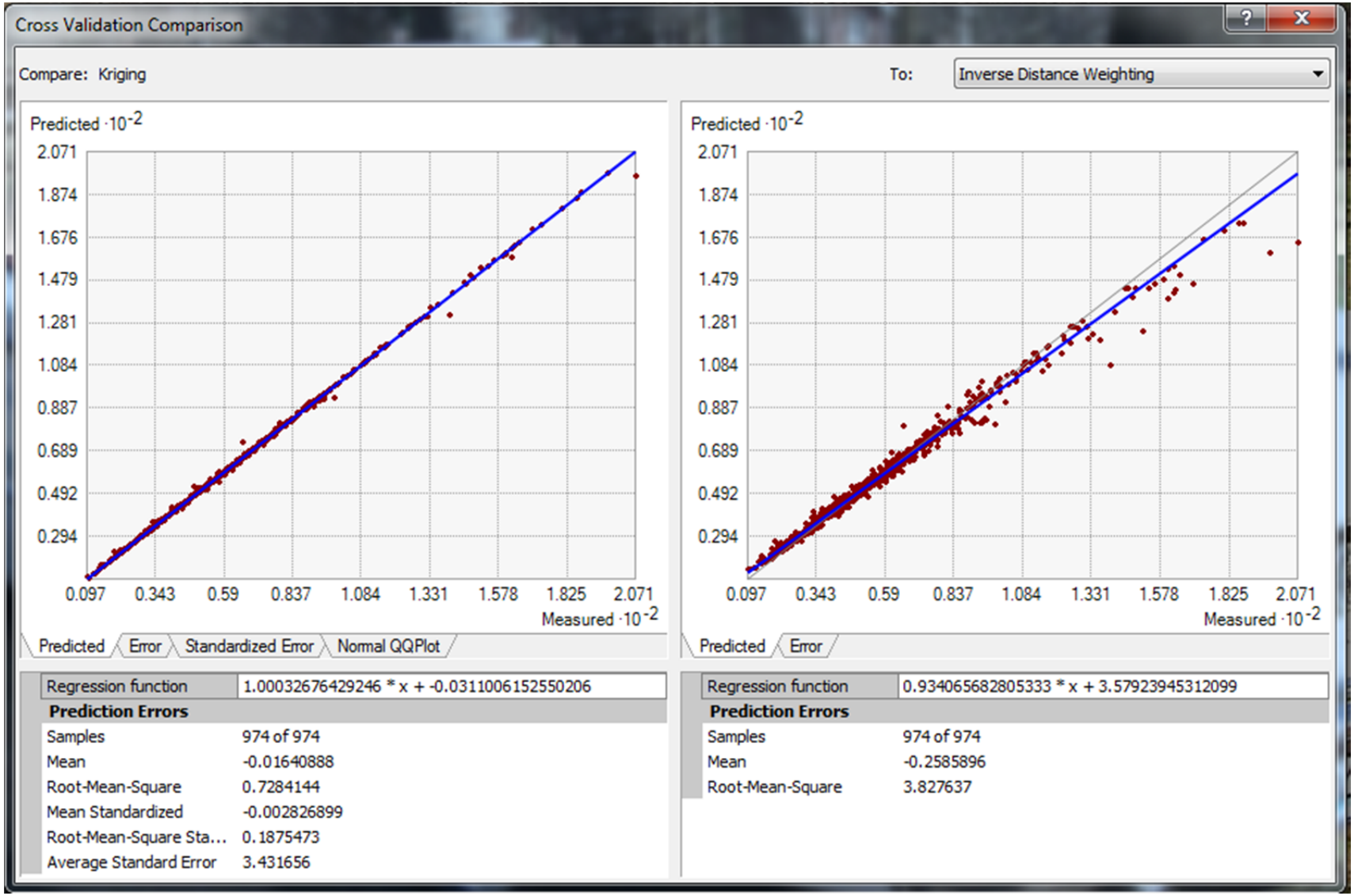
Graph 35: Cross-validation comparison of predicted error for the systematically selected 20% Mg training data set.



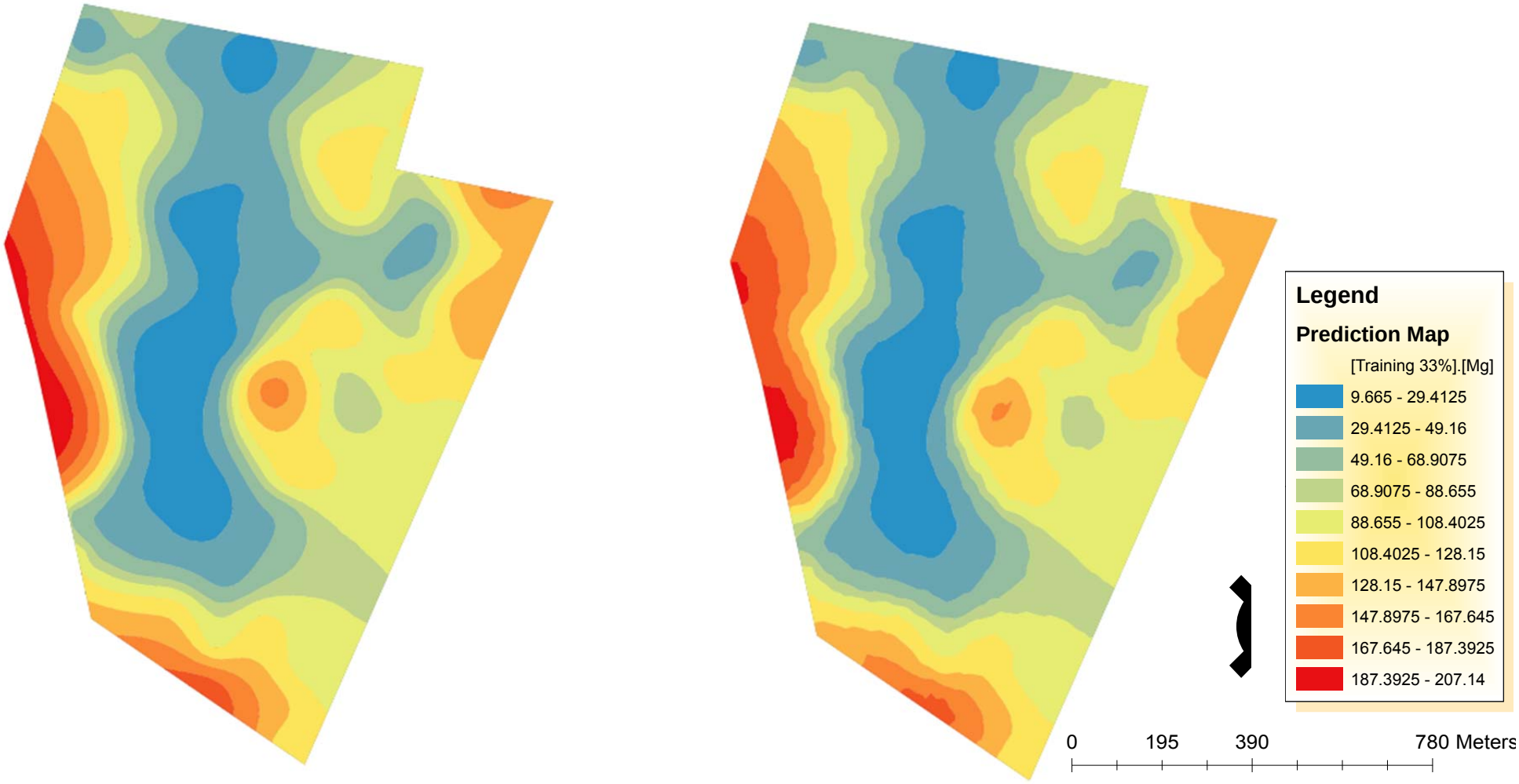
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 51: Prediction maps of systematically selected 25% training and 75% testing (removal of every 4th data point) data sets for Magnesium (Mg).



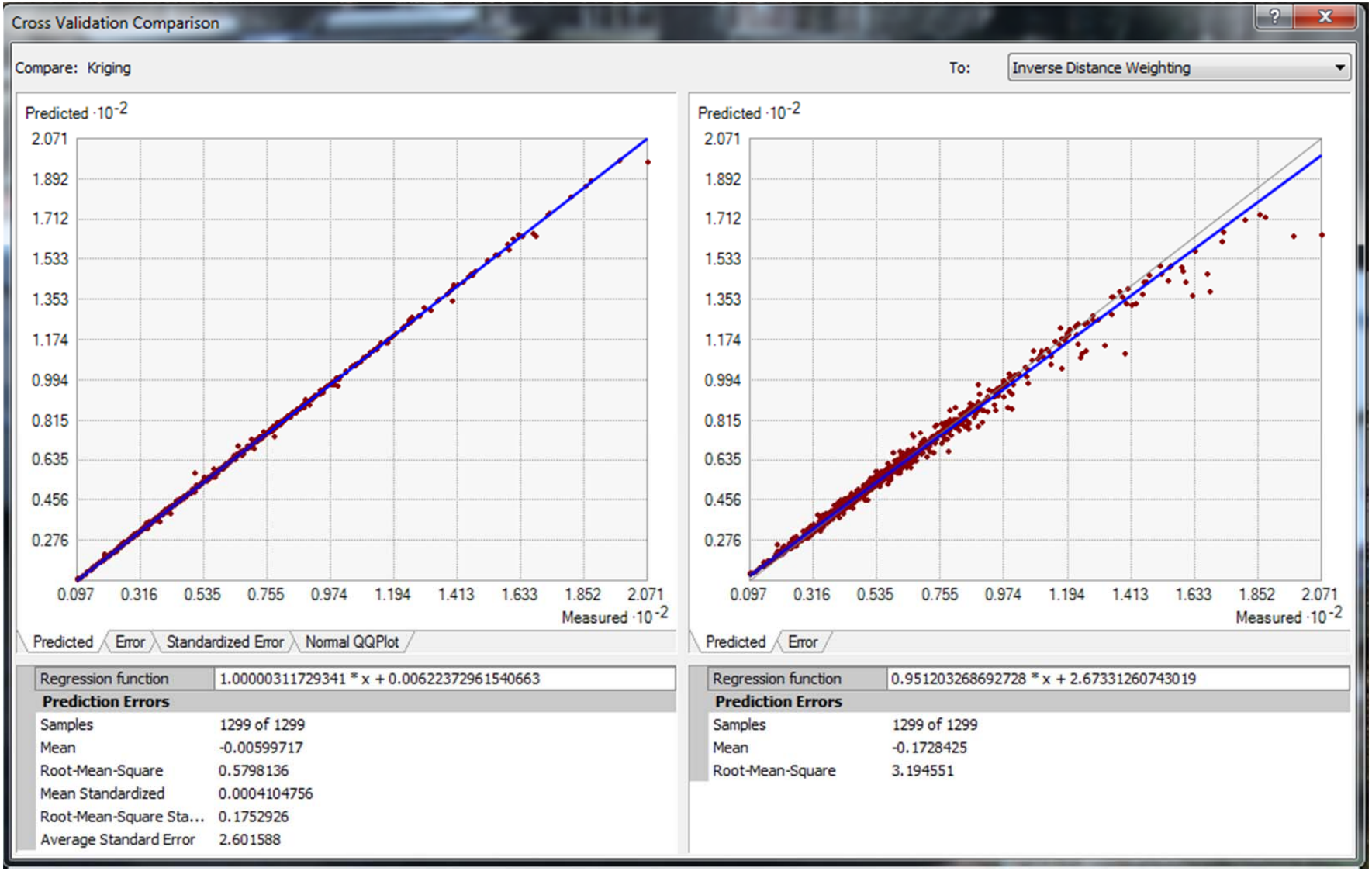
Graph 36: Cross-validation comparison of predicted error for the systematically selected 25% Mg training data set.



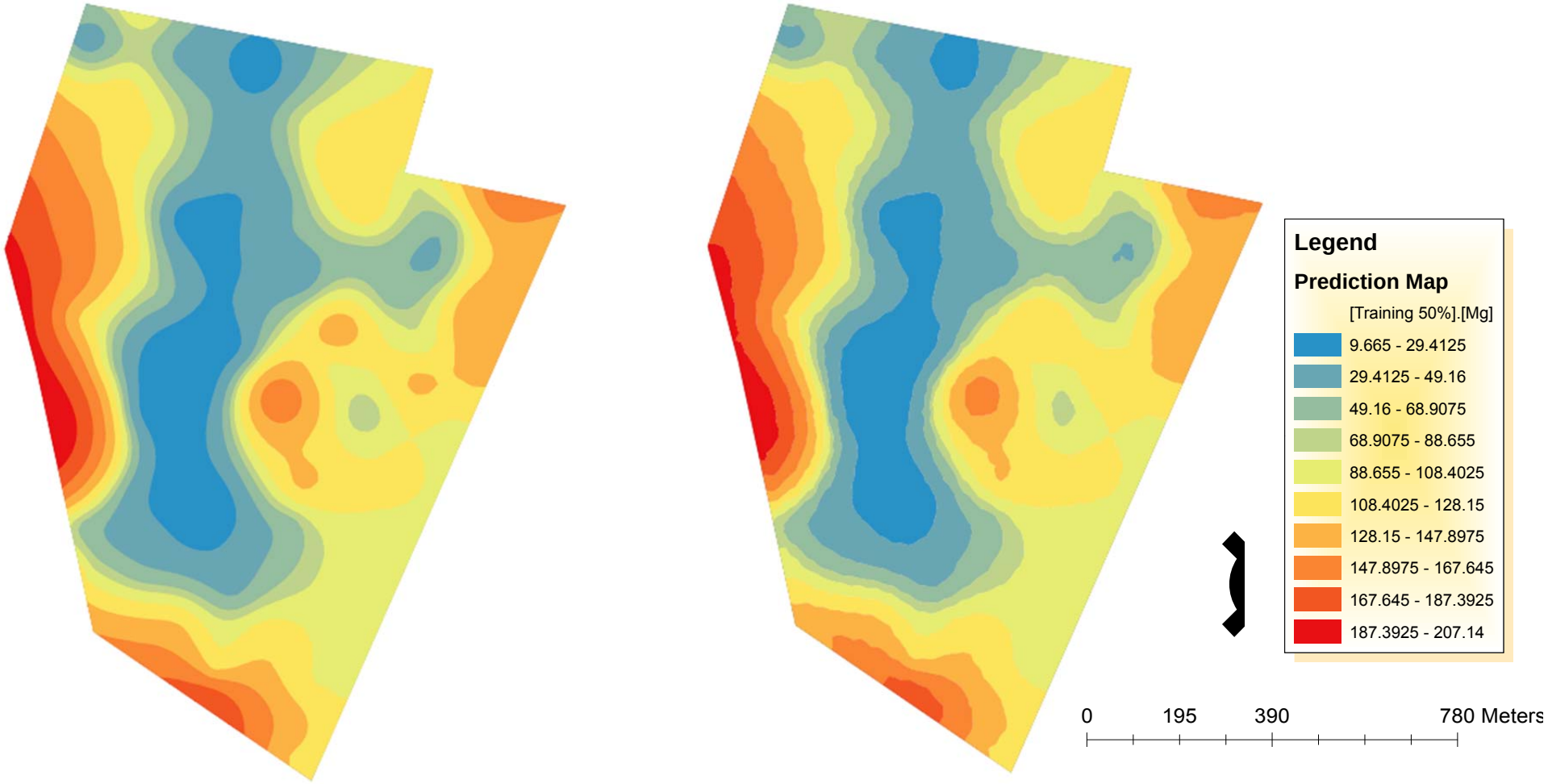
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 52: Prediction maps of systematically selected 33% training and 67% testing (removal of every 3rd data point) data sets for Magnesium (Mg).



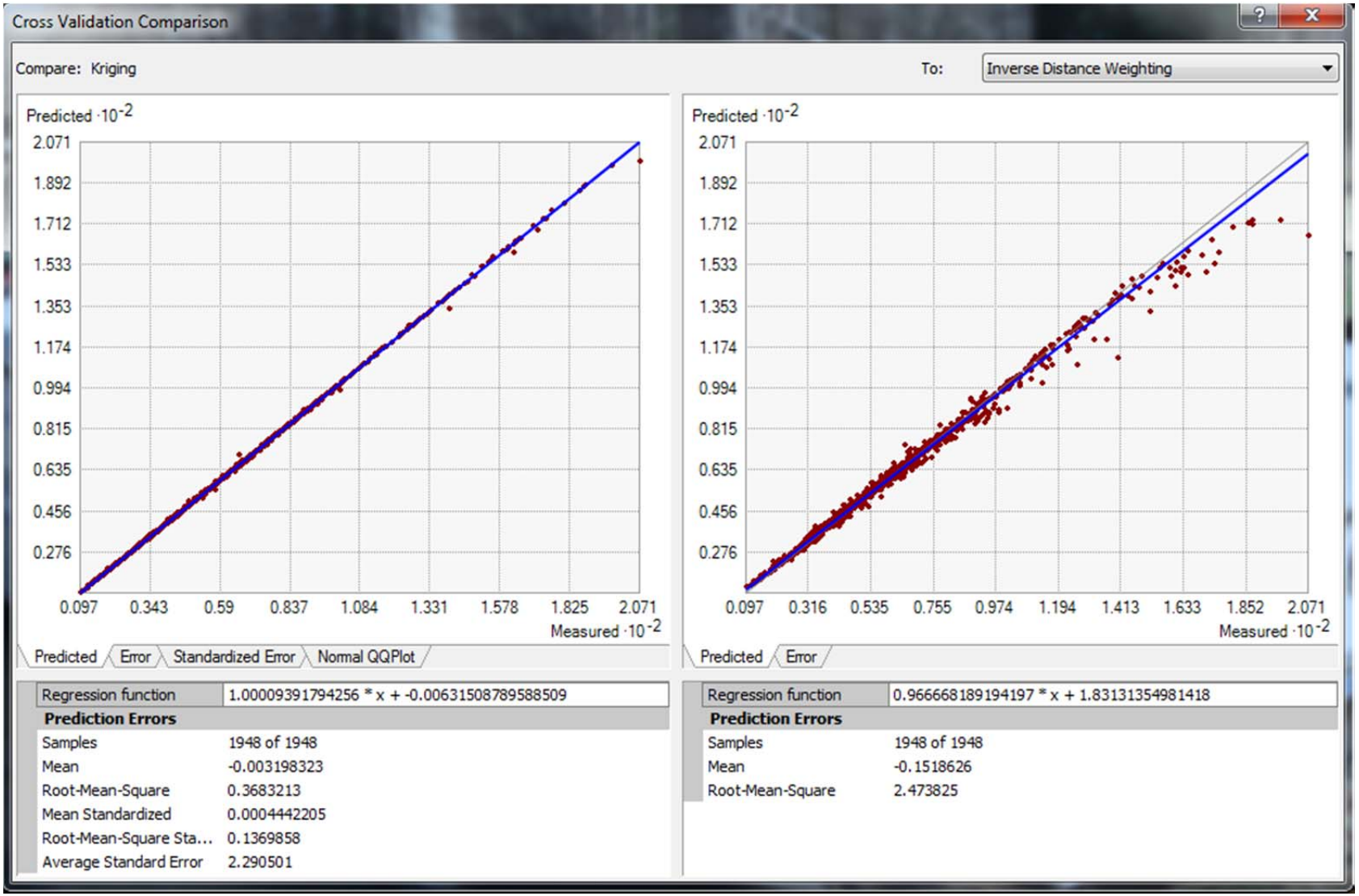
Graph 37: Cross-validation comparison of predicted error for the systematically selected 33% Mg training data set.



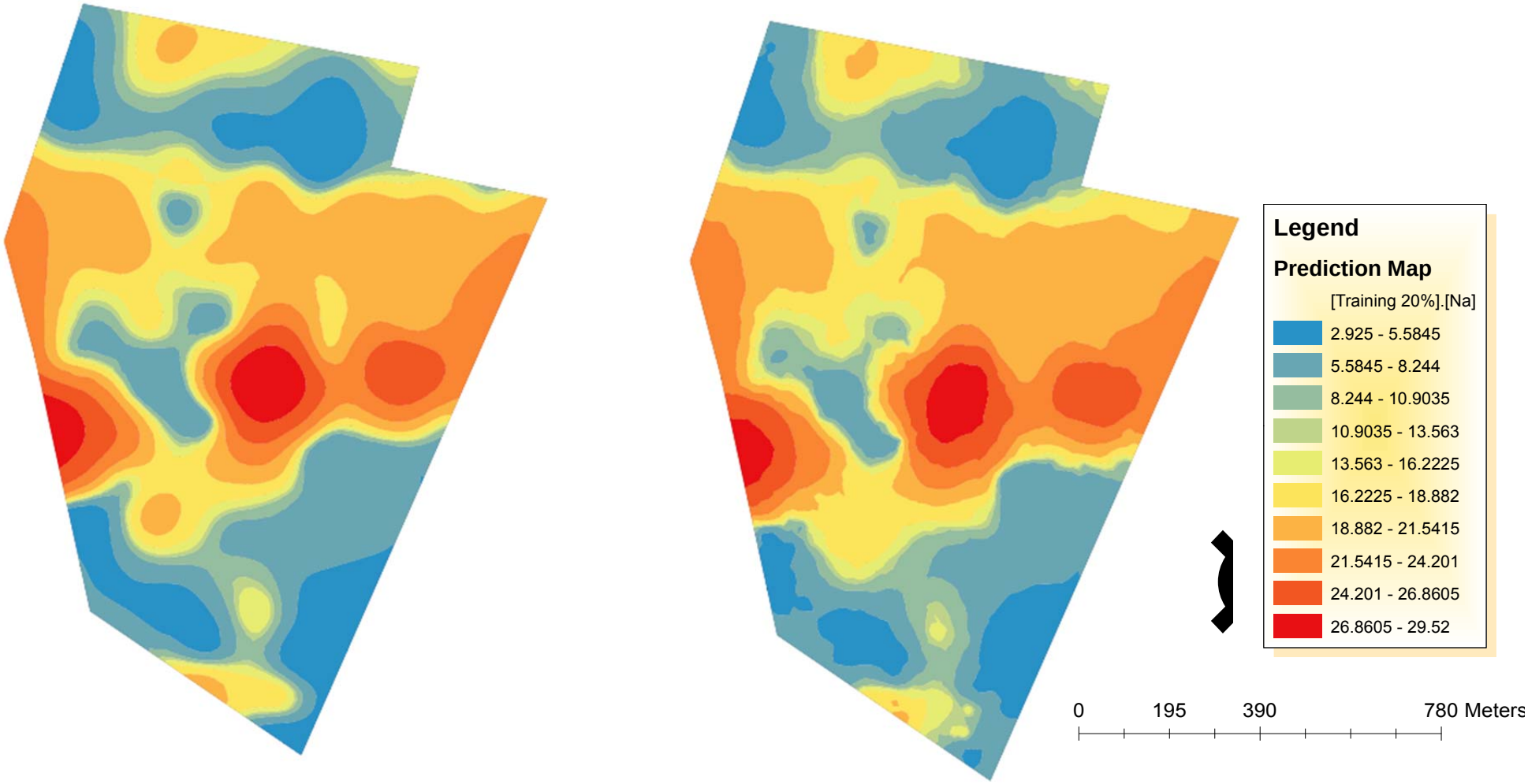
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 53: Prediction maps of systematically selected 50% training and 50% testing (removal of every 2nd data point) data sets for Magnesium (Mg).



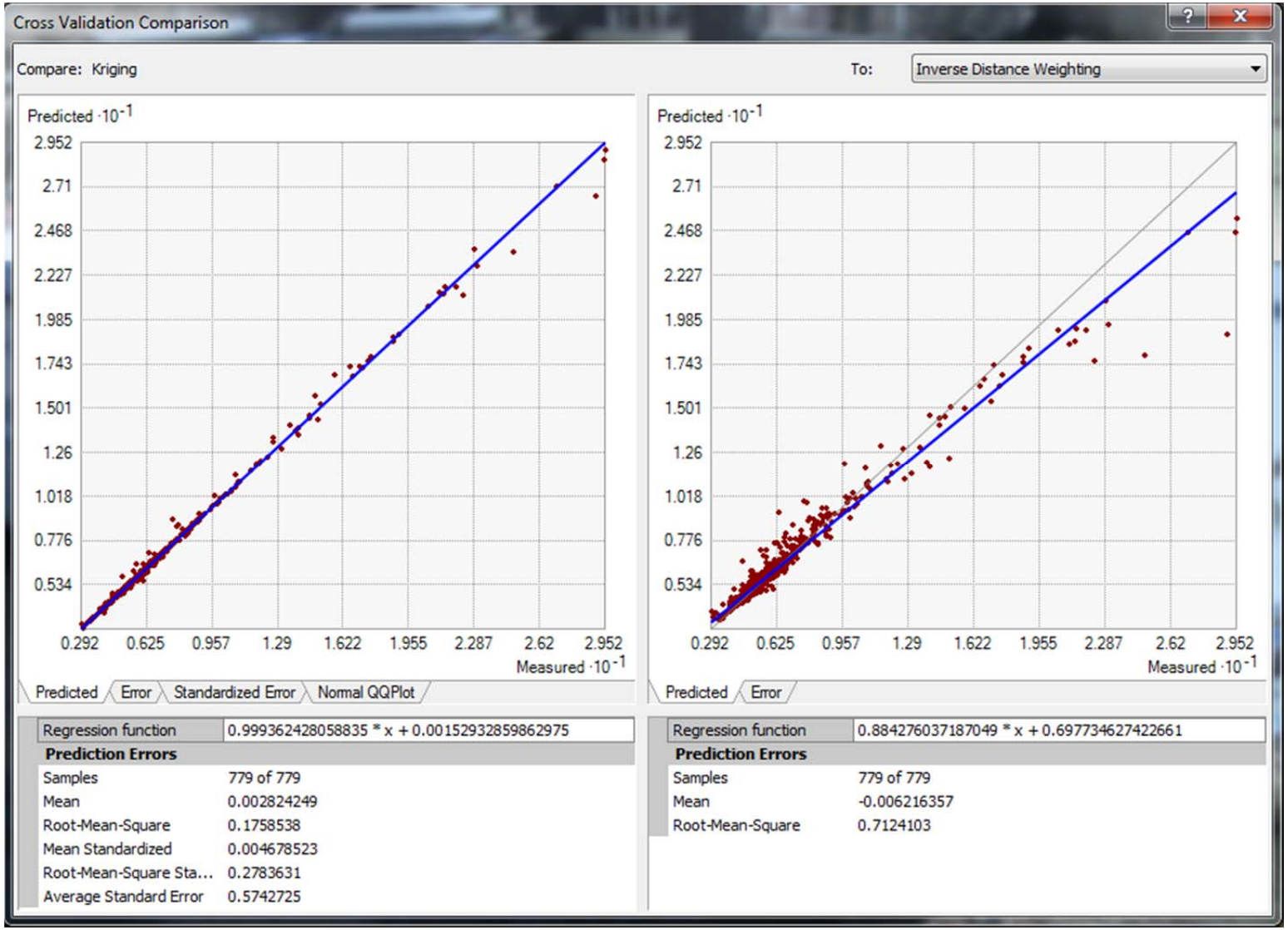
Graph 38: Cross-validation comparison of predicted error for the systematically selected 50% Mg training data set.



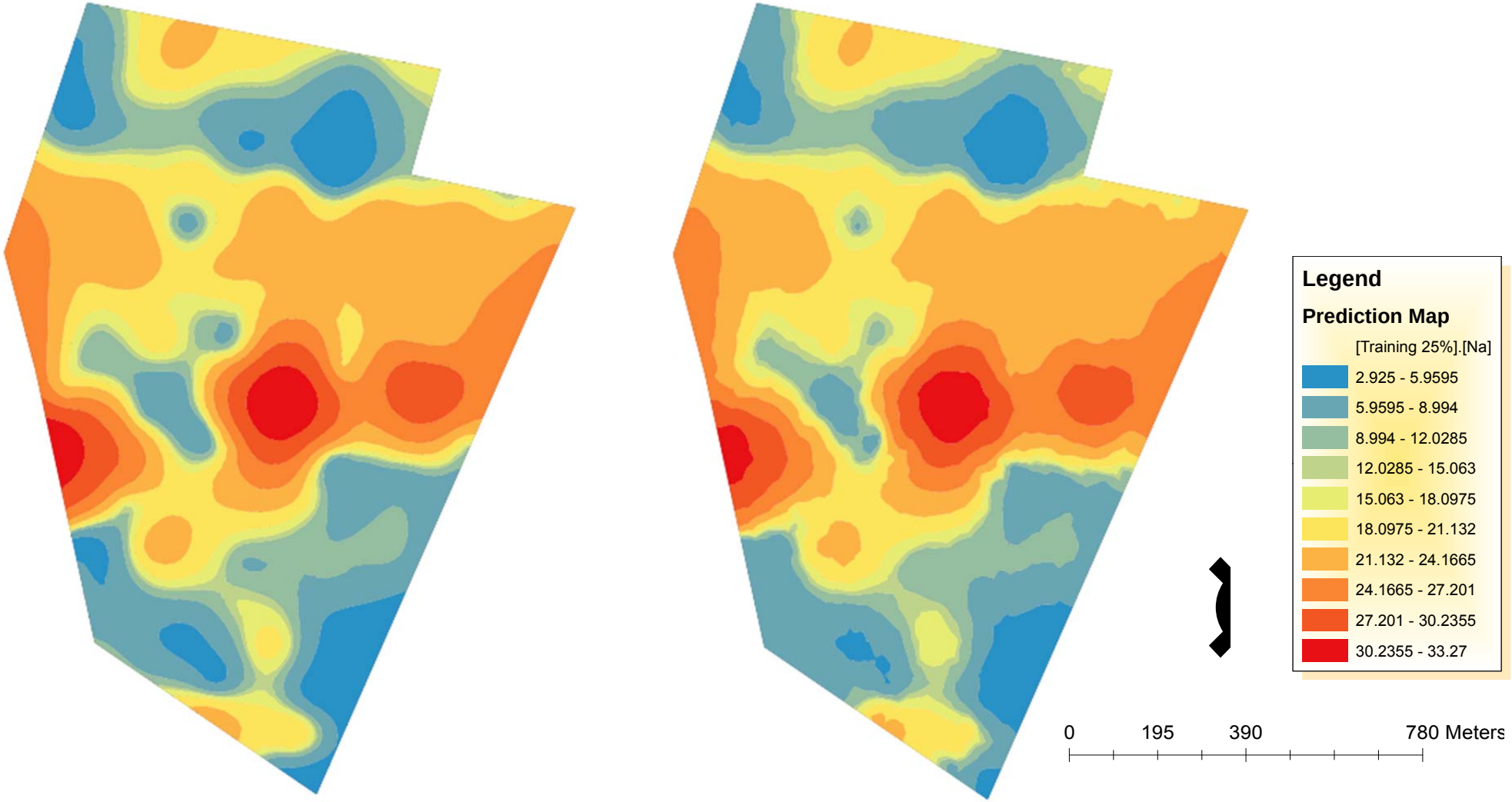
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 54: Prediction maps of systematically selected 20% training and 80% testing (removal of every 5th data point) data sets for Sodium Chloride (NaCl).



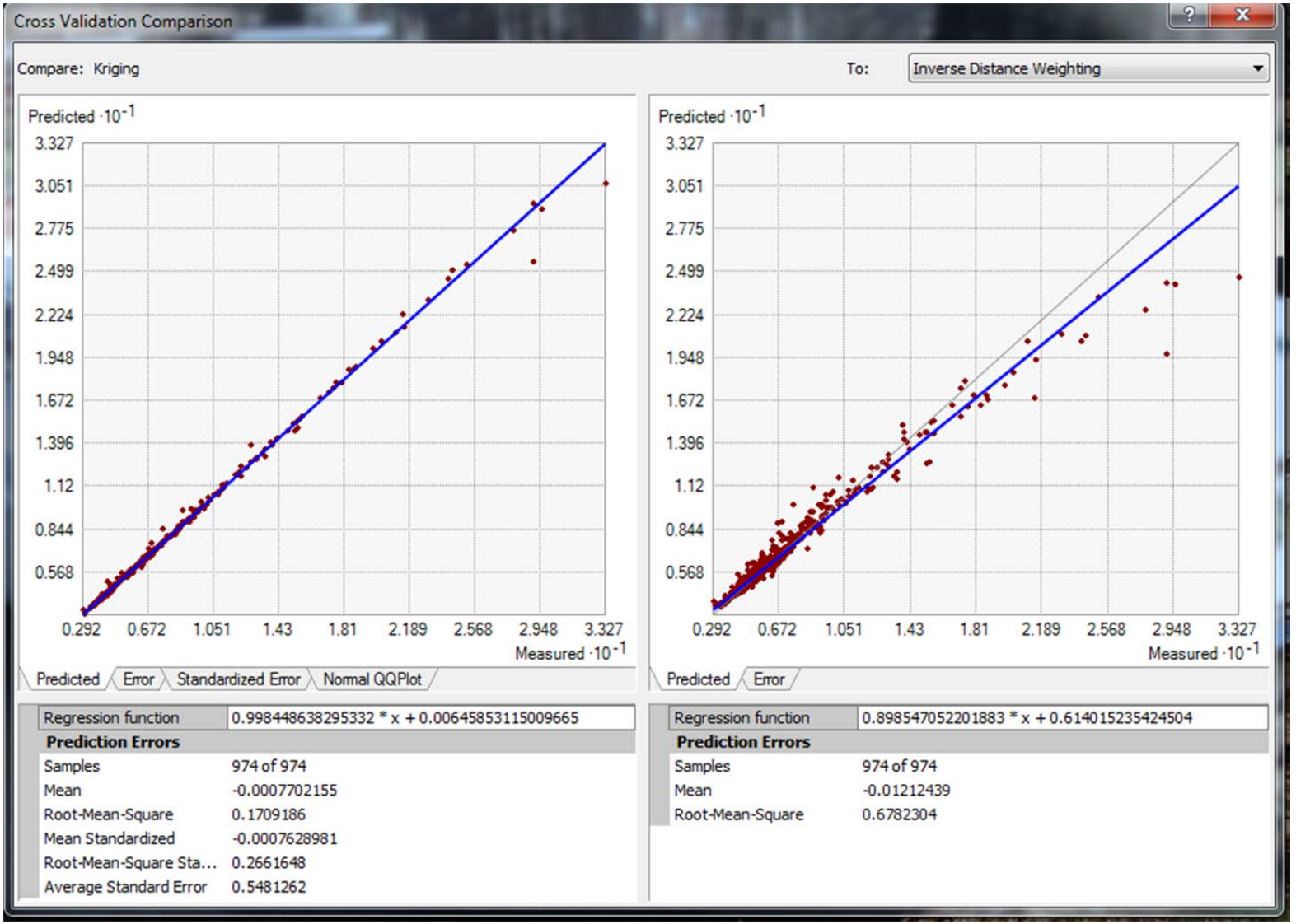
Graph 39: Cross-validation comparison of predicted error for the systematically selected 20% Na training data set.



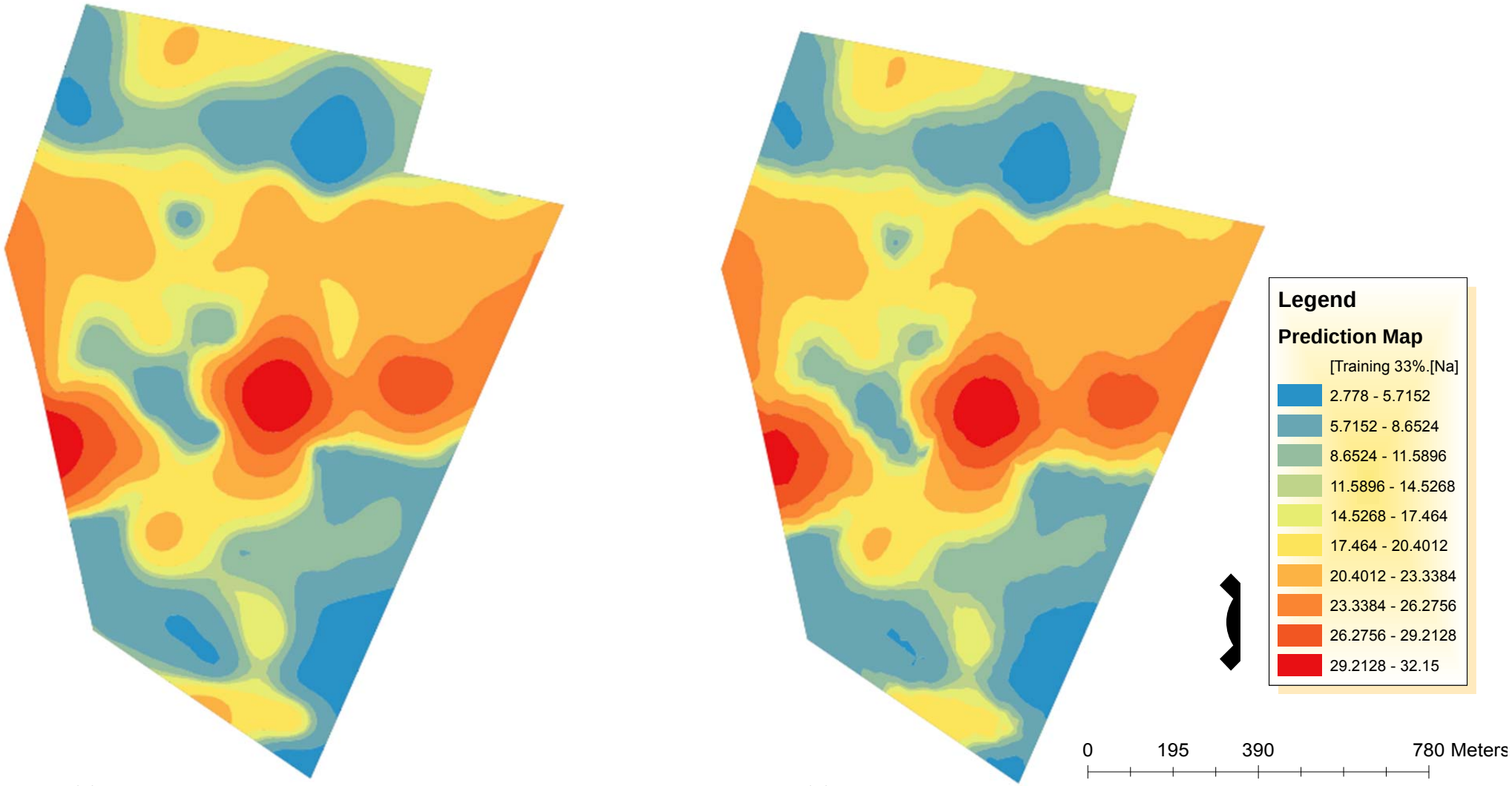
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 55: Prediction maps of systematically selected 25% training and 75% testing (removal of every 4th data point) data sets for Sodium Chloride (NaCl).



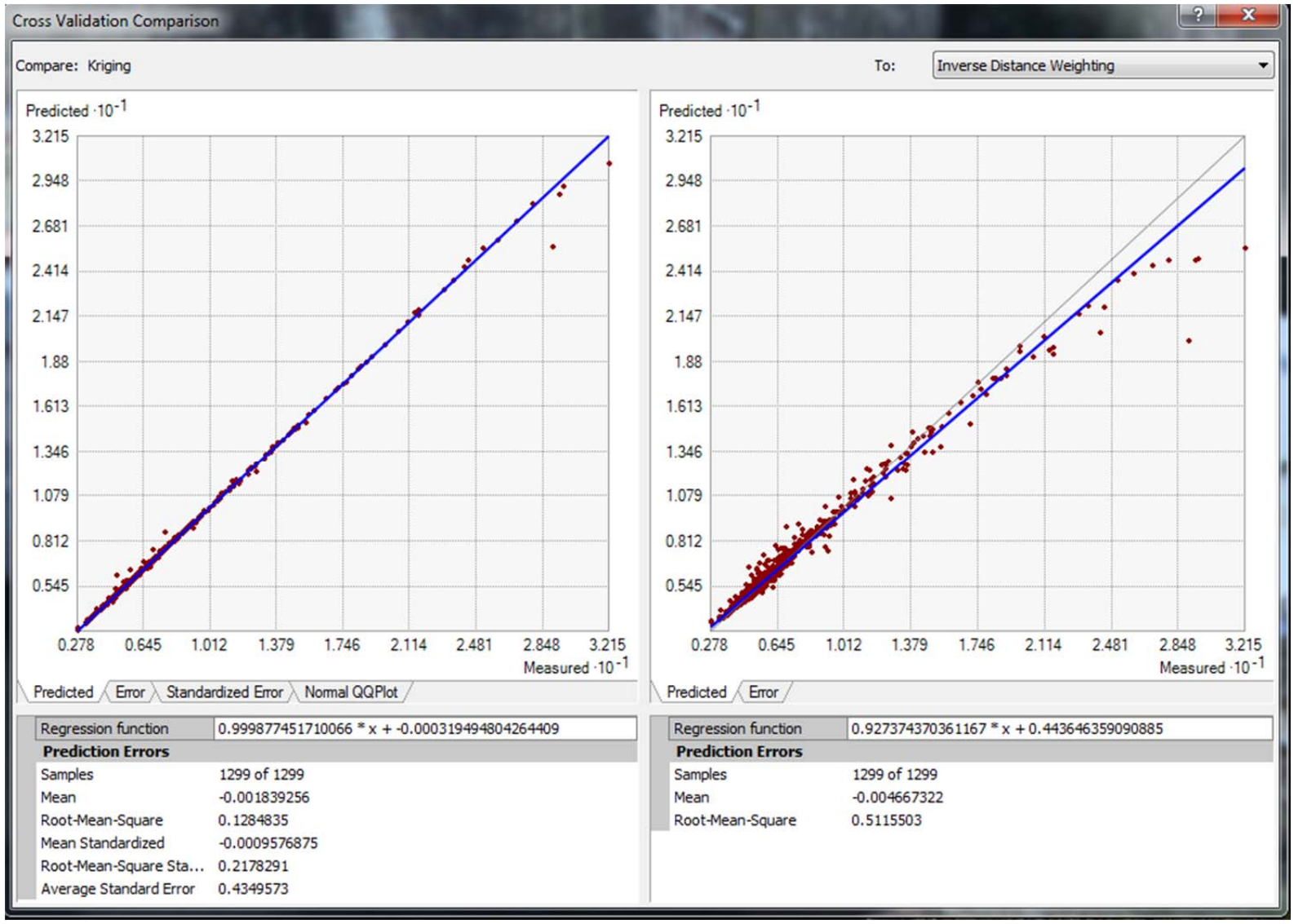
Graph 40: Cross-validation comparison of predicted error for the systematically selected 25% Na training data set.



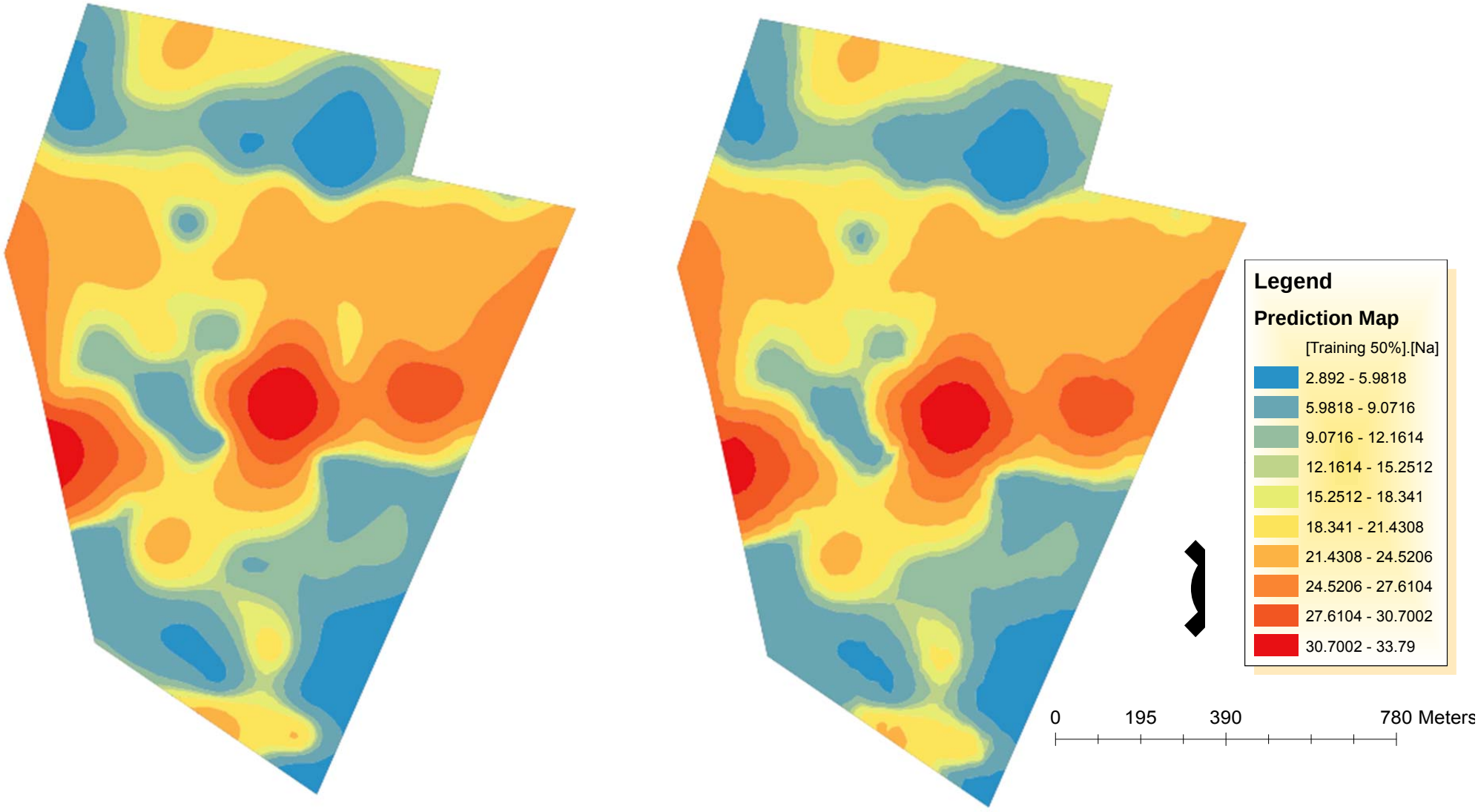
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 56: Prediction maps of systematically selected 33% training and 67% testing (removal of every 3rd data point) data sets for Sodium Chloride (NaCl).



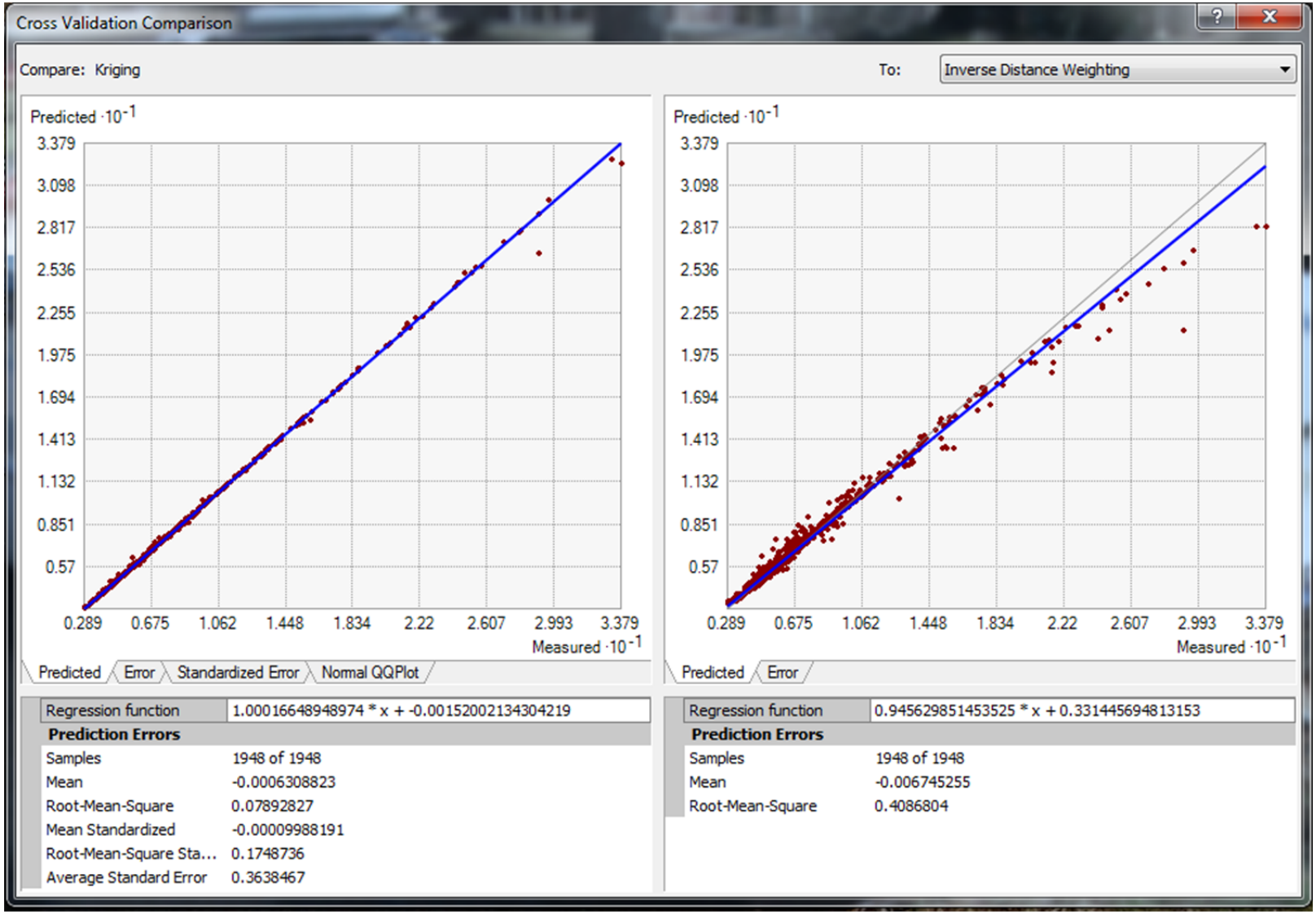
Graph 41: Cross-validation comparison of predicted error for the systematically selected 33% Na training data set.



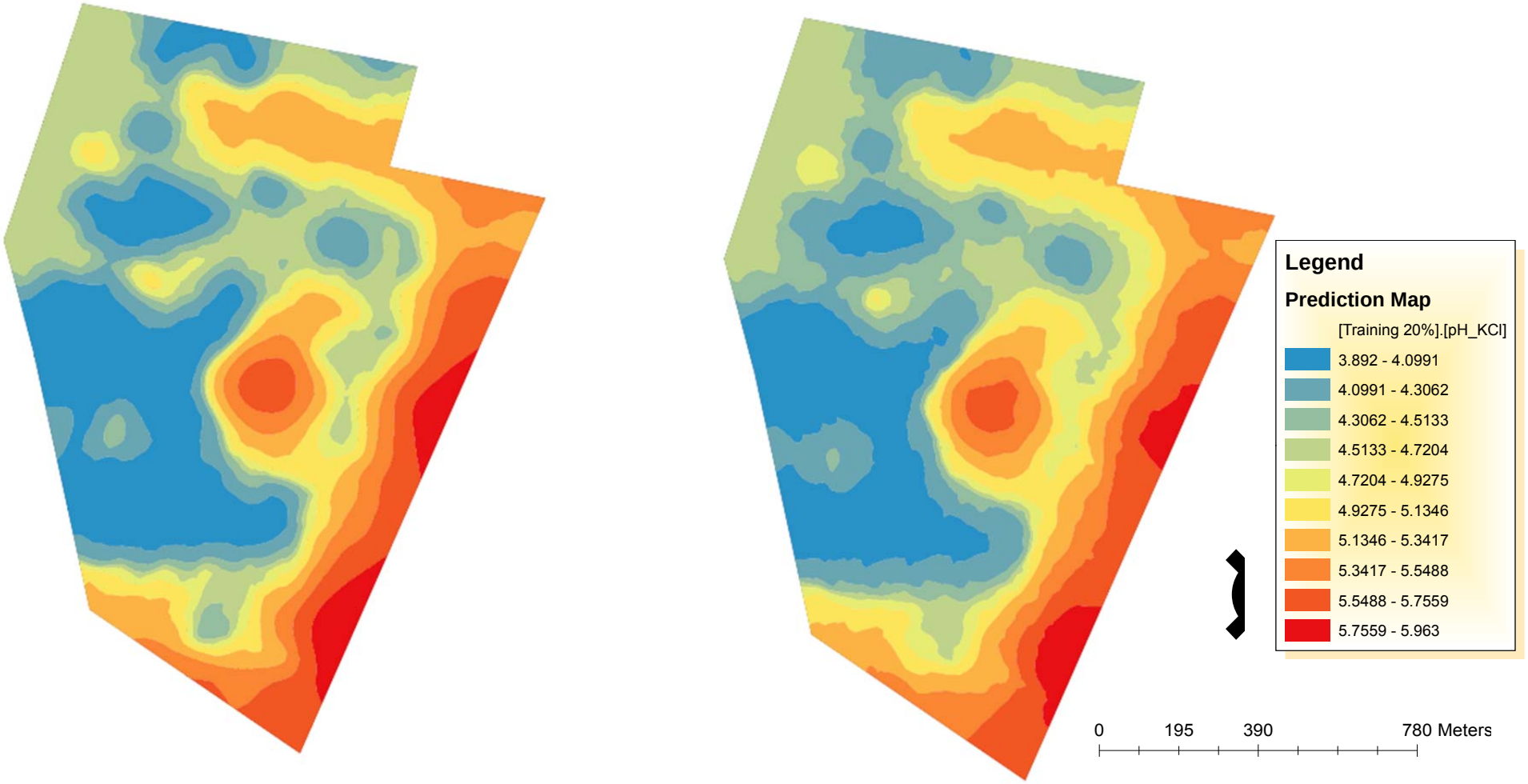
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Fig. 57: Prediction maps of systematically selected 33% training and 67% testing (removal of every 2nd data point) data sets for Sodium Chloride (NaCl).



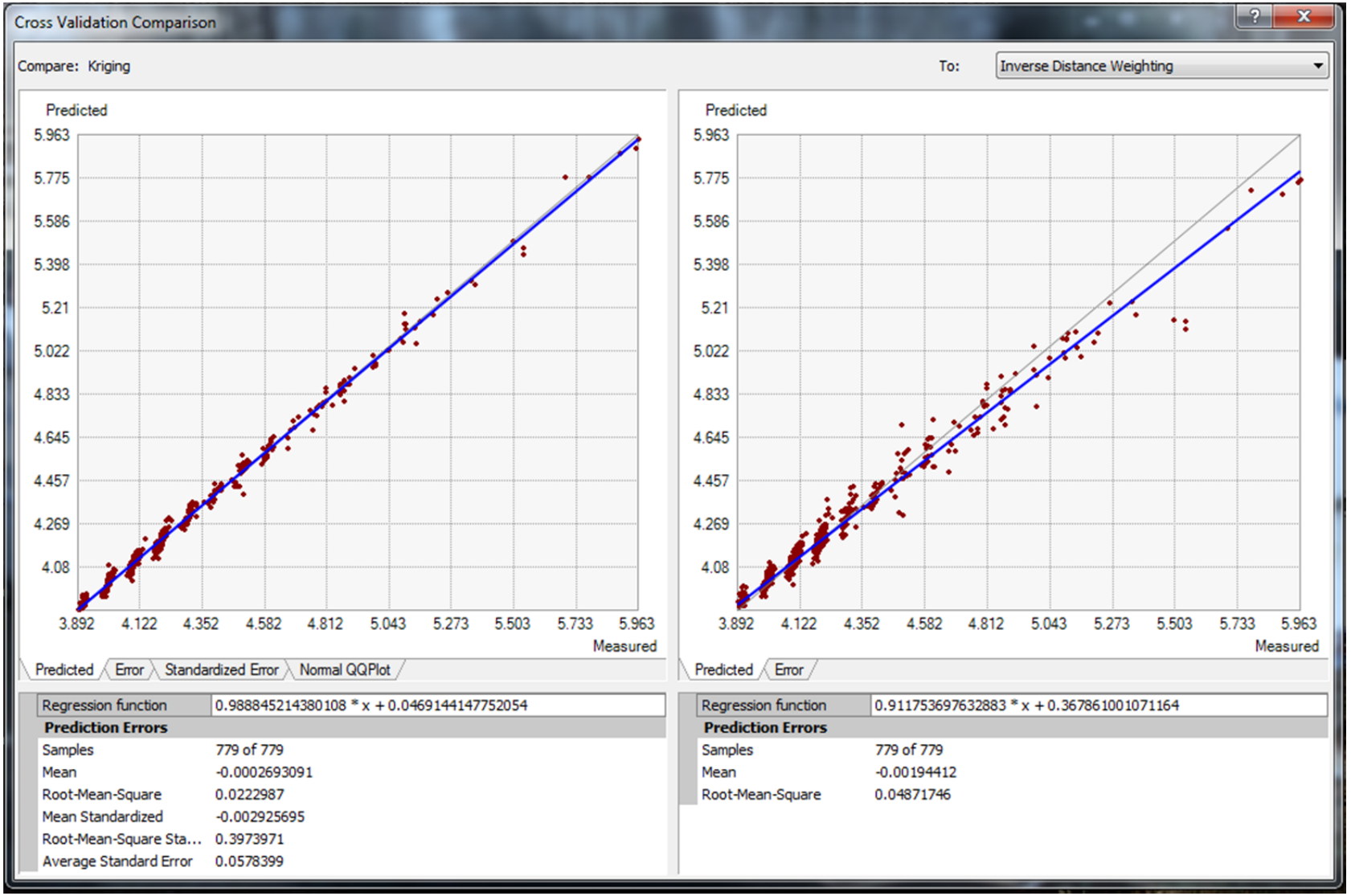
Graph 42: Cross-validation comparison of predicted error for the systematically selected 50% Na training data set.



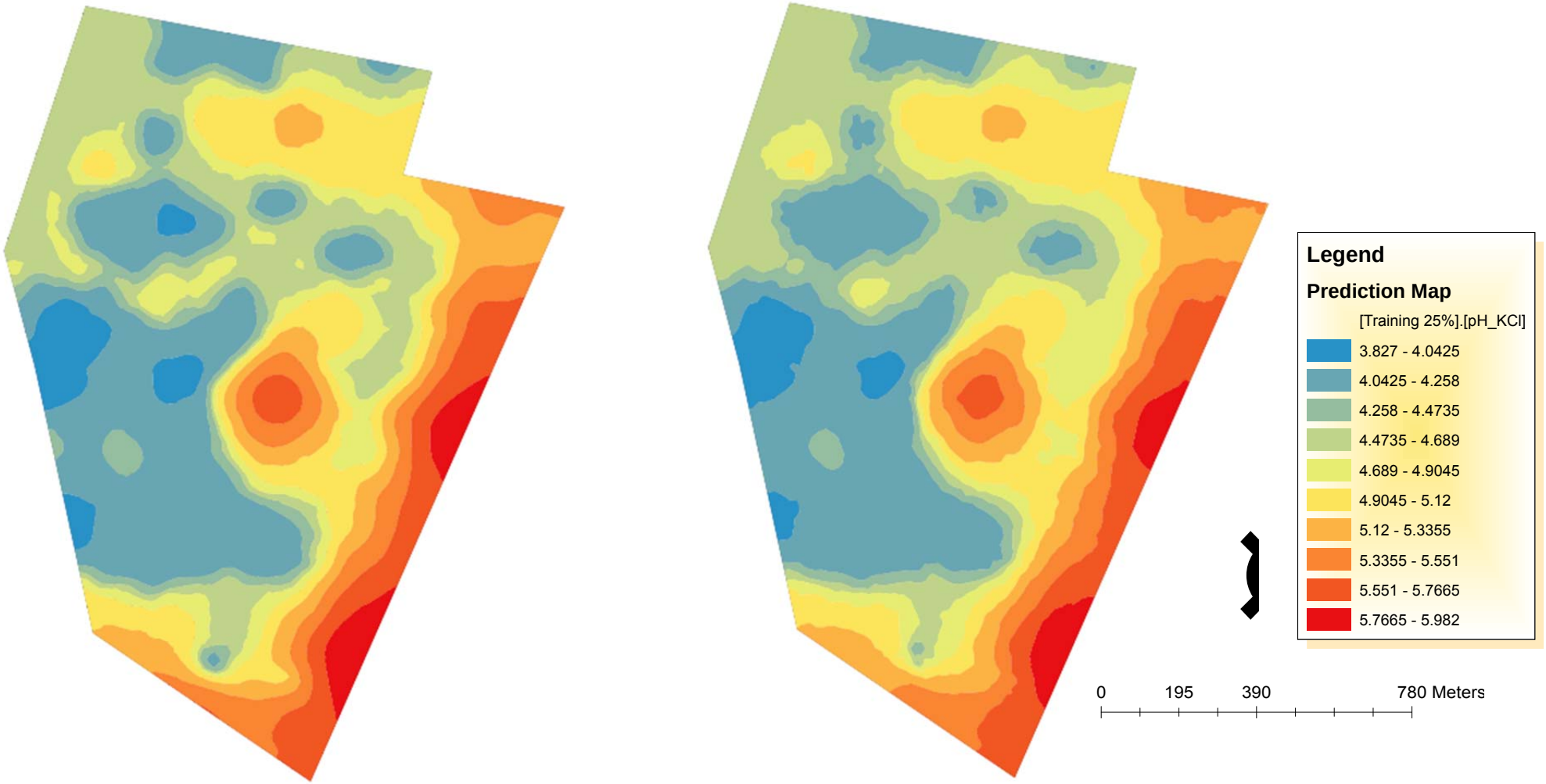
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 58: Prediction maps of systematically selected 20% training and 80% testing (removal of every 5th data point) data sets for the pH of Potassium Chloride (pH-KCl).



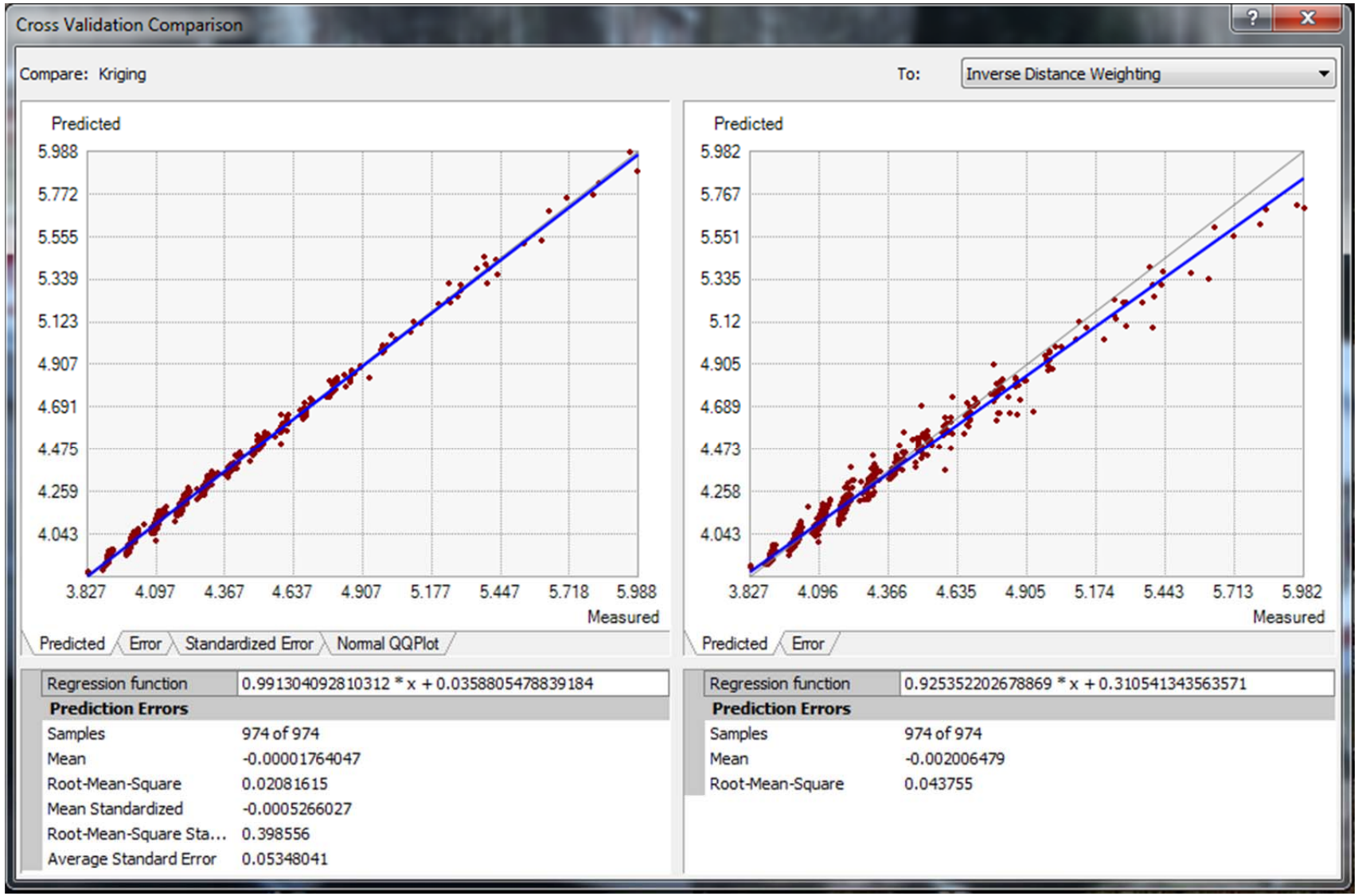
Graph 43: Cross-validation comparison of predicted error for the systematically selected 20% pH-KCl training data set.



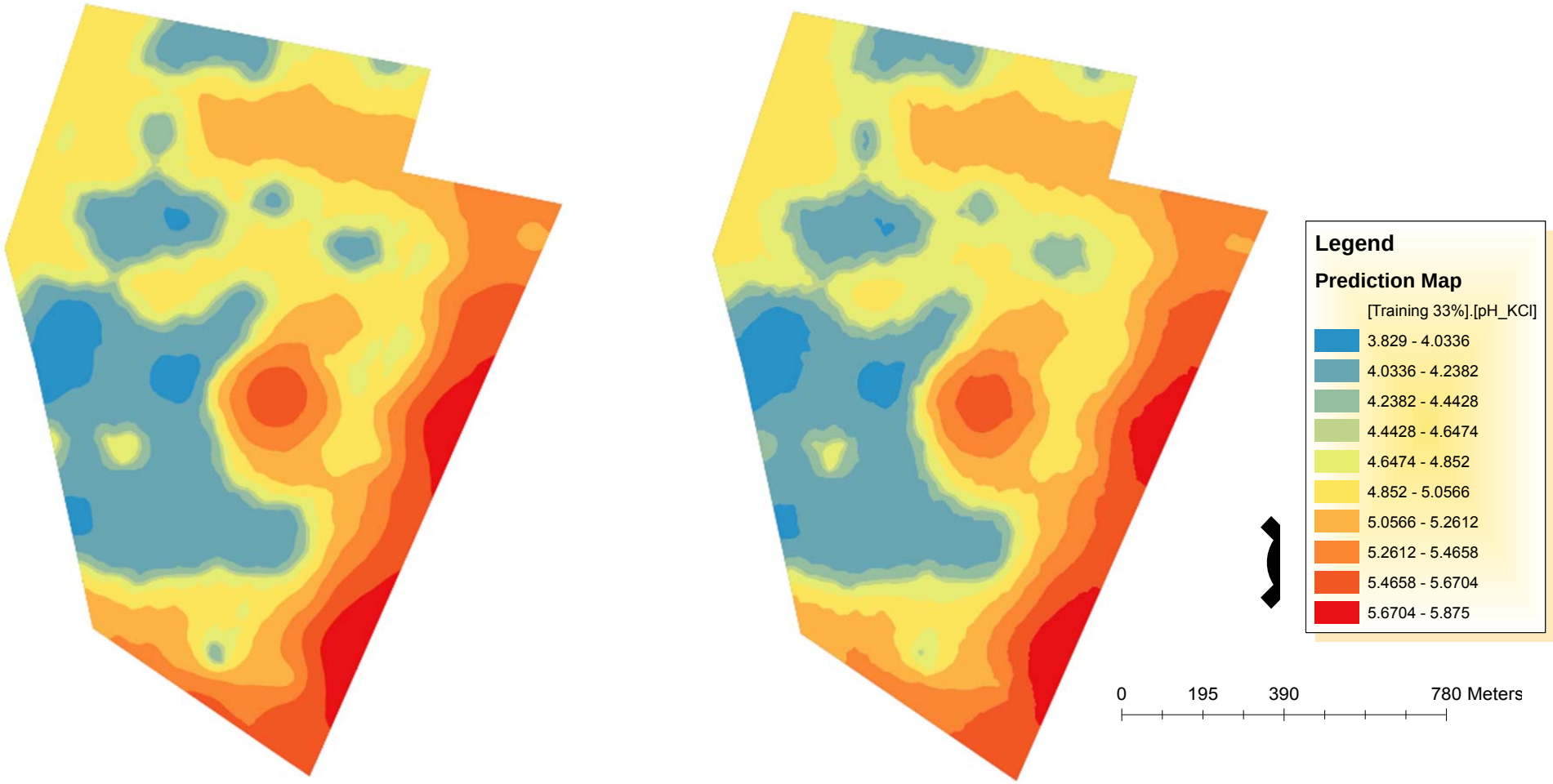
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 59: Prediction maps of systematically selected 25% training and 75% testing (removal of every 4th data point) data sets for the pH of Potassium Chloride (pH-KCl).



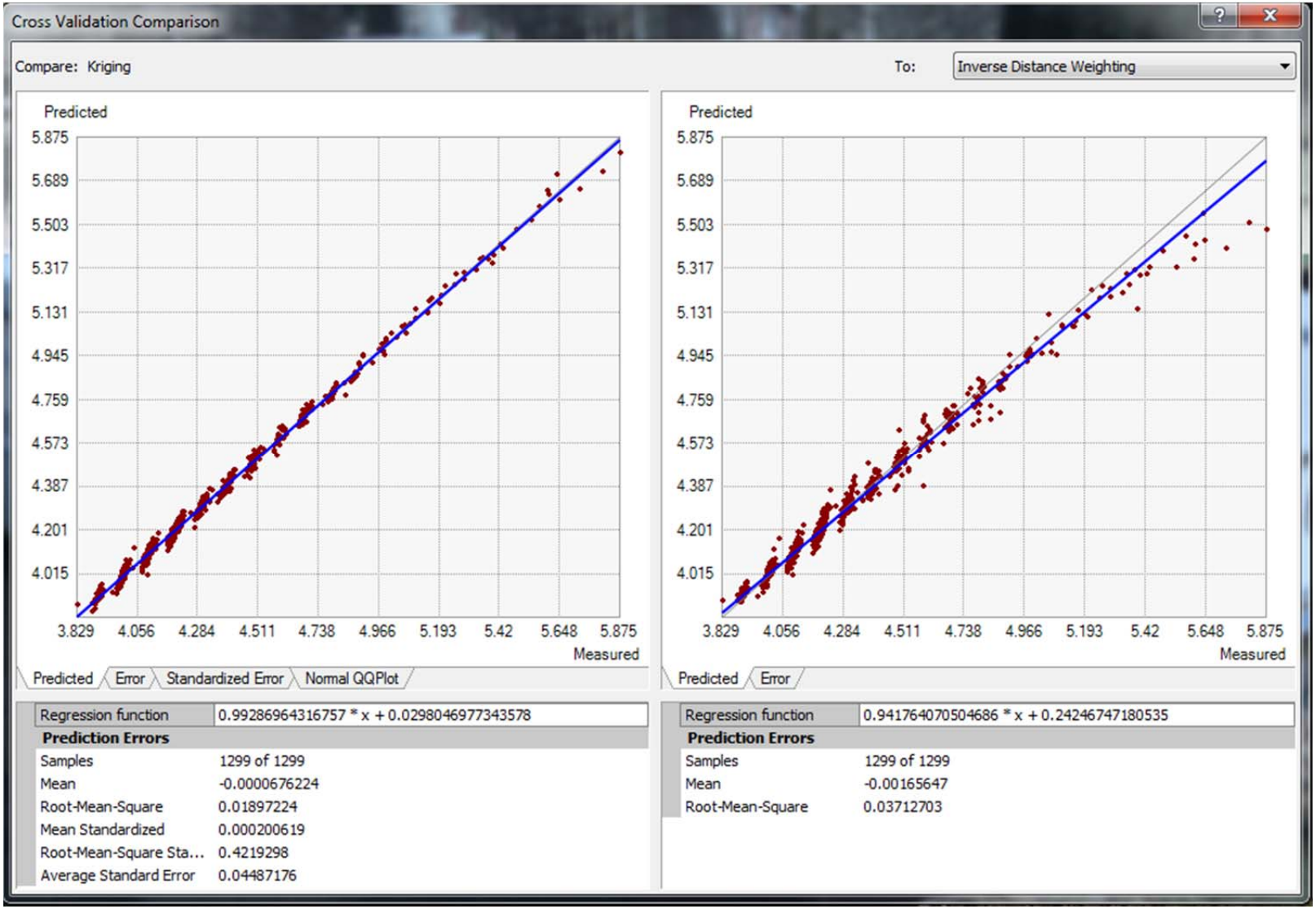
Graph 44: Cross-validation comparison of predicted error for the systematically selected 25% pH-KCl training data set.



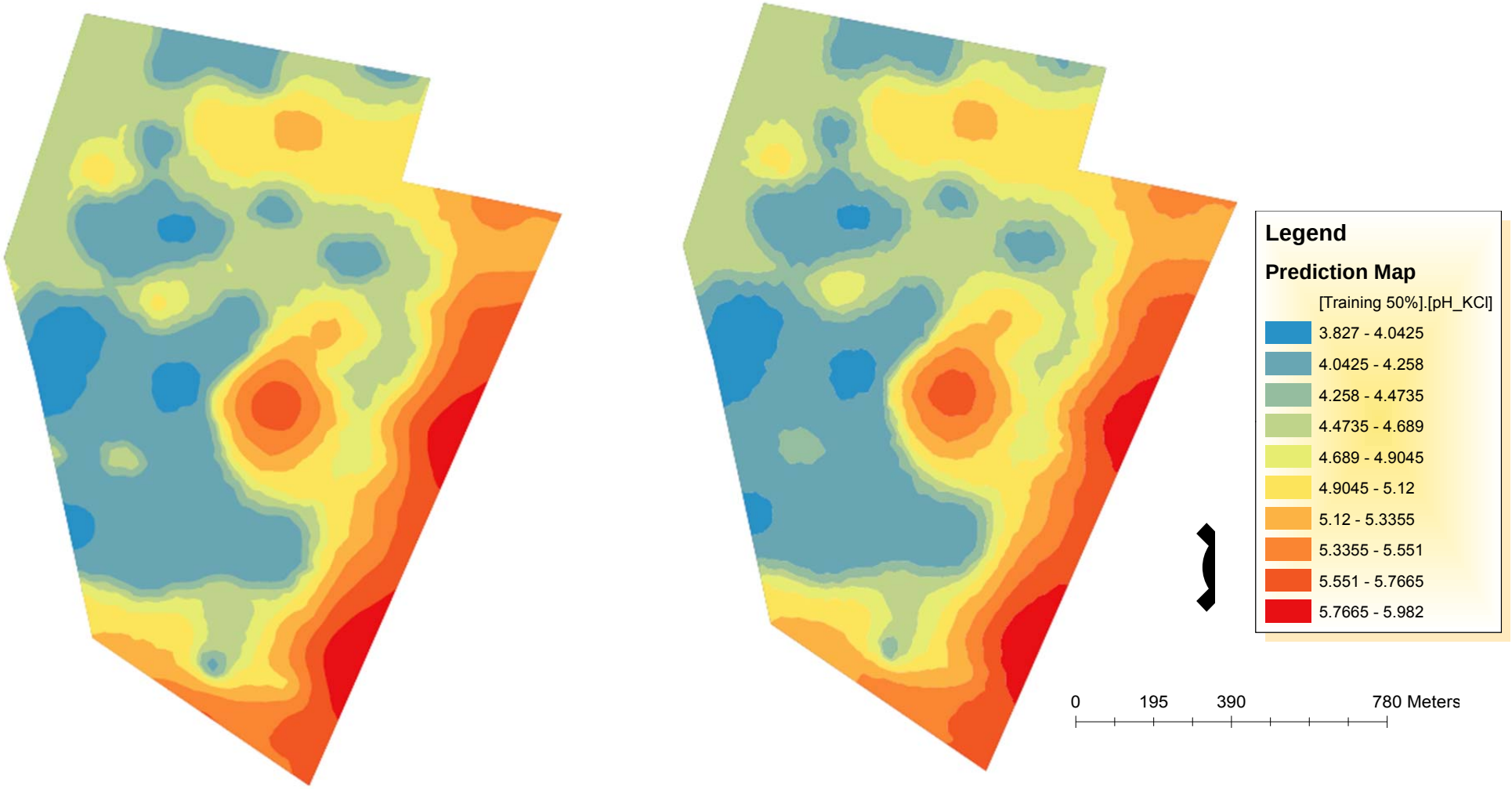
(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 60: Prediction maps of systematically selected 33% training and 67% testing (removal of every 3rd data point) data sets for the pH of Potassium Chloride (pH-KCl).



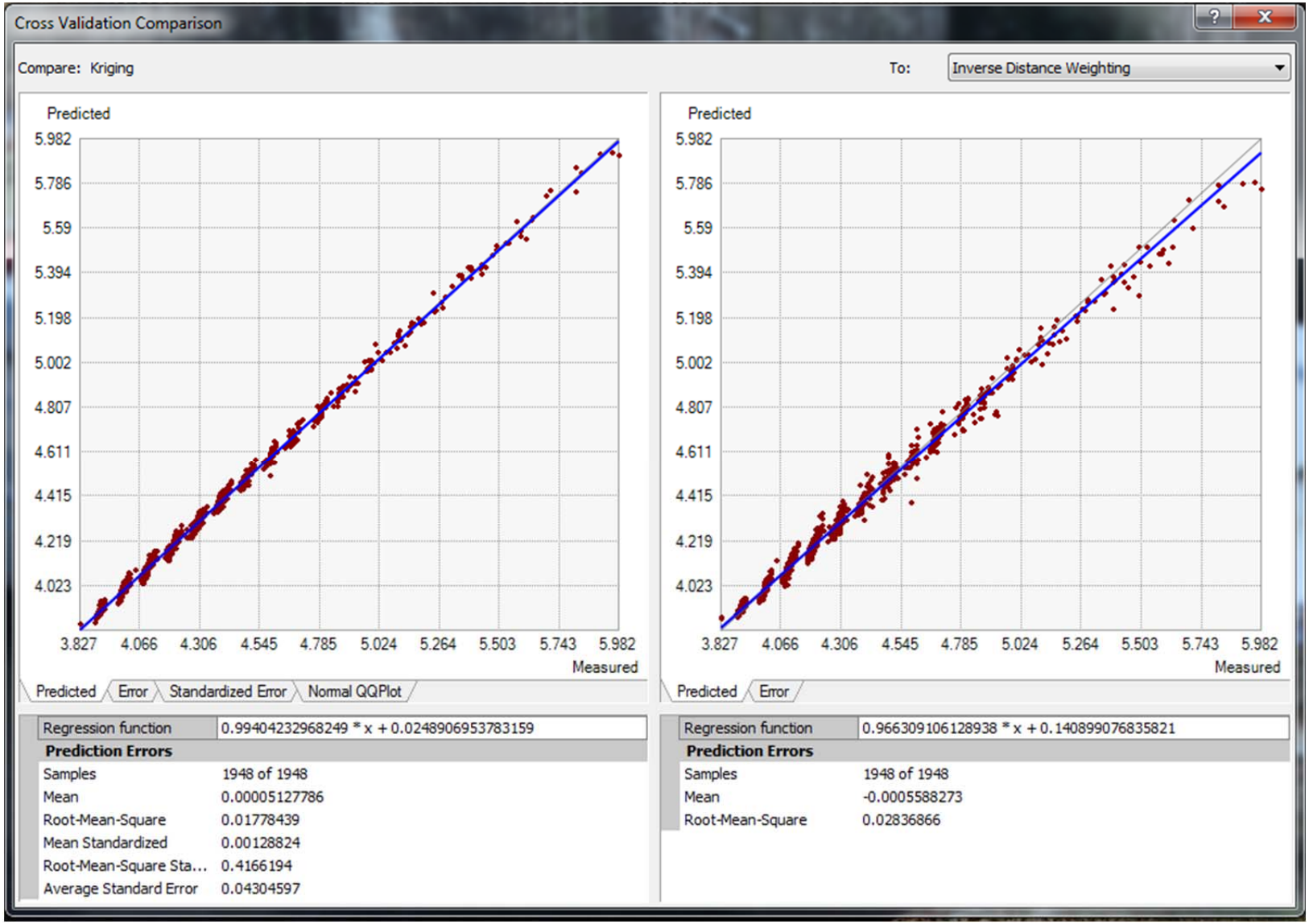
Graph 45: Cross-validation comparison of predicted error for the systematically selected 33% pH-KCl training data set.



(a) Ordinary Kriging

(b) Inverse Distance Weighting

Figure 61: Prediction maps of systematically selected 50% training and 50% testing (removal of every 2nd data point) data sets for the pH of Potassium Chloride (pH-KCl).



Graph 46: Cross-validation comparison of predicted error for the systematically selected 50% pH-KCl training data set.

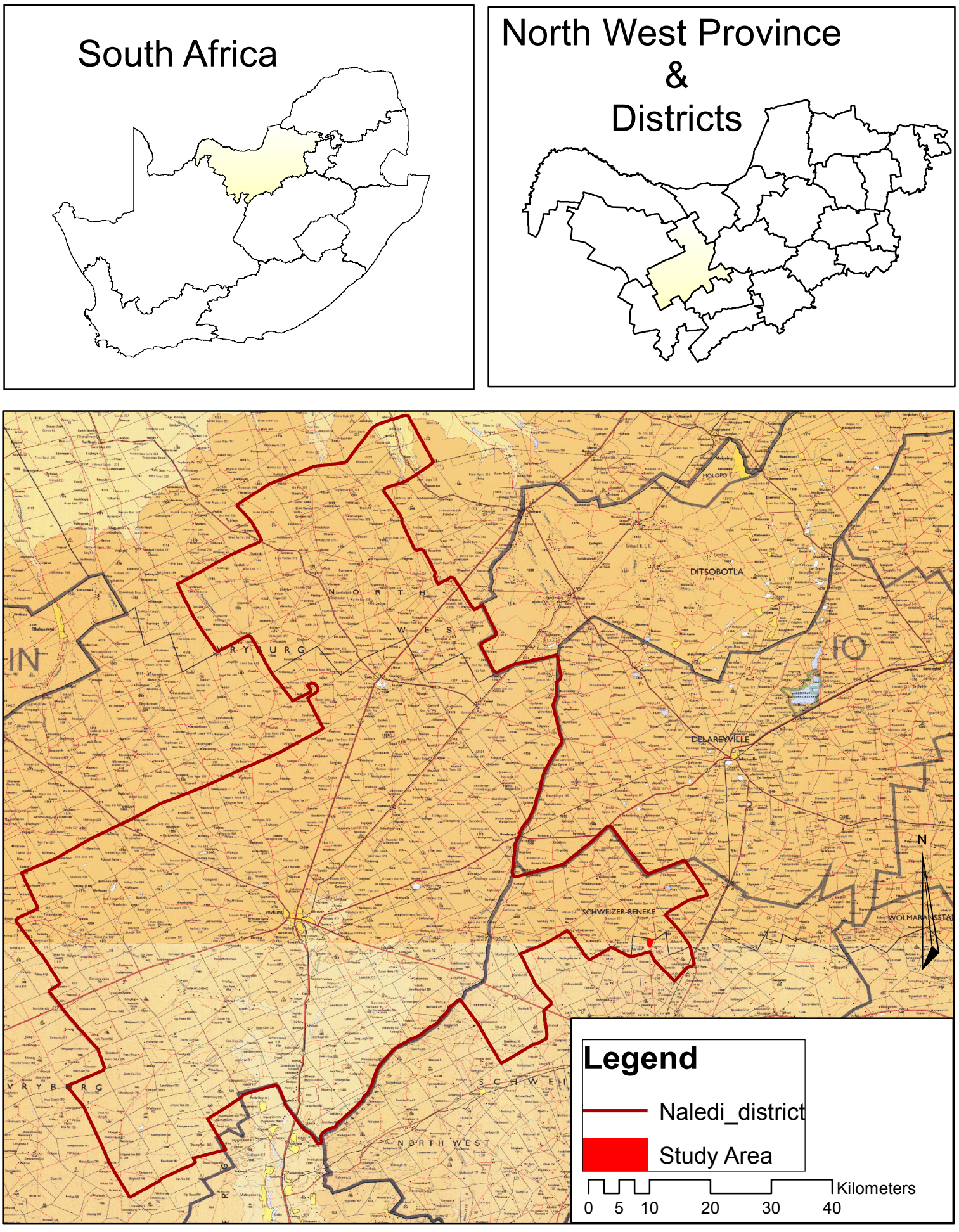


Figure 6: Map of the study location.