UNIVERSITEIT VAN AMSTERDAM

A Comparison of Interpolation Techniques for Spatial Data Prediction

Hamzeh Sheikhhasan
hsheikhh@science.uva.nl

Master's Thesis in Computer Science
Faculty of Science
Universiteit van Amsterdam
The Netherlands

Thesis Supervisors

Emiel van Loon
vanloon@uva.nl

Adam Belloum
adam@science.uva.nl

June 2006

Abstract

The analysis and interpretation of the variety of spatial interpolation and regression techniques became an important topic in the last decade. This process used to be highly human dependent with different individuals approaches, but with the increased availability of digital datasets and with the enormous software and hardware capabilities, this process became more machine dependent. There are a small number of projects that have provided a comparison and superiority of some spatial techniques over others. The objective of this thesis is to provide a comparison between eight interpolation and regression techniques through an automated model that is able to produce different species distribution and abundance maps based on information obtained from sample sites. This automated model will be run under the Grid-based Virtual Laboratory Amsterdam (VLAM-G) which provides access to geographically distributed resources. This will provide extra accessibility and will speedup the performance of the automated model.

Acknowledgments

I would like to thank my parents, my sister Nadia, and my two brothers Ahmad and Belal for all the support and encouragements they gave to me in the last two years. Without their support I wouldn't have reached this stage in my life, I really appreciate it.

I would like to thank Dr. Emiel van Loon for his continues support and for guiding me throughout the whole project. My deepest gratitude to Dr. Adam Belloum for his support and advices throughout the project. My deepest appreciations to the staff of the Computation Biology and Physical Geography department for giving me this great opportunity to be one of their team in the last nine months.

I would like also to thank Adianto Wibisono from the Computer Architecture and Parallel Systems Group for his professional help in managing and designing the workflow. My deepest appreciations to Niels Molenaar for his friendly advices and help and to the International Office of the Faculty of Science.

Finally, I would like to thank all my friends, Dutch and Internationals for the two wonderful years we had together.

List of Tables

List of Figures

Contents

Chapter 1

1 Introduction

In the last years, the analysis and interpretation of spatial datasets became an important topic in geostatistics. In the past, this process was highly human dependent and individuals would take different approaches, this lead to large distinct different solutions. Each case was dependent on the judgment and experience of individuals to select the right interpolation or regression technique.

However, the variety of spatial interpolation and regression technique, the increase availability of digital datasets, and with the increase software and hardware capabilities, the process of interpretation and analysis of spatial datasets became more machine dependent.

Although there are several projects that investigated a large number of spatial interpolation and regression techniques, a small number have provided a comparison and a superiority of some techniques over others. The expanded interest in Geographical Information Systems (GIS) with there wide usage made such a comparison important since it will investigate and show the applicability of these techniques that are embedded in some of these systems.

Nationwide maps of different species abundance and distribution are needed at a high spatial and temporal resolution. Such maps are not easily obtained and not readily available. Intensive fieldwork of thousands of volunteers is needed and can be carried out every few decades since it is not an easy process and needs a lot of preparation.

To be able to produce distribution and abundance maps for different species, information obtained from a small number of sample sites are used. The sample sites do not cover the whole entire area of interest. Several interpolation and regression techniques are available and are able to fill the gabs between these observation sites.

Many species are protected through national and international laws. Several projects concerning constructions of roads and buildings close to species distributions must be reconsidered. Other issues such as climate-change and pollution must be addressed to know their consequences on species populations. Other human-species relationship such as bird-planes collisions cause plane crashes, loss of human lives, and sometimes plane damages.

In order to address and eliminate all these issues and in order to compare the different spatial interpolation and regression techniques used, a toolbox has been created in this thesis to provide a decision support tool. This toolbox is able to produce prediction and distribution maps for different species and will assist in some projects such as Bird Avoidance Model/ Bird Avoidance System (BAMBAS).

In order to provide an extra accessibility to KansK toolbox, which is a toolbox that automates the distribution and prediction maps of different kind of species, and to make the toolbox easy to run and use by scientists, the Grid-based Virtual Laboratory Amsterdam (VLAM-G) provides a generic service for managing data and resources and performing experiments location independent. VLAM-G provides access to geographically distributed powerful resources that will speedup the performance of KansK processes.

The reset of this thesis is organized as follows:

Chapter 2 Several related work are discussed in this chapter. Different spatial interpolation and regression techniques are used in each one of them.

Chapter 3 A statistical background is given for each interpolation or regression technique used in this thesis.

Chapter 4 A description about the methodology used and the workflow design of KansK toolbox.

Chapter 5 The results are presented in this chapter, we will try to decide which is the best interpolation or regression technique to choose for the sample dataset used.

Chapter 6 Conclusions and Future Work.

Chapter 2

2 Related Work

In this chapter, four different spatial case studies are presented. Each case study is investigating a wide variety of spatial interpolation and regression techniques. Some of these techniques are the same as the spatial interpolation techniques we presented in chapter 3. In the case studies, different spatial datasets are used: breeding-birds, rainfall, and temperatures. Breeding-birds dataset is similar to the dataset we used in this thesis which was provided by SOVON – the Dutch Center for Field Ornithology. In each section we are providing a background about each project, spatial interpolation or regression techniques used, and finally the results.

2.1 Bird Avoidance Model/ Bird Avoidance System

Bird Avoidance Model (BAM) project was conducted between 2002-2005 as a joint project between Computational Biology and Physical Geography department at the University of Amsterdam, SOVON – the Dutch Centre for Field Ornithology and the Royal Netherlands Air Force. BAM is used by the Royal Netherlands Air Force as a decision support tool to try to eliminate the risk of bird-aircraft collisions.

Nationwide maps of bird abundance are needed at a high spatial and temporal resolution, for a BAM. Such maps are however not readily available [11]. Intensive fieldwork of thousands of volunteers is needed but can be carried out every few decades because it needs a lot of preparation and collaboration. To be able to produce distribution and abundance maps of birds, information obtained from sample sites, which do not cover the entire area of interest, are used. A technique known as regression-kriging has subsequently been used to fill the gaps between observations sites in space and time.

2.1.1 Test Data

Monitoring of Dutch bird population is in majority part of a governmental monitoring scheme which includes other organisms as well [12]. Fieldwork and data processing is conducted by non-governmental organisations such as SOVON, Statistics Netherlands

and around 3000 volunteers and small number of ornithologists. **Table 2.1** lists some SOVON bird data sources.

Bird datasets from several fieldwork projects have been used to produce the BAM-maps, see **Table 2.1**. One of these datasets is the Point Transect Count. Terrestrial wintering birds are being monitored since 1980 along about 400 transects with 20 observation points. Observers counted all species at each observation point during exactly 5 minutes.

The spatial distribution module includes spatial density maps of the 62 bird species selected as most relevant for flight safety in the Netherlands in bi-weekly intervals, 4 time periods per day and at five altitude layers [11], sample data for the Common Buzzard was used in this project for the production of the density maps.

Table 2.1: SOVON bird data sources

Non-breeding bird data			
Count description	**Abbreviation**	**Months/years**	**Numbers**
Winter bird counts	PTT	Feb. 1993-1997 Aug. 1988-1992 Nov. 1992-1996 Dec. 2000-2004	Observed individuals
Water bird counts	WAV	Monthly 1999-2004	Observed individuals
Casual observations	BSP	Monthly 2000-2005	Observed individuals
Breeding bird data			
Breeding bird counts	BMP	2000-2004	Pairs
Colony Breeders	LSB	2000-2004	Pairs
Rare breeding birds	LSB	2000-2004	Pairs

2.1.2 Statistical modelling

General Additive Model GAM which is a regression model was built for different years for the number of the Common Buzzards. The Poisson distribution with the log link was the appropriate choice for modelling the procedure because the number of Buzzards per point is always positive, many points contains zeros and finally the variance increases while the abundance increases. The best fitting model predictions and standard errors were calculated at 1 km^2 resolution and the predictions were mapped using ArcGIS Geographical Information System.

The difference between the counts and the prediction was calculated. The residuals were spatially interpolated using Ordinary Kriging and producing a 1 km ^{2} residuals map that covers the entire Netherlands. This technique is called Regression Kriging and has two separate steps (GAMs, Kriging). Maps of Buzzard densities are obtained from the addition of the estimated trend from the GAM and the predicted values of the residual. Figure 2.1 shows the final map of the distribution of the Common Buzzards in December 2000 as a result of regression Kriging as mentioned by Sierdsema and van Loon in [12].

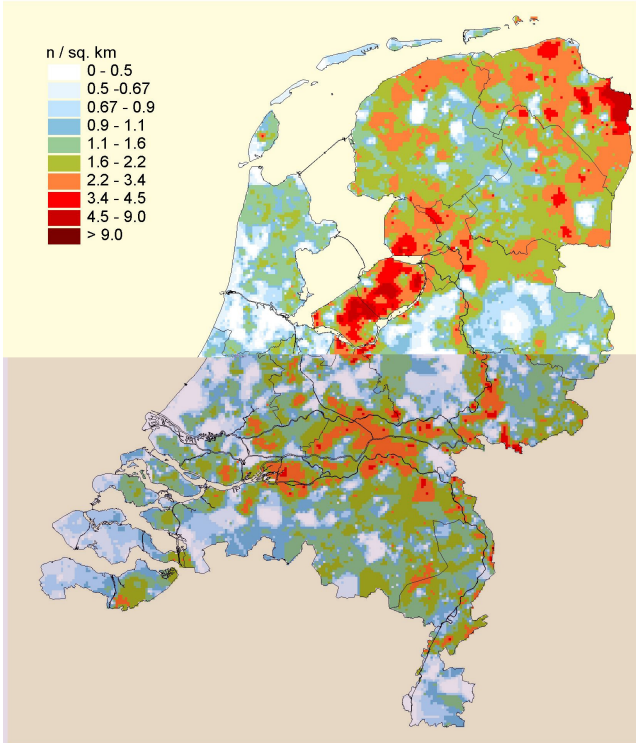


Figure 2.1: The Distribution of the Common Buzzards in December 2000 as a result of regression kriging

2.2 Spatial interpolation techniques for Rainfall Estimation

The analysis and interpolation of spatial data is important and highly human dependent. It is well known that different individuals will take different approaches, yielding a large assortment of distinct solutions [13]. Spatial interpolation technique is being chosen for each case based on the judgement and experience of different individuals

Estimating rainfall at different locations based on meteorological observations results encouraged the development of gridded estimates of rainfall as inputs for spatially distributed hydrologic and management models. Geographic Information Systems (GIS) provide ready-to-use spatial interpolation techniques that need to be investigated through such models.

2.2.1 Spatial interpolation techniques

Twelve interpolation techniques are used to estimate the missing gages values based on the remaining ones. These techniques are available in ArcView GIS software and are ranked based on the best minimum average to the worst maximum average and to other statistics. Figure 2.2 summarizes the twelve spatial interpolation techniques as a result of this study presented by Naoum and Tsanis [13].

- **Spline (Regularized & Tension):** is a form of interpolation where the interpolant is a special type of piecewise polynomial called spline. Spline interpolation is preferred over polynomial interpolation because the interpolation error can be made small when using low degree polynomials for the spline [14].

- **Inverse Distance Weighted (IDW):** is a simple technique for curve fitting, a process of assigning values to unknown points by using values from known points [14].

- **Kriging:** Kriging, as a form of generalized liner regression technique, is used to estimate the value of a property at un-sampled location by referring to neighbouring locations.

- **Trend Surface:** trend creates a floating-point grid by using polynomial regression to fit a least-square surface to the input points. Users are allowed to control the order of the polynomial.

- **Theissen Polygons:** this approach is appropriate when we want to define the region of influence. It is based on the nearest neighbours to a line or a point. For a series of points, the region of influence is represented by a set of polygons called Theissen Polygons. These polygons are the most common approach to model spatial distribution rainfall. The approach is based on defining the area closer to the gage then any alternate gage and the assumption that the best estimate of rainfall on that area is represented by the point measurement at the gage [13]. Using Theissen Polygons, discontinues surfaces were developed defining the rainfall depth at the desired area.

2.2.2 Test Data

This model was applied to a group of rain-gages in Switzerland. The dataset used was taken around the period of Chernobyl Nuclear Power Plant accident in (26^{th} of April 1986). Right after the accident a radioactive plume crossed many European countries causing a radioactive deposition on the ground because of the rainfall. The air pollution Group at Imperial College London prepared the dataset. The dataset was made in 8^{th} of May 1986 and included 467 records of daily rainfall.

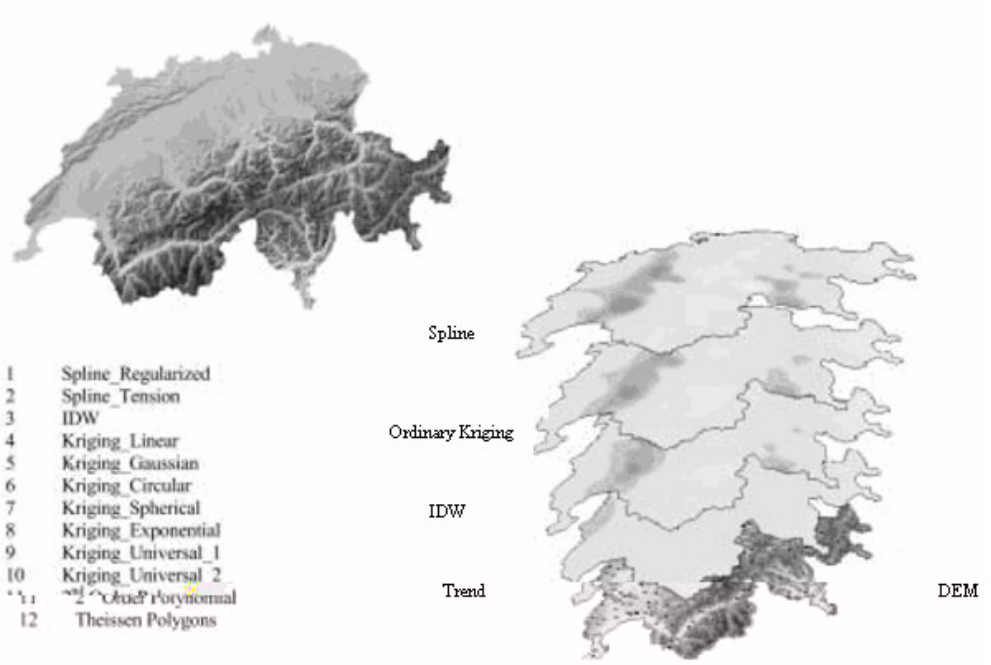


Figure 2.2: Digital elevation model and interpolated rain surfaces in Switzerland

2.2.3 Results

Based on the multiple random selections of gages to eliminate any errors or outliers, no statistical data preparation was done. **Table 2.2** concludes the results of the twelve techniques. Kriging (Exponential and Universal_1) and Inverse Distance are recommended.

Table 2.2: Spatial interpolation techniques results for rainfall

Interpolation Technique	**Type**	**Results**
Spline	Regularized	Poor performance
	Tension	Reliable estimates
Trend Surface	2nd order polynomial	Poor performance
Theissen Polygons	-	Vary from one case to the other
Kriging	Linear	Vary from one case to the other
	Gaussian	
	Circular	
	Universal_2	
	Spherical	Reliable estimates
	Exponential	
	Universal_1	
IDW	-	Reliable estimates

2.3 Spatial interpolation techniques for Temperature Estimation

Landscape scale models such as regeneration, growth, and mortality of forest ecosystems depend on their performance about the accurate estimation of temperature. Temperature prediction at un-sampled sites is of interest to individuals involved in fire management, resource management, and spraying or seeding operations [15]. Scientists are also interested in accurate temperature to study greenhouse effect and global warming. Depending on the data spatial attributes, accuracies among spatial interpolation techniques vary. The choice of spatial interpolator is important especially in mountain regions where variables may change over short scale.

2.3.1 Spatial interpolation techniques

Eight interpolation techniques were used to study two regions in western and eastern North America. Region 1 has stations spread all over the study area, Region 2 has stations surrounding population areas. These techniques are: Inverse Distance Weighted, Spline, Trend Surface, and Kriging were already discussed in **section 2.2.1**, the other four interpolation techniques are explained below:

- **Optimal Inverse Distance Weighted:** is a form of inverse distance weighted where the power parameter is equal to the minimum mean absolute error.

- **Polynomial Regression:** the variable of interest, which is weather, is being fitted to some linear combination of regressor variables (weather station's X, Y, and Z coordinates). There is a chance of increasing multicollinearity because of adding regressor variables, this might decrease the model ability to predict outside the rounded hull of data points. Temperature was fitted to first, second, and third order polynomial models of the X and Y coordinates plus elevation [15].

- **Lapse Rate:** usually temperatures decreases while elevation increases. This relationship for a region is used to predict temperatures at un-sampled sites by using the temperature values of the nearest weather station and the difference in elevation.

- **Cokriging:** Cokriging is an extension of kriging as mentioned in **section 2.2.1**, except it is more intensely sampled. Cokriging estimates a variable from the observations of that variable and the values of related variables at nearby sampling locations.

2.3.2 Results

Interpolation techniques were compared based on the basis of bias, mean absolute error (MAE), and mean squared error (MSE). Other factors where investigated such as the effect of data variance, data correlation with elevation, and lapse rate on MAE. Each technique was repeated several times to obtain the cross validation statistics.

Polynomial regression had the preferred performance and the lowest MAE value among the different techniques ranked. Based on higher correlation between elevation and temperature, Polynomial regression and lapse rate technique gave the preferred performance. Inverse distance weighted, optimal inverse distance, and kriging showed a similar robustness to apriori data range, correlation (between elevation and temperature), and variance [15].

Kriging performed better than optimal inverse distance when the data were anisotropic. However, when data were isotropic, optimal inverse distance was better. **Table 2.3** concludes the results of the eight techniques:

2.4 Minimum Habitat Requirements of Forest-Breeding Birds

Loss of habitat because of human-induced activities is the only threat to the survival of many species and to global biodiversity. Many different organisms are affected in different regions such as amphibians, beetles, butterflies, and small mammals. There is a special concern about the effects of habitat loss on forest birds that breed in the eastern United States and Canada and winter in the Neotopics [16].

This study is concerned about the habitat loss and the species survival threats. Biologists are trying to understand the effect of habitat loss by predicting the minimum amount of habitat necessary for population survival. Life-history knowledge of these species is important since it gives more information about species more sensitive to habitat loss than others.

2.4.1 Test Data

This model was applied to 41 species of forest-breeding birds. The dataset used was taken from the North American Breeding Bird Survey to estimate the ‘proportion presence’ of each of 41 forest bird species over a 10-year window [16]. To calculate the forest percentage covered of the 779 landscapes in central and eastern USA, U.S. Geological Survey (USGS) Land Use and Land Cover (LULC) digital data are used.

Table 2.3: Spatial interpolation techniques results for temperature estimation

Interpolation Technique	Results	Comments
Inverse Distance Weighted	Poor performance.	Results are questionable and temperature peaks because of discontinuities at station locations.
Optimal Inverse Distance Weighted	Better performance than Inverse Distance Weighted and Kriging.	The most preferred technique when the data are not correlated and isotropic.
Trend Surface	Poorest performance.	Broad regional trends because of the bias introduced by multicollinearity.
Polynomial Regression	Preferred performance.	Technique is recommended when the correlations between temperature and elevation are not low.
Spline	Poor performance.	Interpolated values are outside the observed data range. Cubic Spline is not recommended for irregularly-spaced data.
Kriging	Better performance than Inverse Distance Weighted.	When data are anisotropic Kriging is better than Optimal Inverse Distance.
Cokriging	Poor performance.	When temperature and elevation are not correlated, Cokriging is similar to Kriging. This is not expected because elevation component is not significant in Cokriging.
Lapse Rate	Reasonable performance.	It is preferable over Cokriging based on visual plausibility and adherence. When elevation and temperature are not correlated then the technique is degraded into a nearest neighbour technique.

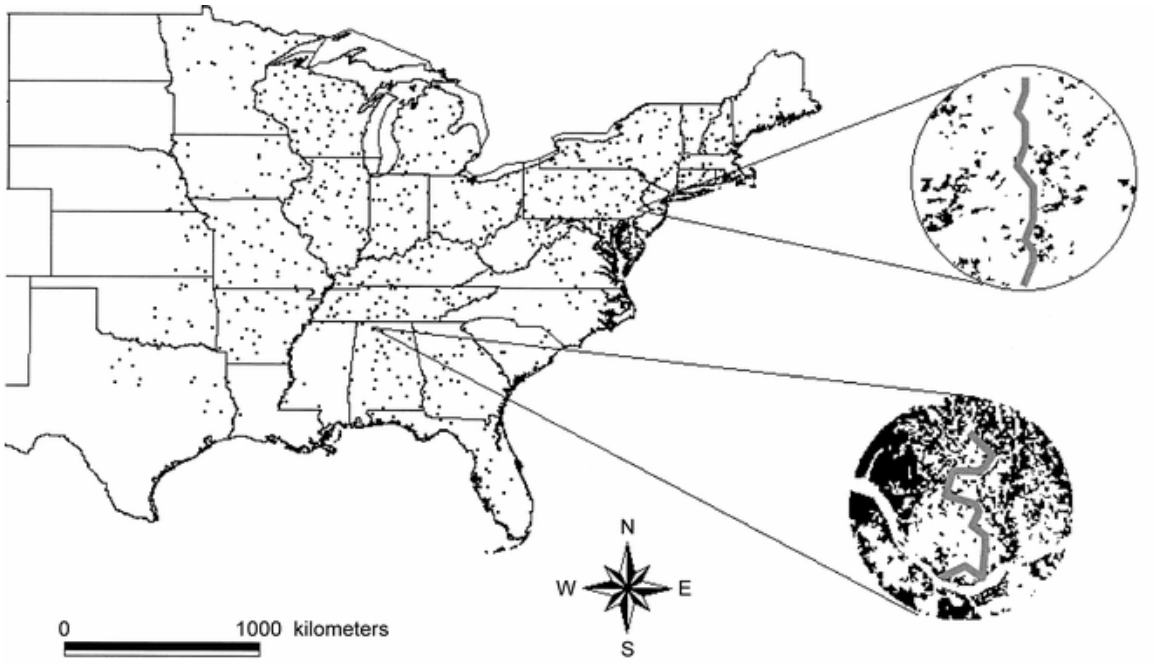


Figure 2.3: Study area of 779 landscapes in central and eastern United States

Figure 2.3 (Vance et al. [16]) shows 779 landscapes, each one of them has a radius of 19.7 km and is centred in a breeding bird survey route. Zigzag line is the survey route, each route is 39.4 km with a total of 50 stops, each stop is 3 minutes and conducted at 0.8 km interval. All birds heard or seen by the observers at 0.4 km radius are counted. Black areas are forests, and white areas are other landscapes.

2.4.2 Data Analysis and Results

Locally weighted regression, or *Loess*, is used to smooth the dependent variables in a moving window fashion by fitting a local regression that was weighted by the distance of the data points within a specified neighbourhood from a point x on the independent axis [16]. Points close to x have large weights and points far from x have smaller weights as explained by Cleveland in [6]. Loess curves were fit using default parameters from SAS, local linear multiple regression, and normal weight function. This model provides good compromise between data goodness of fit and Loess curves smoothness.

Smoothed regression curves are used to estimate the minimum habitat for each of the 41 forest bird species. Species have a 50% probability of presence in the landscape which is considered a tendency of occupancy. Figure 2.4 (Vance et al. [16]) demonstrates the estimation of minimum habitat required using Loess smoothed data. Figure 2.4(a) presents the 'normal' Loess curve where the minimum habitat amount of 50% presence can be predicted directly from the map which is 26.5% for this species.

Figure 2.4(b) represents one kind of species that is always had a presence of less than 50% regardless of the presence of other habitats. Figure 2.4(c) represents the opposite situation where the species never reached the 50% presence with a minimum habitat of 99%. Figure 2.4(d) represents the situation of more than one habitat, the minimum habitat amount needed here is 53% in which the species have a chance of 50% probability of presence.

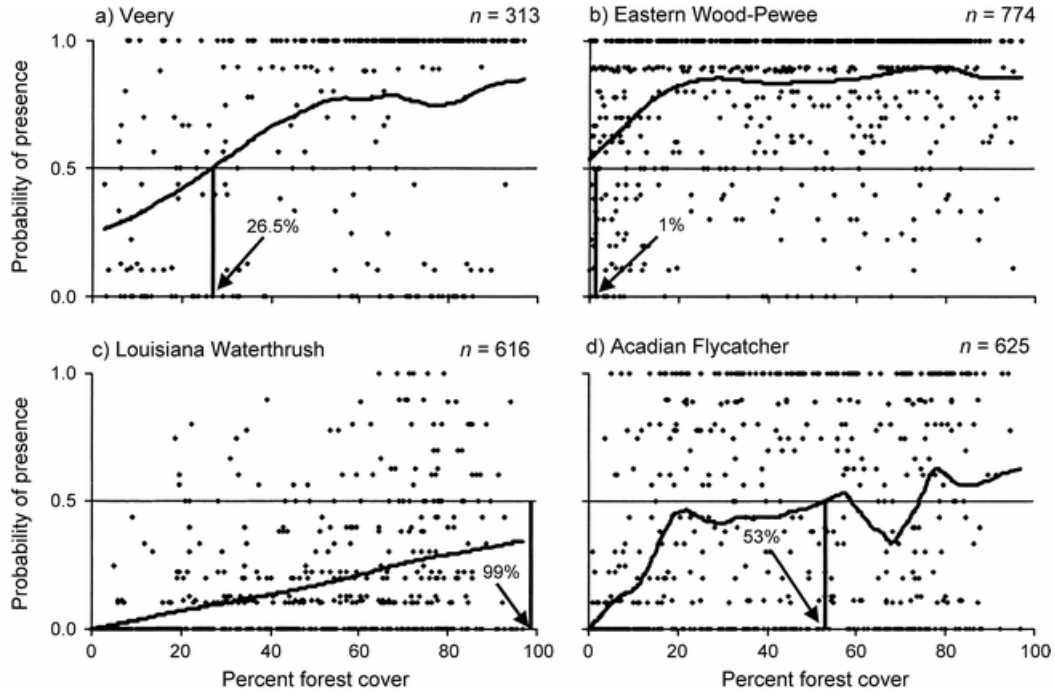


Figure 2.4: Results of different Loess curves

2.5 Summary

This chapter presented several spatial interpolation techniques used in different case studies. BAM project recommended using a combination of Poisson and regression kriging on the basis of exploratory analyses and expert knowledge. This combination is the appropriate choice since count data often follows a Poisson distribution.

In the rainfall estimation case, it appeared that increasing the number of gages available for interpolation enhanced the performance of several techniques especially inverse distance weighted where it is known to provide good results for dense networks such as this study. Ordinary kriging (exponential and universal_1) showed consistent performance and provided reliable estimates regardless of the number of gages or the cell size used in the interpolation [13].

The temperatures estimation study recommended using polynomial regression, kriging, and optimal inverse distance weighted. For these techniques, the results indicated that they have small temperatures variances and temperatures ranges which tend to decrease the interpolator Mean Absolute Error (MAE) and they also have high correlation between the temperature and elevation. These two factors had a strong influence on the predictor performance.

Finally, the minimum habitat study used locally weighted regression to smooth the data since with nonlinear regression it is not necessary to choose a model in-advance, and this allows the data to estimate the best regression surface.

The different spatial cases described in this chapter used different interpolation techniques and different datasets. Three cases recommended using kriging interpolation; two cases recommended using inverse distance weighted interpolation. Polynomial regression and locally weighted regression were recommended to be used by two different cases. From this chapter, we expect that kriging, inverse distance weighted, polynomial, and locally weighted regression will most probably give a reasonable performance in KansK toolbox.

In the next chapter, the statistical background of eight spatial interpolation and regression techniques are presented. These techniques are similar to some of the techniques presented in this chapter.

Chapter 3

3 Statistical Background

In this chapter we will go through the statistical background of each interpolation and regression techniques that are investigated in this thesis.

3.1 Akima Linear interpolation (LIN)

A method of bivariate interpolation and smooth surface fitting for z values at irregularly distributed data points in the x-y plane was presented by Akima in 1978 [1]. Akima's design of the method assumes that the resulting surface will pass through all the irregularly spaced data points.

The method first triangulates the x-y using the Max-Min angle triangulation suggested by Lawson in 1972 [2]. This triangulation works as follows. For each quadrilateral consisting of a four-point set (**p1**, **p2**, **p3**, **p4**) with each internal angle smaller than π, two triangles are created from the partitioning of the quadrilateral based on the choice that maximizes the minimum interior angles or the choice of the shorter diagonal. New points are added in this way forming new triangles, each point, **p5**, is connected to the closest pair of points and lies outside the quadrilateral. Figure 3.1 is an example that illustrates the Max-Min angle triangulation.

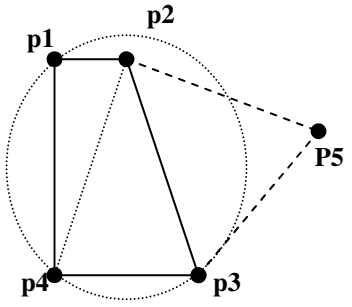


Figure 3.1: Max-Min angle triangulation

A fifth-degree bivariate polynomial in x and y is used to interpolate the z values in a triangle at any point (x, y):

$$z(x,y)=\sum_{i=0}^{5}\sum_{j=0}^{5-i}\mathbf{q}_{ij}\,\mathbf{x}^{i}\mathbf{y}^{j} \qquad \textbf{(3.1)}$$

To determine the 21 coefficients $\mathbf{q}_{ij}$ [3], the values of **(eq. 3.1)** and its first-order and second-order partial derivatives are provided at each vertex of the triangle at points (**p1**, **p2, p4**), this yields to 18 independent conditions:

- The z values:

  z1, z2, and **z3** ;

- The first order partial derivatives:

$$\left.\frac{\partial z}{\partial x}\right|_{p=p1},\left.\frac{\partial z}{\partial x}\right|_{p=p2},\left.\frac{\partial z}{\partial x}\right|_{p=p3},\left.\frac{\partial z}{\partial y}\right|_{p=p1},\left.\frac{\partial z}{\partial y}\right|_{p=p2} and \left.\frac{\partial z}{\partial y}\right|_{p=p3};$$

- The second order partial derivatives:

$$\left.\frac{\partial^2 z}{\partial x^2}\right|_{p=p1},\left.\frac{\partial^2 z}{\partial x^2}\right|_{p=p2},\left.\frac{\partial^2 z}{\partial x^2}\right|_{p=p3},\left.\frac{\partial^2 z}{\partial y^2}\right|_{p=p1},\left.\frac{\partial^2 z}{\partial y^2}\right|_{p=p2},\left.\frac{\partial^2 z}{\partial y^2}\right|_{p=p3},\left.\frac{\partial^2 z}{\partial xy}\right|_{p=p1},\left.\frac{\partial^2 z}{\partial xy}\right|_{p=p2} and \left.\frac{\partial^2 z}{\partial xy}\right|_{p=p3};$$

- The partial derivative of the function differentiated in the normal direction to each side, $\mathbf{n}_{s}$, of the triangle [3]. This is a third-degree polynomial, at most, in the variable measured in the direction of the side of the triangle [1]. Additional three independent conditions are added:

$$\left.\frac{\partial z}{\partial n}\right|_{n_{s1}},\left.\frac{\partial z}{\partial n}\right|_{n_{s2}} and \left.\frac{\partial z}{\partial n}\right|_{n_{s3}}.$$

Partial derivatives are evaluated for every data point $\mathbf{p}_k$ by determining the vector normal to the surface of each data point, this can be performed by calculating the vector product of $\mathbf{p}_k = (\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k)$ and two close neighbors $\mathbf{p}i$ and $\mathbf{p}j$:

$$\mathbf{dz}_{ikj} = (x_i - x_k)(y_j - y_k) - (y_i - y_k)(x_j - x_k) \quad \textbf{(3.2)}$$

$$\mathbf{dx}_{ikj} = (y_i - y_k)(z_j - z_k) - (z_i - z_k)(y_j - y_k) \quad \textbf{(3.3)}$$

$$\mathbf{dy}_{ikj} = (z_i - z_k)(x_j - x_k) - (x_i - x_k)(z_j - z_k) \quad \textbf{(3.4)}$$

where $\mathbf{dz}_{ikj}$, $\mathbf{dx}_{ikj}$ and $\mathbf{dy}_{ikj}$ are the three components of the vector product and must be positive to assure that the vector normal to the surface is also positive [3].

Gradient vectors in the direction of x and y are evaluated through the following equations:

$$\left.\frac{\partial z}{\partial x}\right|_{p=pk} = \sum_{i=0}^{nc-2}\sum_{j=j+1}^{nc-1}\mathbf{dx}_{ijk} \Big/ \sum_{i=0}^{nc-2}\sum_{j=j+1}^{nc-1}\mathbf{dz}_{ijk} \quad \textbf{(3.5)}$$

$$\left.\frac{\partial z}{\partial y}\right|_{p=pk} = \sum_{i=0}^{nc-2}\sum_{j=j+1}^{nc-1}\mathbf{dy}_{ijk} \Big/ \sum_{i=0}^{nc-2}\sum_{j=j+1}^{nc-1}\mathbf{dz}_{ijk} \quad \textbf{(3.6)}$$

Where **nc** is the closest neighbors. To calculate the second order partial derivatives, the same approach can be used.

3.2 Inverse Distance Weighted (IDW) interpolation

Another technique that is frequently used to interpolate scattered points is Inverse Distance Weighted interpolation (IDW). IDW assumes that a point to estimate is influenced most by nearby points, hence each observed point has an associated weight that is inversely proportional to the distance to the point to be estimated. Figure 3.2 illustrates the IDW interpolation, scatter points (**p1, p2, p3, p4, p5**) are within the search radius of the estimated point **z**.

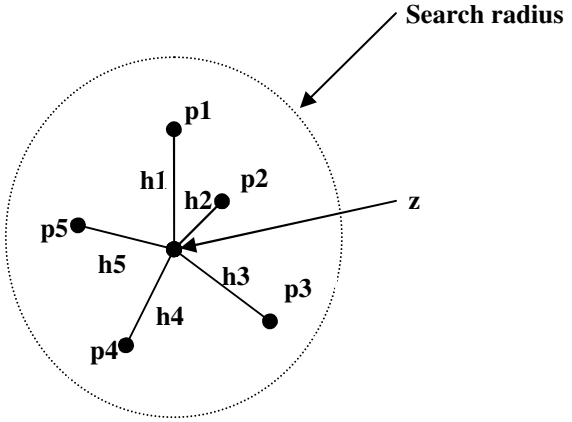


Figure 3.2: IDW interpolation

Inverse distance weighted interpolation method was first presented in its simple form by Donald Shepard in 1968 [4]:

$$z(x, y) = \sum_{i=1}^{n} w_i f_i \tag{3.7}$$

where **n** is the total number of observations, f_i are the observed values, and w_i is the weighted associated with each observation point, relative to an observation at (x, y). Weights are calculated using the following weight function:

$$\mathbf{w}_i = \frac{h_i^{-p}}{\sum_{i=1}^{n} h_i^{-p}} \tag{3.8}$$

where p is the power parameter that defines the rate of the reduction of the weights as distance increases [5], h_i is the distance between the observation point and the point to be estimated:

$$\mathbf{h}_i = \sqrt{(x - x_i)^2 + (y - y_i)^2} \tag{3.9}$$

where (x, y) and (x_i, y_i) are the coordinates of the interpolation point and the scatter point. The weight function **(eq. 3.8)** approaches zero when the distance from the scatter points increases.

3.3 Locally Weighted Regression (LWR)

A technique used to smooth a scatterplot (x_i, y_i), where $i = 1, \ldots, n$, is Locally Weighted Regression (LWR), or *loess*. The observed value z_i at (x_k, y_k) is the value of a polynomial fit to the data using weighed least squares, where the weight of (x_i, y_i) is large if (x_i, y_i) is close to (x_k, y_k) and small if it is not [6], The smoothing procedure has been designed to accommodate data for which

$$z_i = B_0 - B_1 x_k y_k - B_2 x - B_3 y - B_4 x^2 - B_5 y^2 + \varepsilon_i \tag{3.10}$$

where $B_i(x_i, y_i)$ are the estimates of the parameters in a polynomial regression of degree d and the ε_i are random variables with mean 0 and constant scale. Neighboring points of (x_i, y_i) are used to form the observed value z_i, the weights $w_k(x_i, y_i)$ decrease as the distance between (x_k, y_k) and (x_i, y_i) increases [6]. Figure 3.3 is an example that illustrates the LWR.

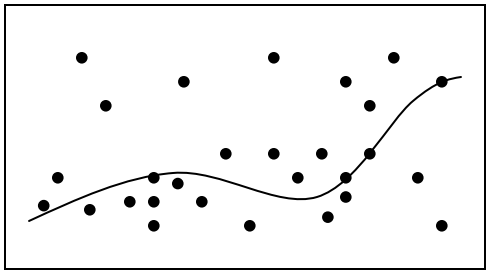


Figure 3.3: Scatterplot of robust smoothed data

Locally weighted regression and robust locally weighted regression are estimated by following these operations:

- Compute the estimate $\hat{B}_j(x_i, y_i)$, j = 1, …, d, of the parameters in a polynomial regression of degree d as mentioned by Cleveland [6]. The estimate $\hat{B}_j(x_i, y_i)$ are the values of the B_j that minimizes:

$$\sum_{k=1}^{n} w_k(x_i, y_i)\left(z_k - B_0 - B_1 x_k y_k - B_2 x - B_3 y - B_4 x^2 - B_5 y^2\right)^2 \qquad \textbf{(3.11)}$$

  z is the observed value and $\left(B_0 - B_1 x_k y_k - B_2 x - B_3 y - B_4 x^2 - B_5 y^2\right)$ are the fitted values of the locally weighted regression at (x_i, y_i).

- Let Q be the bisquare weight function and $e_i = z_k - z_i$ be the residuals from the current observed values. Let s be the median of $|e_i|$ and the robustness weights:

$$\delta_k = Q(e_k / 6s) \qquad \textbf{(3.12)}$$

- Compute new z_i by fitting a dth degree polynomial with weight $\delta_k w_k(x_i, y_i)$ at (x_k, y_k). This Step and the step of **(eq. 3.12)** will be repeated and the final z_i are robust locally weighted regression observed values.

The weight function ***W*** is used to define weights for all (x_k, y_k), where $k = 1, \ldots, n$:

$$w_k(x_i, y_i) = W\left(h_i^{-1}\left((x_i - x_k)^2 + (y_i - y_k)^2\right)^{1/2}\right) \quad \textbf{(3.13)}$$

where h_i is the distance beyond which $W(x) = 0$.

3.4 Ordinary Kriging (OK)

Kriging named after the South African mining engineer D. G. Krige. It was designed originally to accurately predict ore reserves as mentioned by Davis [7]. Kriging, as a form of generalized liner regression, is used to estimate the value of a property at un-sampled

(mp niKedipb
1 w

surface when reaching the *Range* which is the maximum neighbourhood over which observation points are selected to estimate the estimation point.

The calculation of all possible combination of different pairs at different distances will be performed. The initial distance used is called *Lag* and will increase by the same amount though the dataset. Every point is compared to all other points to determine the variance between the points that have the same *Lag* distance and their geographical orientation. This process is repeated until all distance possibilities are analyzed.

In Figure 3.4, each small circle represents a pair of points. The big circles are the averages obtained from the ranges. Using Ordinary Kriging the semivariogram needs to be reduced to a mathematical function so that it will be evaluated at any distance and this is implemented by the solid black line.

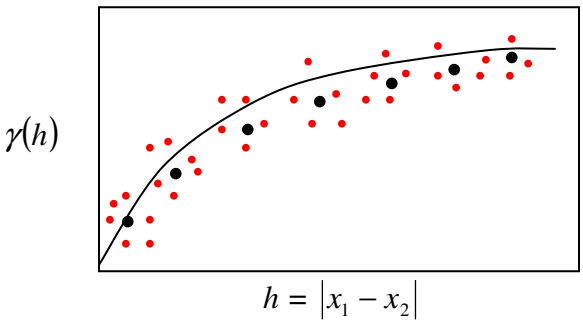


Figure 3.4: Semivariogram

W and **B** are estimated from the spatial covariance function of the semivariogram model γ, and Λ is the weights after inserting the Lagrange Multiplier μ which increases the number of unknown coefficients to be estimated, s is the observed point $(\mathbf{x}_k, y_k)$.

$$\Lambda = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \\ \mu \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \gamma(\mathbf{s}_1,\mathbf{s}_1) & \gamma(\mathbf{s}_1,\mathbf{s}_2) & \cdots & \gamma(\mathbf{s}_1,\mathbf{s}_k) & 1 \\ \gamma(\mathbf{s}_2,\mathbf{s}_1) & \gamma(\mathbf{s}_2,\mathbf{s}_1) & \cdots & \gamma(\mathbf{s}_2,\mathbf{s}_k) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{s}_k,\mathbf{s}_1) & \gamma(\mathbf{s}_k,\mathbf{s}_2) & \cdots & \gamma(\mathbf{s}_k,\mathbf{s}_k) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \gamma(\mathbf{s}_0,\mathbf{s}_1) \\ \gamma(\mathbf{s}_0,\mathbf{s}_2) \\ \vdots \\ \gamma(\mathbf{s}_0,\mathbf{s}_k) \\ 1 \end{bmatrix}$$

Ordinary Kriging weights are estimated using the following matrix equation:

$$\Lambda = W^{-1}\mathrm{B} \tag{3.17}$$

A vector **Y** of the *k* observations around locations x_0 is needed to estimate the $\hat{z}(x_0)$ of the regionalized variable desired:

$$\mathbf{Y} = \begin{bmatrix} z(\mathbf{s}_1) \\ z(\mathbf{s}_2) \\ \vdots \\ z(\mathbf{s}_k) \\ 1 \end{bmatrix}$$

The Ordinary Kriging estimate of the estimated value at location x_0 [7]:

$$\hat{z}(x_0) = \mathrm{Y}^{'}\Lambda \qquad \textbf{(3.18)}$$

3.5 Generalized Linear Models (GLM)

Generalized Linear Models (GLM) extend the classical linear regression model to data that have a non-normal distributed residuals. GLM supports a wide variety of explanatory variables next to (x, y) coordinates such as vector, matrices, and lists. The linear model assumes that the components of **Y** are independent Normal variables with the constant variance σ^2 and

$$E(Y) = \mu \text{ where } \mu = XB \qquad \textbf{(3.19)}$$

where the *Bs* are the unknown parameters that need to be estimated from the data [8]. The generalization of the linear model by rearranging **(eq. 3.19)** will produce the following three parts:

- **The random component:** each component of **Y** has independent Normal distribution and constant variance σ^2.
- **The systematic component:** a linear predictor η, which is a linear sum of the effects of explanatory variables x_j, is produced:

$$\eta = \sum_{1}^{p} x_j B_j \qquad \textbf{(3.20)}$$

Where *x* are the values of the different *p* values.

- **The link function:** is the relationship between the random and systematic components. The link function $g(\bullet)$, that relates the mean value of **Y** to its linear predictor [8]:

$$\eta = g(\mu) \tag{3.21}$$

Each component of **Y** has a distribution that is a member of the exponential family and has the following form:

$$f_Y(y,\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{3.22}$$

where functions $a(\bullet)$, $b(\bullet)$, and $c(\bullet)$ will depend in ϕ, if it is known then this is an exponential–family model with the parameter of interest θ which is called the canonical link function. If ϕ is unknown this may leads to a two-parameter exponential family.

The exponential family contains a large number of useful distributions, for example among continuous distributions there are Normal, inverse-Gaussian, and Gamma and among discrete distributions there are Binomial, Poisson and negative Binomial. **Table 3.1** lists some probability distributions that are used with GLM:

Table 3.1: GLM probability distributions

	Range of *y*	***F(y)***	***Canonical link:*** $\theta(\mu)$
Binomial $B(k,\mu)$	{0, …, k}	$\binom{k}{y}\mu^y(1-\mu)^{k-y}$	logit = $\log\left(\frac{\mu}{k-\mu}\right)$
Poisson $P(\mu)$	{0, 1, 2, …}	$\frac{\mu^y}{y!}e^{-\mu}$	log = $\log(\mu)$
Normal $N(\mu,\sigma^2)$	$\{-\infty,\infty\}$	$\frac{\exp\{-(y-\mu)^2/(2\sigma^2)\}}{\sigma\sqrt{2\pi}}$	identity = μ

Different canonical link functions are available, each distribution has a link function **Table 3.1**. log link is used where negative fitted values are forbidden and logit link is used where proportion of data is needed. Canonical link functions are the default options and will only occur when the linear predictor η is equal to the canonical parameter θ as defined in **(eq. 3.22)**.

Two distributions will be discussed in the following sections, the Normal (Gaussian) distribution as an example of continuous distributions and the Poisson distribution as an example of discrete distributions:

3.5.1 The Normal Distribution

Normal or Gaussian distribution, a bell-shape curve, is considered to be an important probability model in statistics. A normal process results when a number of unrelated, continuous random variables are added together [10]. Normal distribution is important because it is the most used statistical distribution since it arises naturally in many physical situations by the accumulation of many independent errors. A standard normal distribution is shown in Figure 3.5.

The normalized sum **Z** of mutually independent random variables in which the mean $\mu = 0$ and the variance $\sigma^2 = 1$ is:

$$Z = \frac{X_1 + X_2 + ... + X_n}{\sqrt{n}} \tag{3.23}$$

As **n** becomes large, **Z** is likely to be normally distributed according to the Central Limit Theorem which states:

For any distribution with a finite variance, the mean of a random sample from that distribution tend to be normally distributed [9].

Normally distributed random variables are standardized by employing the following transform:

$$Z = \frac{y - \mu}{\sigma} \tag{3.24}$$

A standardized normal Probability Density Function (PDF) is shown in Figure 3.5.The PDF is centered around the mean μ and its distribution spread is determined by the variance σ. The standardized normal distribution maybe viewed as the special case **N**(0,1) [10].

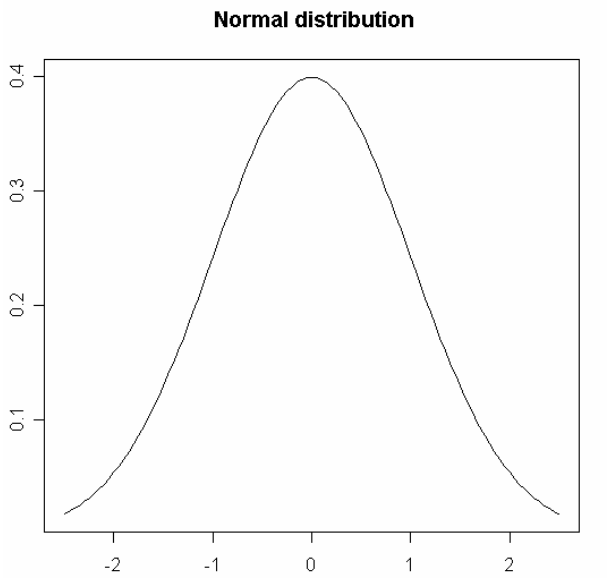


Figure 3.5: The standard normal distribution

3.5.2 The Poisson Distribution

A Poisson process describes the total number of independent events occurring during a specified observation period in which the event arrival rate is fixed [10]. Poisson PDF is shown in Figure 3.6, each arrival is causing a 'jump' of unit magnitude.

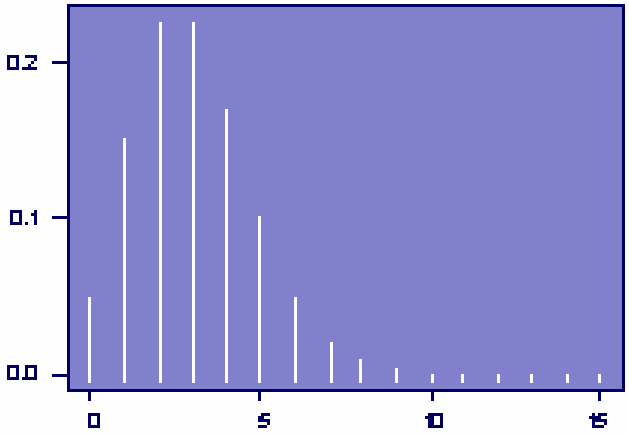


Figure 3.6: Poisson Probability Density Function (mean = 3)

The Probability Density Function (PDF) of Poisson process has only one parameter which is the arriving rate λ. Adding the terms of Poisson PMF will provide the following:

$$\sum_{y=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^{y}}{y!} = e^{-\lambda t}\left(1 + \frac{\lambda t}{1!} + \frac{(\lambda t)^{2}}{2!} + \frac{(\lambda t)^{3}}{3!} + \ldots\right) \quad \textbf{(3.25)}$$

where ***t*** is time and the expansion on the right hand side of the **(eq. 3.25)** is the Taylor series for $e^{\lambda t}$. As a result, Poisson distribution is $e^{-\lambda t}e^{\lambda t} = 1$ and this is a required condition for all probability distributions.

The expected value of the Poisson distribution is calculated by multiplying every term by ***y***:

$$E[N_t] = \lambda t \quad \textbf{(3.26)}$$

If λt increases without limit, the coefficient of skewness will reach zero and the Poisson distribution will become normal in appearance. The Poisson PDF can be expressed in the following compact way:

$$P_N\{y\} = \frac{e^{-\mu}\mu^{y}}{y!} \quad \textbf{(3.27)}$$

where $y = 0, 1, 2, \ldots,$ and $\mu \geq 0$ **(Table 3.1)**. In general, $\boldsymbol{P}(\mu)$ denotes a random variable with a Poisson distribution and $\mu = \lambda t$.

3.6 Normal and Poisson Regressions followed by Ordinary Kriging (Regression Kriging)

Regression kriging is carried out by combining two steps. First, the Generalized Linear Model (GLM) is fit with Poisson and Normal errors as discussed in section 3.5, and the observations for all data points are calculated. Then the Residuals are calculated and interpolated using ordinary kriging as discussed in section 3.4. Regression kriging is likely to enhance the spatial prediction because the error in the regression-prediction contains a spatial structure that can be described using ordinary kriging.

Chapter 4

4 KansK & Scientific Workflow

Scientific workflows are modules of data that flow through processing components. Each module represents the ways data being exported and imported through different software packages. Scientific workflows provide visual communication of different components and reproducibility of workflow components in different experiments.

Scientific workflows often manage large and heterogeneous data that can be computationally intensive. It constructs and executes complex scientific experiments such as the different processes in KansK toolbox, which is a toolbox that automates the distribution and prediction maps of different kind of species. Scientific workflows provide data reusability for different datasets, some basic processes can be replaced with a newly developed ones.

In the first section of this chapter, KansK toolbox methodology for spatial data prediction is presented, more information about the sample datasets used will be given. The second section discusses the need for scientific workflow management systems. Finally, in the third section, KansK will be presented as a sample case to be performed within a scientific workflow management system.

4.1 KansK Toolbox

KansK toolbox was created to automate the distribution and the prediction maps of different kind of species over the entire Netherlands. The datasets used were provided by SOVON - the Dutch Centre for Field Ornithology, these datasets are sample datasets for six different breeding-bird species distributed around the Netherlands. The unit modeled for the datasets is pairs/km^2. **Table 4.1** lists the environmental predictive variables of the datasets.

In KansK toolbox, eight spatial interpolation and regression techniques are examined. These techniques are explained in detail in chapter 3. One of the most important reasons of creating KansK toolbox is to investigate and evaluate these techniques. Based on the datasets we are using, the best spatial interpolation and regression techniques will be chosen.

KansK toolbox was developed using R. R is a programming language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible [20]. Some of R benefits are:

- R is free, open-source, and runs on UNIX, Windows and Macintosh.
- R provides an excellent built-in help.
- R provides good graphing capabilities.
- R has many built-in statistical functions and can easily call other user-written functions, C and Java functions.

Table 4.1: Environmental predictive variables

Variable	Description
Physical Geographical sub-regions in km^2	A division of the Netherlands in Physical Geographical Regions based on the major soil type of the region. Division of sub-regions has been made based also on the soil type, general land use, and broad spatial trends in densities.
X- and Y- coordinates	
Land use: top classification in km^2	Based on data from the top10-vector land use map, satellite information for different sources has been composed. 36 legend items have been used such as forest, marshland, urban and agricultural.
Land use: higher order in km^2	Higher order ecotopes such as forest and marsh.
Land use: lower order in km^2	Lower order ecotopes such as deciduous and coniferous forests.
Ground water table in km^2	Information of ground water levels are stored based on the soil types, there are 63 legend items in the ground water map that are combined in 6 classes from very wet to very dry.
Openness landscape	Based on the top 10-vector several maps has been made to describe the Dutch landscape. One of these maps describes the openness of the landscape on the scale of square kilometers. The 1 km-grid was not enough to describe the openness of the landscape. Therefore km-grid was interpolated using kriging to 25 meter-grid version with continues scale of 0-100.
Nature grassland in km^2	Grasslands are managed as nature reserve and hold different bird compositions compared to ordinary, intensively used grasslands. A vector map has been made of all the grasslands

4.1.1 Methodology proposed for spatial data prediction

In this section we present the methodology proposed for implementing KansK toolbox as a useful decision support tool for spatial data prediction. Figure 4.1 shows the structure of the methodology proposed for spatial data prediction using interpolation and regression techniques.

In KansK toolbox there are two main processes that are concerned with finding the appropriate spatial interpolation or regression technique to be used for spatial data prediction. In the first process, *data evaluation*, based on the species chosen, part of the data concerning that species is returned. This part is used to evaluate the model with a Calibration/ Validation cycle. This process will be repeated several times through a bootstrap procedure. Several error measurements will be monitored for each interpolation or regression technique. In the second process, *data prediction*, based on the results of data evaluation, the best interpolation or regression technique will be used for further species prediction. The methodology proposed processes are explained in details below:

- In Figure 4.1, the first process, *process data preparation*, is a common process and is needed by the other three processes, data need to be interpolated are prepared in this process. Process data preparation consists of three main steps. The first step creates two object files, one for evaluation data (FIT data) and the other is for prediction data. Each file will be return as an object dataset with the same environmental predictive variables listed in **Table 4.1**. These variables are subject to change based on the dataset being used.

  In the second step, *prepare data files*, different text files are created, these text files contain information regarding the datasets created in the first step, the interpolation and regression techniques for each bootstrap, years and months of the records of the datasets, and the environmental predictive variables used with each interpolation or regression technique. These text files are used as count references in other processes in KansK toolbox.

  In the third step, *prepare species run*, an object for the chosen species is created. This object is created based on the data returned from the first and the second steps. The object dataset is changed to km^2 and then passed to the other processes to be interpolated by the interpolation and regression techniques. In this step we can also produce distribution maps of the environmental predictive variables.

- In the second process, *process data evaluation*, FIT object dataset will be passed from the first process and the bootstrap number will be set. This dataset will be split into 80% calibration data and 20% validation data. These calibration/ validation data are then being interpolated using the eight interpolation and regression techniques explained in details in chapter 3.

As a result of the second process, a number of text files are returned. For each bootstrap a group of Residuals and RMSE (Root Mean Square Error) for each interpolation or regression technique are stored as text files.

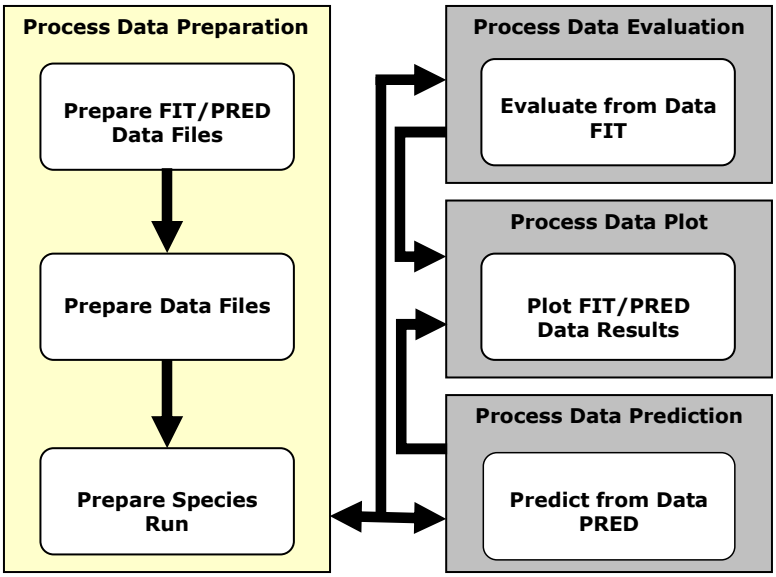


Figure 4.1: The methodology proposed for spatial data prediction

- In the third process, *process data prediction*, PRED object dataset will be passed from the first process. This dataset will be interpolated using the eight interpolation and regression techniques explained in details in chapter 3.

  As a result of the third process a number of map files are returned. The first group of map files will be visualized in R and the second group of map files have ArcView Geographical Information System (GIS) format.

  In the second and the third processes, the datasets are interpolated using eight interpolation and regression techniques. Figure 4.2 shows the flow of these processes. If Ordinary Kriging (OK) is chosen, the best semivariogram is returned first and then the datasets are interpolated. However, when using the Generalized Linear Models (GLM), the estimation values of the best significant values are chosen. Based on the estimation values, the environmental predictive variables are chosen and interpolated. More information about semivariogram and GLM are available in chapter 3.

- In the fourth process, *process data plot*, the results of the second and the third processes will be plot. Several kinds of plots and maps will be used. The results of the second, the third and the fourth processes will be discussed in details in chapter 5.

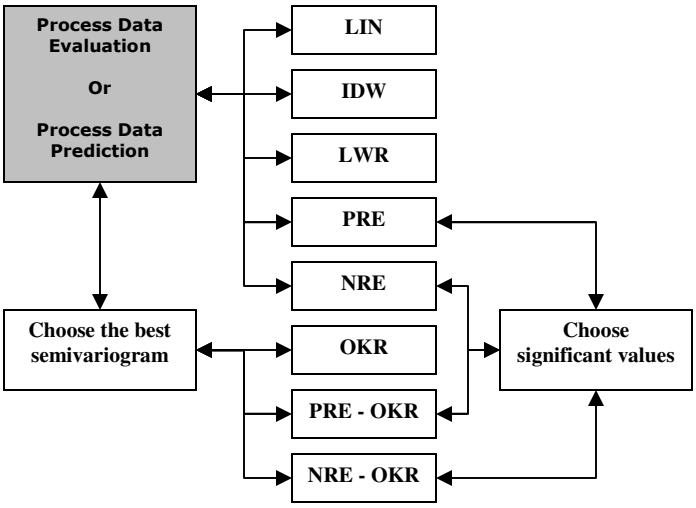


Figure 4.2: (Calibration/Validation/Prediction) method in the methodology proposed

4.2 Scientific Workflows and Management Systems

Grid environments break down barriers to communication and interaction, and make valuable resources accessible among groups of trusted users (called Virtual Organizations (VO)) [17]. Large numbers of people and sharable resources are involved in such environments to be able to make data and computing intensive scientific experiments possible. Scientific workflows consists of experiment routines and processes, each successful experiment is considered as a workflow template that will be considered as a reusable resource for studying other related problems.

Scientific Workflow Management Systems (SWMS) were introduced as system that is able to manage the dependencies between scientific experiment processes and to orchestrate resources runtime behavior. Scientific workflows are important for spatial systems like KansK for several reasons:

- The increasing demand on using the Grid computing capacity for running computation intensive processes such as the interpolation and regression techniques used in the (Calibration/Validation/Prediction) method, Figure 4.2.

- The integration of different environments into a single environment which deals with different software packages with expanding CPU processing capabilities such as integrating R for statistical computing and graphics into the scientific workflow.

- The discovery of distributed software resources and the availability of the Grid resources.

To be effective in running KansK toolbox, a scientific workflow management system needs to fulfill general quality criteria apart from meeting the specific requirements from the application domain [18]. Some of the generic quality requirements are:

- Flexible and generic modeling mechanism which is not domain specific and distributed resources are captured during runtime.
- A friendly environment which will allow the user to customize the configuration of the environment.
- User support and assistance at different levels.

Several scientific workflow systems have been introduced in the last decade and increasingly being used by scientists to construct and execute complex scientific analysis. Such analyses are typically data-centric and involve "gluing" together data retrieval, computation, and visualization components into a single executable analysis pipeline [19]. Some of these components contain computational intensive processes such as the interpolation and regression techniques discussed earlier. These processes are part of other software applications such as R scripts. Scientific workflows, such as VLAM-G, provide a way to compose and configure all these components.

4.3 Run KansK toolbox in VLAM-G

The Virtual Laboratory for e-Science (VL-e) project aims to realize a Grid enabled generic framework where scientists from different domains can share their knowledge and resources, and perform domain specific research [18]. The domains are: food informatics, medical diagnosis and imaging, bio-diversity, bio-informatics, high energy physics, and tele-science.

VLAM-G (Virtual Laboratory Amsterdam for Grid) is used as a prototype for the shared framework. It provides a service for managing data and resources, performing experiments location independent, and utilizing Grid resources transparently. VLAM-G also gives the scientists an access to geographically distributed resources, this is needed in KansK toolbox since scientists might run the toolbox location independent.

Experiment processes in VLAM-G are modeled explicitly using three elements: *physical entities* which are the instruments to be used, *activities* to be performed by the scientists, and *data elements* which are the input/output of the activities [17]. Each experiment has three levels: *Process Flow Template* (PFT) which is an abstract description of dependences, *Studies* which are instantiations of template, and *Experiment Topology* which is a set of self-contained software *modules* that performs computer tasks.

Domain experts and scientists are able to define *Studies* by instantiating a PFT through VLAM-G environment GUI. Scientists are able to define an experiment and execute it through an engine called *Run Time System* (RTS).

Modules, which are software entities, can be run in any computational resource within VL-e testbed. Each resource might contain different software libraries. The data stream between them is represented by connected arrows which represent the flow of data. Figure 4.3 shows the interface of VLAM-G, the PFT is displayed on the upper right side and KansK experiment is displayed on the lower right side.

VL-e provides support for R for statistical computing and graphics. In the experiment editor the scientist can define the elements composing the experiment. Such experiments are represented by *modules* that are implemented in R and connected to each others by arrows, each *module* has number of input/output ports. Figure 4.3 shows the experiment of KansK process data evaluation which matches with the first, second, and the fourth process of the methodology proposed in section 4.1.1.

KansK R-modules can be used in multiple experiments. Process data prediction can get advantage of data reusability since it is close to process data evaluation. A new experiment can be created or data prediction module can be added to the current experiment. Each module is being run in parallel with other modules on different computational resources, this will speed up the processing time for each module. One future advantage of this design is to run each interpolation or regression modules, as presented in Figure 4.2, in parallel by using the large number of Grid resources.

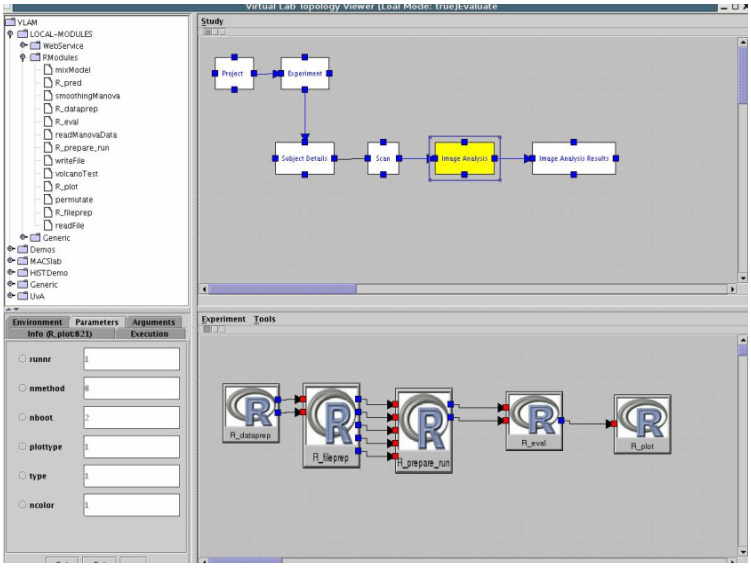


Figure 4.3: The interface of VLAM-G for composing a PFT and evaluation process experiment in KansK toolbox

Chapter 5

5 Results

In this chapter we discuss the results obtained by running KansK toolbox, which dataset we used, and an analysis and discussion will be presented.

In the first section we present the datasets used. We used three different tests to investigate the interpolation and regression techniques. Residuals results, which are the differences between observed and predicted values, are presented in section two. Root Mean Square Error (RMSE), RMSE are commonly used measures of success for numeric prediction, are presented in section three. Finally, in the fourth section, prediction maps using different spatial interpolation techniques are presented.

5.1 Datasets and bootstraps

KansK toolbox is created to work with different kind of spatial datasets. We tested KansK toolbox using six bird species datasets provided by SOVON, the bird species are listed in **Table 5.1**. In this chapter, the results of the Common Buzzard are presented in details as a sample result of these sample datasets.

Table 5.1: Bird species used in KansK toolbox

Dutch Name	**English Name**	**Scientific Name**
Buizerd	Common Buzzard	Buteo buteo
Scholekster	Eurasian Oystercatcher	Haematopus Ostralegus
Grutto	Black-tailed godwit	Limosa limosa
Houtduif	Common Wood Pigeon	Columba palumbus
Blauwborst	Blue throat	Luscinia svecica
Spreeuw	Common Starling	Sturnus vulgaris

To obtain the results of KansK toolbox, several tests have been made to return the Residuals and the RMSE for each bird species. Bootstraps are used to repeatedly sample the data in order to estimate confidence intervals for various parameters [9].

Each test was made using 100 bootstraps, each bootstrap has different random indexes for different chosen specie's dataset. Residuals and RMSE are averaged and presented in the next sections.

5.2 Residuals

The first test we used to investigate the spatial interpolation and regression techniques is the Residuals. Residuals are the experimental errors obtained using a model to predict the individual observations. The general formula for Residuals is:

Residuals = response variable – predicted values

The predicted values are calculated from the chosen model after estimating all unknown parameter from the experimental data. Residuals estimation and analysis give us a good indication about the assumptions and the choice of the models if they are reasonable and appropriate.

There are several graphical assumptions to examine the Residuals. Residuals should be similar to the normal bell-shaped discussed in details in section 3.5.1. Residuals have been examined in this thesis using two kind of tests, Residuals Normality and Residuals Homoscedasticity. The two tests are presented in the next sections.

5.2.1 Test of Residuals Normality

The first test we used to investigate the Residuals of the spatial interpolation and regression techniques is the Test of Residuals Normality. There are two common plots to display the Residuals Normality:

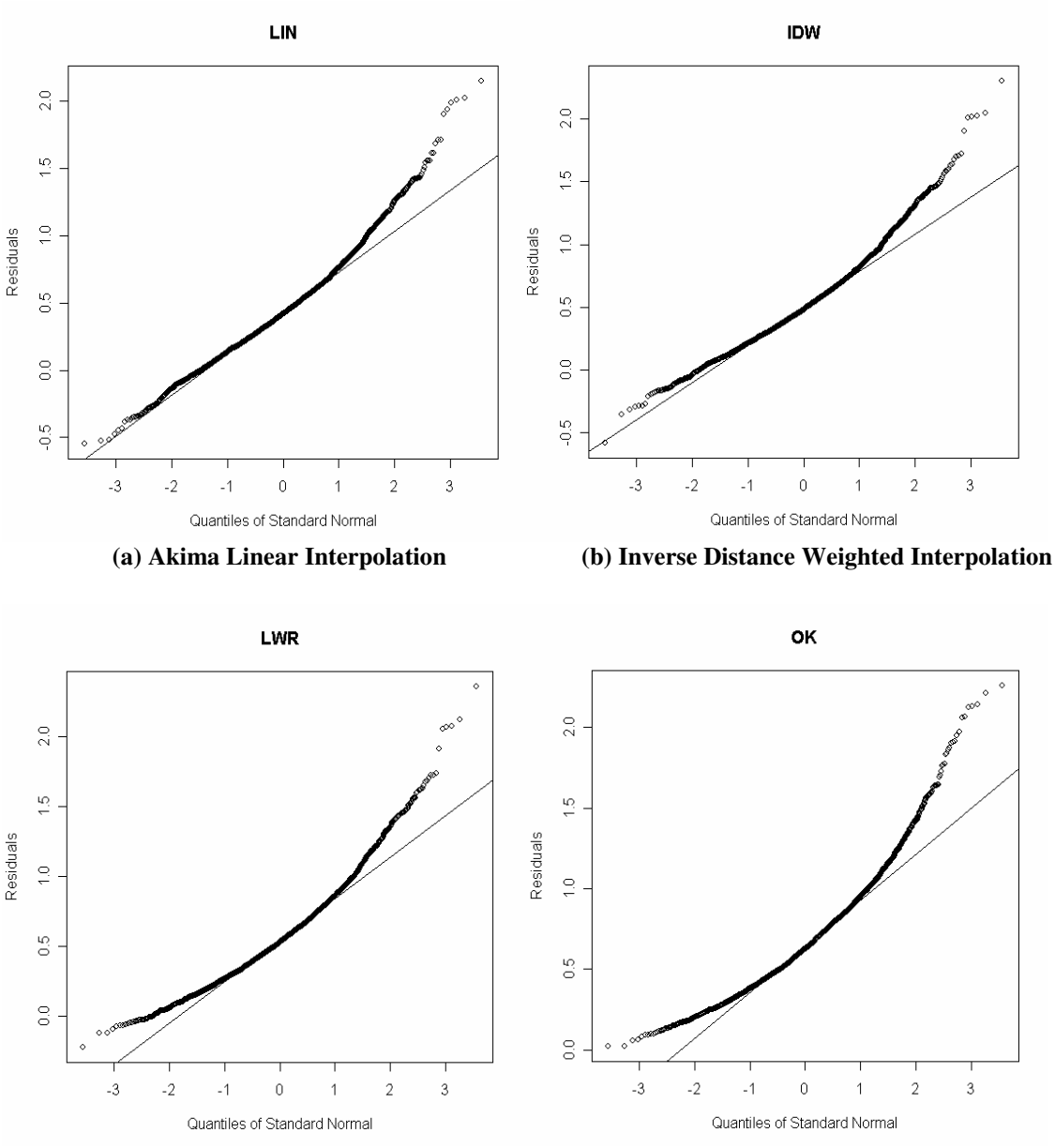
- **Histograms:** the range is split into equal-sized bins, each bin represents the count of the number of points from the chosen dataset that fall in it. The vertical axis is the Frequency and the Horizontal axis is the response variable, see Figure 3.5.

- **Normal Probability Plots:** it is a way to assess if the dataset is approximately normally distributed or not. The data are plotted against a theoretical normal distribution in a way the points will fall roughly on a **straight line**. This straight line passes through the first and the third quantiles of the dataset. The more the points deviate from the straight line, the less reasonable is the distribution. It is possible to tell from the Normal probability plots the following:

- Weather the distribution has longer or shorter tails than the normal distribution.

- Weather the distribution is skewed and in which direction.

In KansK toolbox, the normal probability plots are chosen to assess the normality of the data. Figure 5.1 and Figure 5.2 show the normal distributions of the eight spatial interpolation and regression techniques discussed in details in chapter 3.

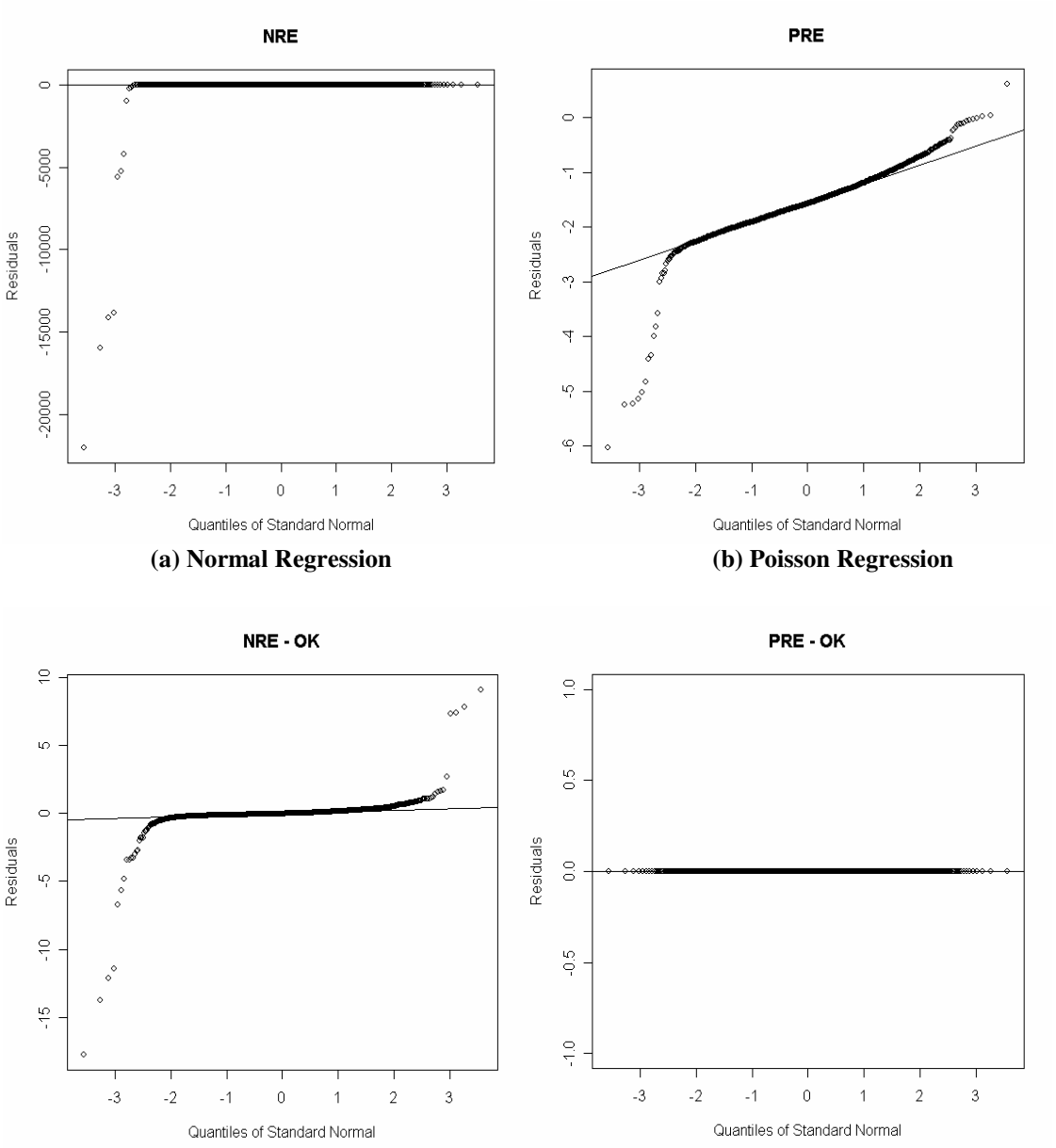


(a) Akima Linear Interpolation **(b) Inverse Distance Weighted Interpolation**

(c) Locally Weighted Regression **(d) Ordinary Kriging**

Figure 5.1: Residuals - normal probability plots for spatial interpolation techniques for the Common Buzzard *Buteo buteo*

The normal plots of the Residuals of the four spatial interpolation techniques are close to the straight line. Figure 5.1(a, b) have some outliers, which are some data points that are higher or lower than the reset of the data points, from the straight line which is common in normal probability. Figure 5.1(c, d) appear to deviate to the left from both sides of the straight line. This indicates a non-normal distribution behaviour in both the first and the third quantiles and a long tail to the right with a positive skewness in these techniques.



(a) Normal Regression **(b) Poisson Regression**

(c) Normal Regression – Kriging **(d) Poisson Regression - Kriging**

Figure 5.2: Residuals - normal probability plots for spatial regression techniques for the Common Buzzard *Buteo buteo*

Figure 5.2(a) has some extreme negative outliers, this indicates that the predicted values are less than the validation values for these techniques. Figure 5.2(b, c) have also extreme outliers from both sides of the straight line, they also have the ***S*** shape and this indicates that these two models have short tails relative to the normal distribution. Figure 5.2(d) has a straight line because this technique failed to return any Residuals results.

The test of Residuals Normality for the spatial interpolation techniques showed that Akima linear interpolation and inverse distance weighted interpolation have more reasonable results than the locally weighted regression and the ordinary kriging. Figure 5.1(c, d) showed that the two techniques have a non-normal distribution in both side of the straight line with positive skewness.

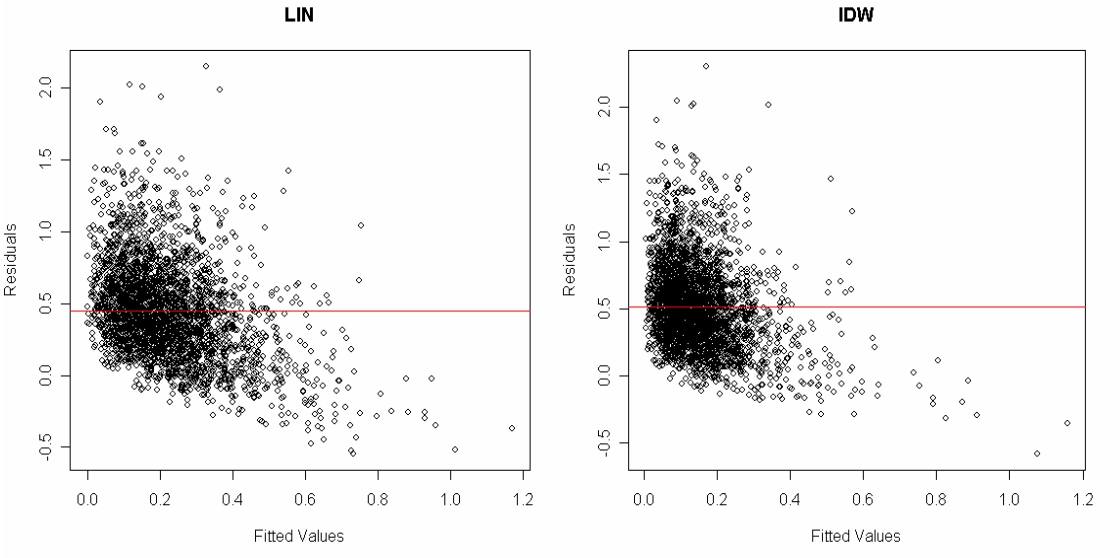
For spatial regression techniques, the Test of Residuals Normality showed that the four techniques have some extreme negative outliers and *S* shape and this indicates a short tail and a failure for these techniques to model this dataset. Because of this result, we won't present any further graphical results for these regression techniques.

5.2.2 Test of Homoscedasticity

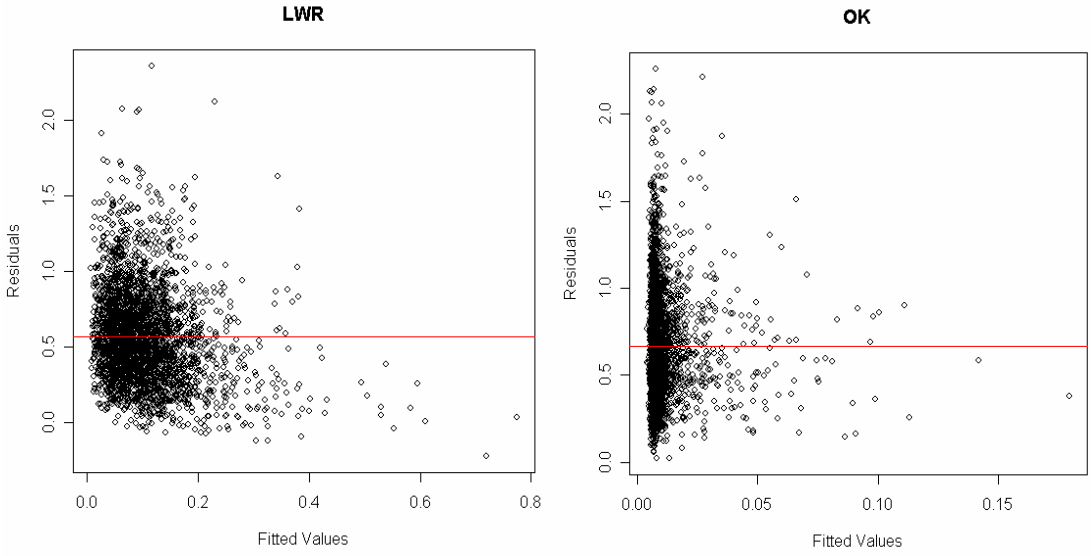
The second test we used to investigate the Residuals of spatial interpolation and regression techniques is the Homoscedasticity. Homoscedasticity means that the residuals are approximately equal for all the predicted dependent variables. Dataset is homoscedastic if the residuals plot has the same width for all values of the predicted dependent variables. However, dataset is heteroscedastic if a cluster of points gets wider as the values of the predicted dependent variables get larger.

Figure 5.3(a, b, c, d) show the scatterplots for the linear interpolation, the inverse distance weighted interpolation, the locally weighted regression, and ordinary kriging. For each interpolation technique the Residuals are scattered around the mean line which is biased to the positive side for the four techniques, this indicates that there are around 0.5 less species predicted in each observation point. There are more outliers for the small predicted fitted values.

We have indicated from this test that the four spatial interpolation techniques gave a similar behaviour of being heteroscedastic and biased. **Table 5.2** shows that locally weighted regression and ordinary kriging are more biased than Akima linear interpolation and inverse distance weighted. The biases for the regression techniques are not valid since they are negative and not close to the biases of the interpolation techniques and this also indicate a failure for these techniques. More tests are applied to this dataset in the coming sections to investigate more the behaviour of these techniques.



(a) Akima Linear Interpolation **(b) Inverse Distance Weighted Interpolation**



(c) Locally Weighted Regression **(d) Ordinary Kriging**

Figure 5.3: (Residuals vs. Fitted Values) - Scatter plots for spatial interpolation techniques for the Common Buzzard *Buteo buteo*

5.3 Root Mean Square Error

The second test we used to investigate the spatial interpolation and regression techniques is the Root Mean Square Error (RMSE). RMSE is commonly used to measure the success of numeric prediction. It is a measure of the average error across a map and is used in digitising to give an approximate measure of the difference between the real-world coordinates and the registration points on the digital layer [21].

RMSE is the square root of the square of the difference between estimated point and interpolated observation point divided by the total number of the observation points:

$$\mathbf{RMSE} = \frac{\left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2}}{(n)}$$

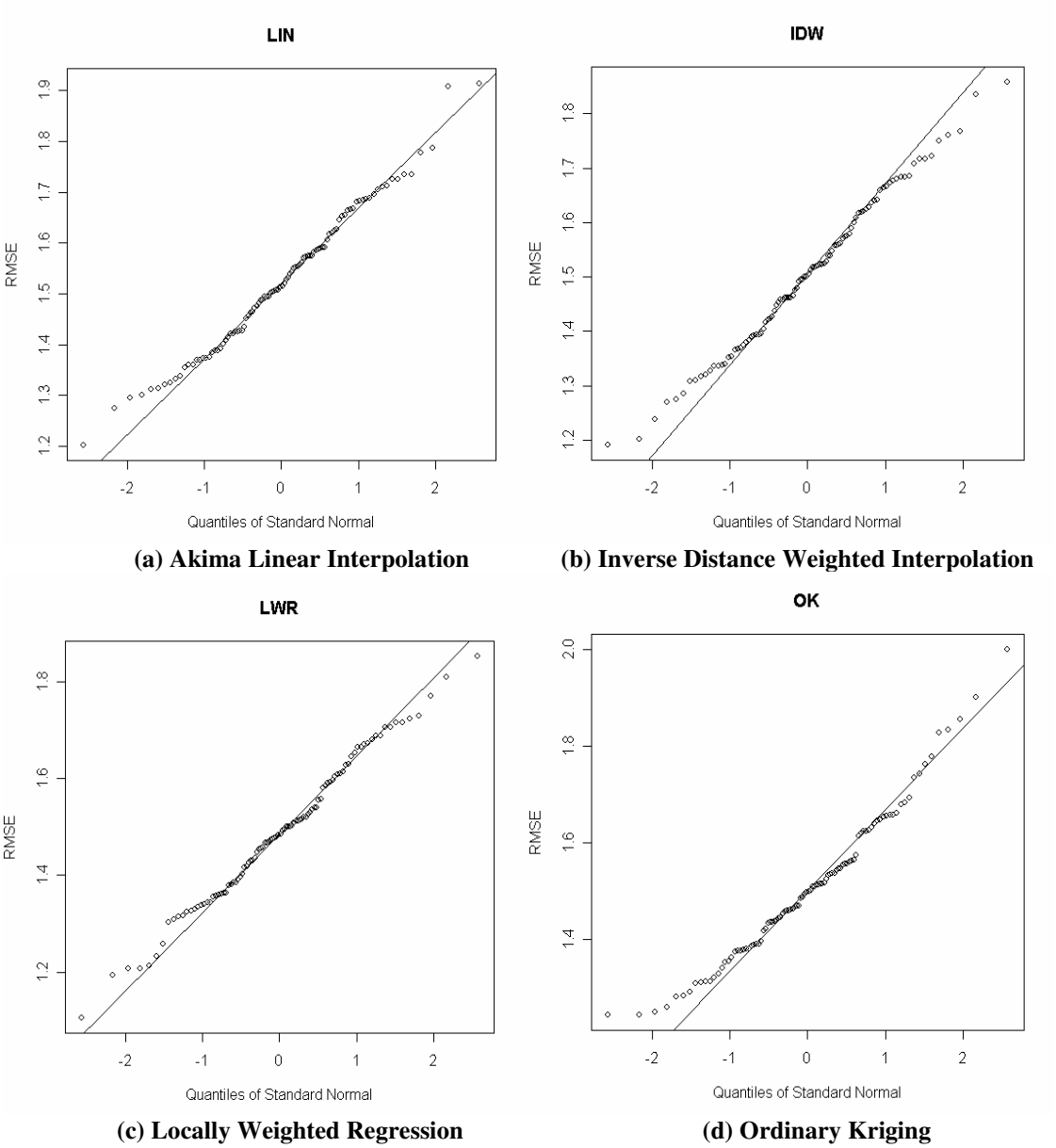
where x_i is the i-th point in array X, y_i is the y-th point in array Y, and n is the total number of points in array X and Y.

Figure 5.4 shows the RMSE normal probability plots for the spatial interpolation techniques of the Common Buzzard. The normal plots for the four techniques are close to the straight line with some outliers which is common in normal probability.

Table 5.2: (RMSE/ Bias) results for spatial interpolation and regression techniques for the Common Buzzard *Buteo buteo*

Interpolation/Regression	Bias	RMSE
Akima Linear Interpolation	0.4461777	1.524020
Inverse Distance Weighted Interpolation	0.5166599	1.504721
Locally Weighted Regression	0.5657069	1.486310
Ordinary Kriging	0.6693147	1.507431
Normal Regression	-2710.592	370907.0
Poisson Regression	-2259.087	2.376199
Normal Regression - Ordinary Kriging	-1936.364	3.208888
Poisson Regression - Ordinary Kriging	-1694.318	NA

Table 5.2 shows the average results of the RMSE for the eight interpolation and regression techniques used in KansK toolbox. The first four techniques have close results to each other and indicate a success. However, the last four techniques have high RMSE results. These techniques performed poorly for the Common Buzzard dataset and the other datasets, this is because these datasets are not normally distributed, and this indicates that regression techniques for such datasets are not appropriate.



(a) Akima Linear Interpolation **(b) Inverse Distance Weighted Interpolation**

(c) Locally Weighted Regression **(d) Ordinary Kriging**

Figure 5.4: RMSE normal probability plots for spatial interpolation techniques for the Common Buzzard *Buteo buteo*

Both Residuals and RMSE tests showed close results for the spatial interpolation techniques with an advantage to Akima linear interpolation and inverse distance weighted interpolation. In the next section, we present the prediction maps of the Common Buzzard species.

5.4 Prediction Maps

The third test we used to investigate the spatial interpolation and regression techniques is the Prediction Maps. In this section, prediction maps for the Common Buzzard species are presented. These maps are created using the spatial interpolation techniques and are comparable to the prediction map of the Common Buzzard of the BAMBAS project presented in details in chapter 2.

Common Buzzards are mostly concentrated in High and Riverine Netherlands with less population in Low Netherlands and fewer populations in the north part of Low Netherlands, Figure 5.5 shows the distribution of the Common Buzzard around the Netherlands.

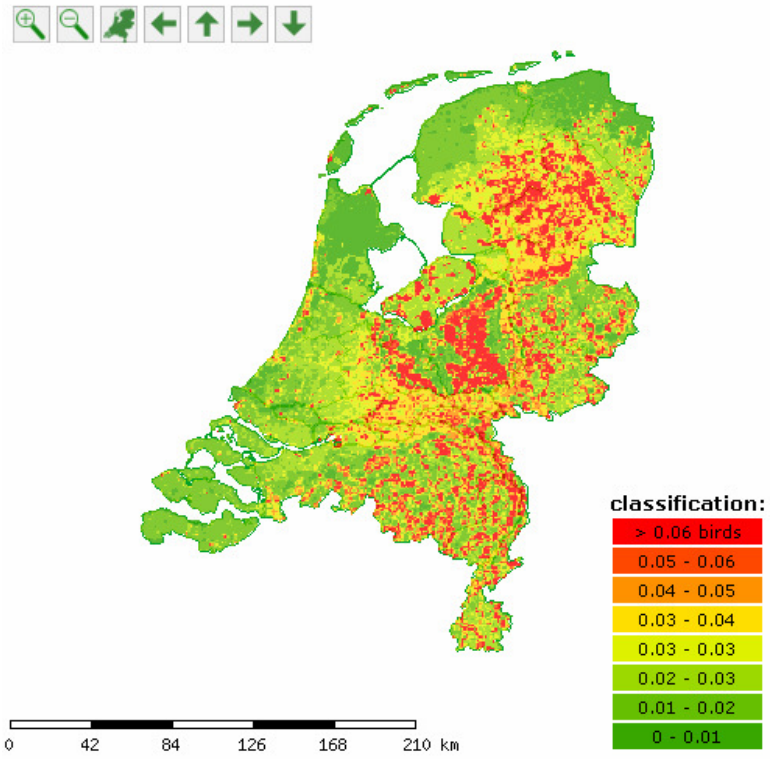
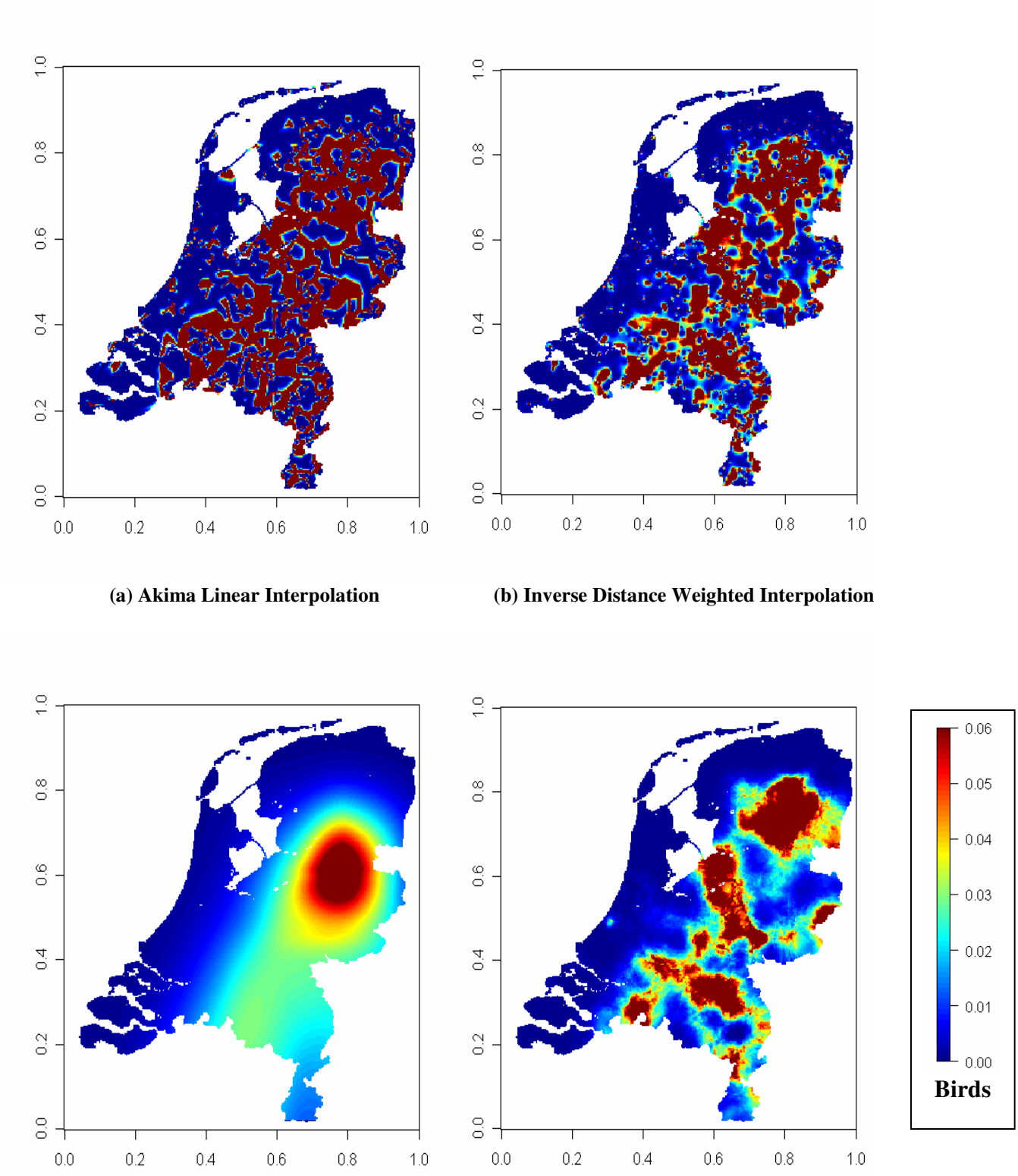


Figure 5.5: Common Buzzard prediction map - BAMBAS project



(a) Akima Linear Interpolation **(b) Inverse Distance Weighted Interpolation**

(c) Locally Weighted Regression **(d) Ordinary Kriging**

Figure 5.6: Common Buzzard prediction map – KansK toolbox

Figure 5.6 shows the prediction maps for the Common Buzzard using KansK toolbox. The Residuals and the RMSE results recommended using Akima linear interpolation and inverse distance weighted interpolation. The prediction maps for both interpolation techniques are close to each other, Figure 5.6(a, b). However, Akima linear interpolation has connected regions because of the way the interpolation technique works, which triangulate the estimated points as discussed in section 3.1.

Locally weighted regression makes contours of the Common Buzzard species densities, Figure 5.6 (c) shows that concentration of the Common Buzzards are in the red contour and the surrounding contours have less densities. Ordinary Kriging is making regions of high densities of the Common Buzzards Figure 5.6 (d), and these regions are covering several empty or less density smaller regions.

The maps give the final results in favour for inverse distance weighted interpolation since it is smoother than the linear interpolation and the other interpolation techniques. It is also close to the prediction map of BAMBAS project for the Common Buzzard species Figure 5.5. The results obtained in this chapter also validate the expectation we reached in chapter 2 that kriging, inverse distance weighted and locally weighted regression will give reasonable performance. Other datasets prediction maps are presented in Appendix B.

Chapter 6

6 Conclusions and Future Work

6.1 Conclusions

In this thesis, we compared different spatial interpolation and regression techniques for spatial data prediction. We concluded that it is important to investigate such techniques because of the expanding interest of the spatial prediction and the limited knowledge about the applicability of these techniques.

We started this thesis by investigating several projects and cases where interpolation and regression techniques were used for different spatial datasets. These projects and case studies indicated that inverse distance weighted interpolation, kriging interpolation, and locally weighted regression were relatively suitable techniques. All of these techniques showed a good performance and this gave us a good indication that these techniques will work well for other different datasets.

A methodology overview of KansK toolbox has been given, KansK toolbox investigated a computational intensive spatial interpolation and regression techniques that are written in R script, the Grid-based Workflow Management Systems such as VLAM-G provide a way to compose and configure all these techniques. The benefits of using VLAM-G are uncountable, the future users of KansK toolbox will be able to benefit from the geographically distributed resources available which will speedup the performance of KansK modules. Another important benefit is that the scientist will be able to reuse the proposed experiment for different datasets and will be able to change some of the basic components with newly developed ones.

We presented our results for one dataset of the six datasets tested. The dataset is for the Common Buzzard. We used three kinds of measurements, Residuals, RMSE, and prediction maps. Residuals and RMSE showed reasonable results for the interpolation techniques and bad results for the regression techniques. Regression Residuals results are not normally distributed for the datasets used. Interpolation Residuals and RMSE were in favour towards Akima linear and inverse distance weighted interpolations. Predication maps made it clear that inverse distance weighted interpolations produced smoother maps which are close to the results of BAMBAS project.

6.2 Future work

This toolbox is a building block for a decision support tool. This toolbox can be expanded easily to support other interpolation techniques. Future research may also include an analysis of the datasets used to try to know more about the unexpected behavior of the regression techniques.

More tests and analysis have to be performed on other species datasets for the entire region of interest. It is important to monitor if the behavior and the differences of these interpolation techniques stay the same or become better or even worse.

From a reusability point of view, it would be easier for the user to be able to merge different topology experiments. The proposed experiments in KansK toolbox are executed separately. The intermediate data have to be saved then reloaded to be used in the next phase of the experiment. This lack of control flow in VLAM-G adds an unnecessary burden to the end user. Finally, a future good research is trying to run this toolbox in different workflow systems and making a comparison out of it.

Bibliography

[1] Akima, H. 1978, 'A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points', *ACM Transactions on Mathematical Software*, vol. 4, no. 2, pp. 148-159.

[2] Lawson, C. L. 1972, 'Generation of a Triangular Grid with Application to Contour Plotting', *Technical Memorandum*, section 914, no. 299, pp. 3-6.

[3] Felgueiras, C. A. & Goodchild, M. F. 1995, 'A comparison of three TIN surface modeling methods and associated algorithms', *National Center for Geographic Information and Analysis*, Santa Barbara, CA, pp. 6-8.

[4] Shepard, D. 1968, 'A two-dimensional interpolation function for irregularly-spred data', *23rd ACM National Conference*, pp. 517-424.

[5] Peralvo, M. 2003, 'Influence of DEM interpolation methods in drainage analysis', [online] available at: http://www.crwr.utexas.edu/gis/gishydro04/

[6] Cleveland, W. S. 1979, 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829-836.

[7] Davis, J. C. 2002, *Statistics and Data Analysis in Geology*, John Wiley & Sons, New York, pp. 416-428.

[8] McCullagh, P. & Nelder, J.A. 1989, *Generalized Linear Models: Monographs on Statistics and Applied Probability*, 2nd edn, Chapman and Hall, London.

[9] Crawley, M. J. 2002, *Statistical Computing: An Introduction to Data Analysis using S-Plus*, John Wiley & Sons Ltd, Chichester.

[10] Ott, W. R. 1995, *Environmental Statistics and Data Analysis*, Lewis Publishers, Boca Raton.

[11] Shamoun, J., Sierdsema, H., van Loon, E., van Gasteren, H., Bouten, W. & Sluiter, F. 2005, 'Linking Horizontal and Vertical Models to Predict 3D + time Distributions of Bird Densities', *International Bird Strike Committee*, Athens.

[12] Sierdsema, H. & van Loon, E. 2006, 'Filling the Gaps: Using Count Survey Data to Predict Bird Density Distribution Patterns and Estimate Population Sizes', *Turkish Journal of Zoology (Accepted).*

[13] Naoum, S. & Tsanis, I.K. 2004, 'Ranking Spatial Interpolation Techniques Using a GIS-Based DSS', *Global Nest: the Int. J.*, vol. 6, no.1, pp. 1-20.

[14] Wikipedia: *the Free Encyclopedia*, Available at: http://www.wikipedia.org

[15] Collins, F. C. & Bolstad, P. V. 1996, 'A Comparison of Spatial Interpolation Techniques in Temperature Estimation', *The Third International Conference/Workshop on Integrating GIS and Environmental Modeling*, Santa Fe, NM.

[16] Vance, M. D., Fahrig, L. & Flather, C. H. 2003, 'Effect of Reproductive Rate on Minimum Habitat Requirements of Forest-Breeding Birds', *Ecology*, vol. 84, no. 10, pp. 2643-2653.

[17] Zhao, Z., Belloum, A., Yakali, H., Sloot, P. & Hertzberger, B. 2005, 'Dynamic Workflow in a Grid Enabled Problem Solving Environment', *The 5th International Conference on Computer and Information Technology (CIT2005)*, Shanghai, China.

[18] Zhao, Z., Belloum, A., Wibisono, A., Terpstra, F., de Boer, P. T., Sloot, P. & Hertzberger, B. 2005, 'Scientific Workflow Management: between Generality and Applicability', *QSIC*, pp. 357-364.

[19] Bowers, S., Ngu, A. H. H., Ludascher, B. & Critchlow, T. 2006, 'Enabling Scientific Workflow Reuse through Structures Composition of Dataflow and Control-Flow', *The 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, Atlanta, GA.

[20] The R Project for Statistical Computing: *Introduction to R*, Available at: http://www.r-project.org

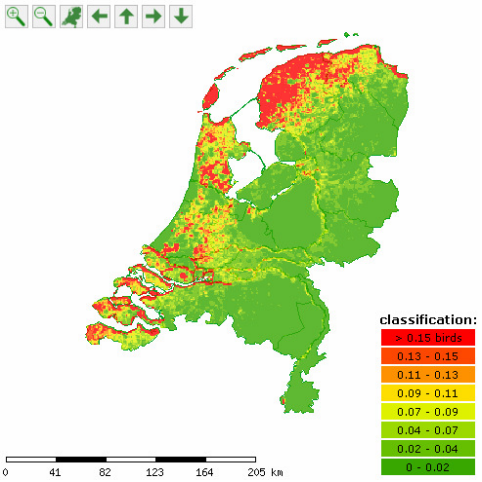
[21] A place in history: *a guide to using GIS in historical research*, Available at: http://hds.essex.ac.uk/g2gp/gis/sect101.asp

Appendices

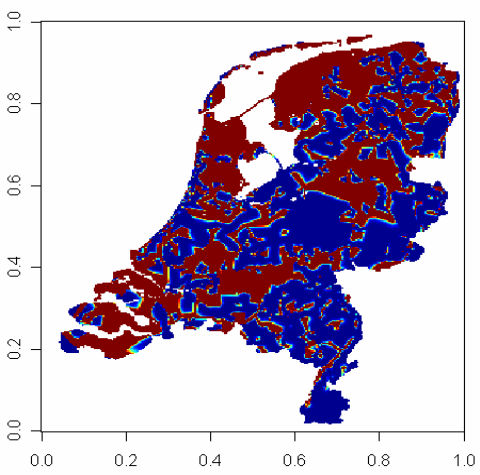
Appendix A – Acronym and Abbreviations

ArcGIS	Arc- Geographical Information Systems
BAMBAS	Bird Avoidance Model/ Bird Avoidance System
GAM	General Additive Model
GIS	Geographical Information Systems
GLM	Generalized Linear Models
IDW	Inverse Distance Weighted
KansK	Kans Kaart
LIN	Linear Interpolations
LULC	Land Use and Land Cover
LWR	Locally Weighted Regression
MAE	Mean Absolute Error
MSE	Mean Squared Error
NKR	Normal Regression followed by Ordinary Kriging
NRE	Normal Regression
OK	Ordinary Kriging
PDF	Probability Density Function
PFT	Process Flow Template
PKR	Normal Regression followed by Ordinary Kriging
PRE	Poisson Regression
RMSE	Root Mean Square Error
RTS	Run Time System
SOVON	the Dutch Centre for Field Ornithology
SWMS	Scientific Workflow Management Systems
USGS	U.S. Geological Survey
VLAM-G	Virtual Laboratory Amsterdam for Grid
VL-e	Virtual Laboratory for e-Science
VO	Virtual Organizations

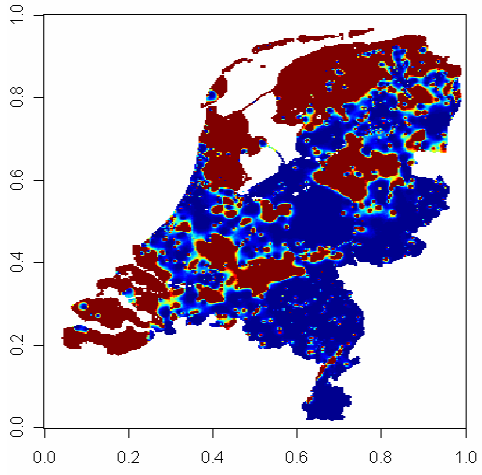
Appendix B – Prediction Maps (Scholekster)



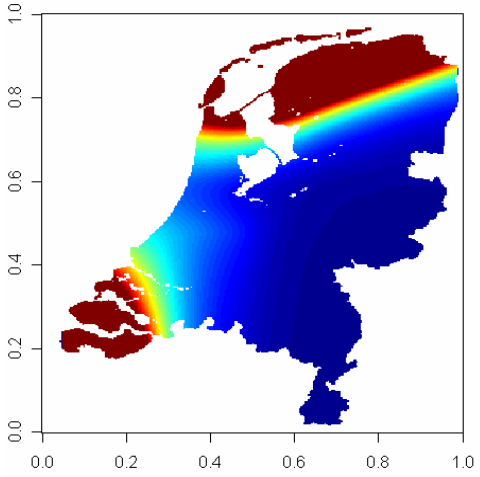
BAMBAS



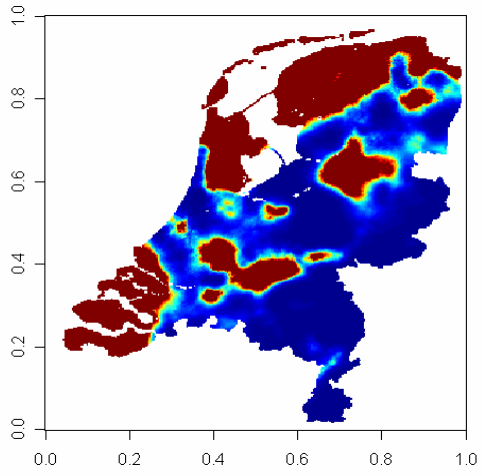
(a) Akima Linear Interpolation



(b) Inverse Distance Weighted Interpolation

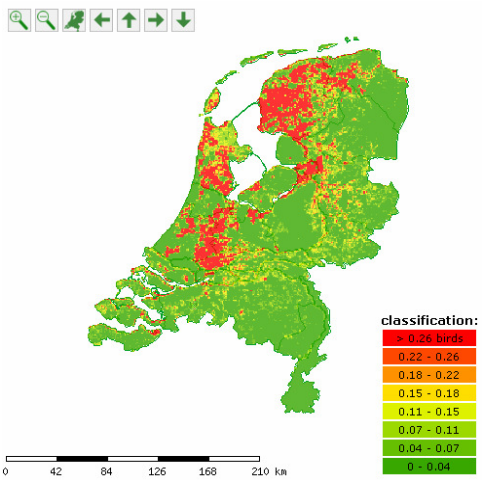


(c) Locally Weighted Regression

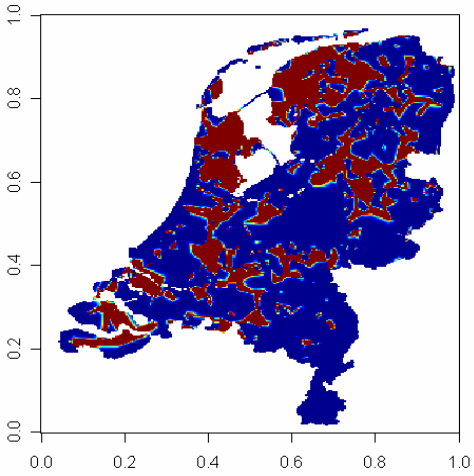


(d) Ordinary Kriging

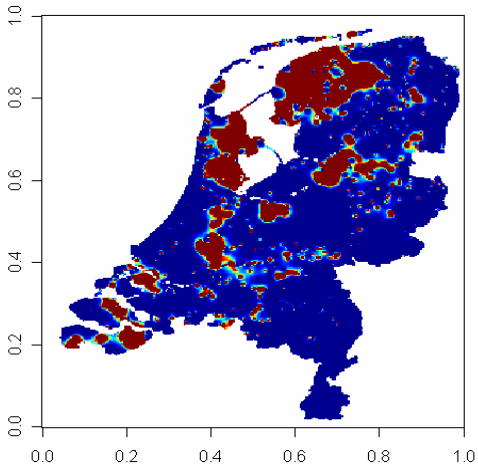
Prediction Maps (Grutto)



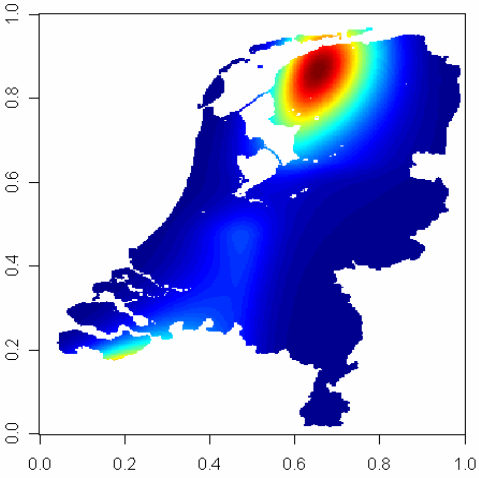
BAMBAS



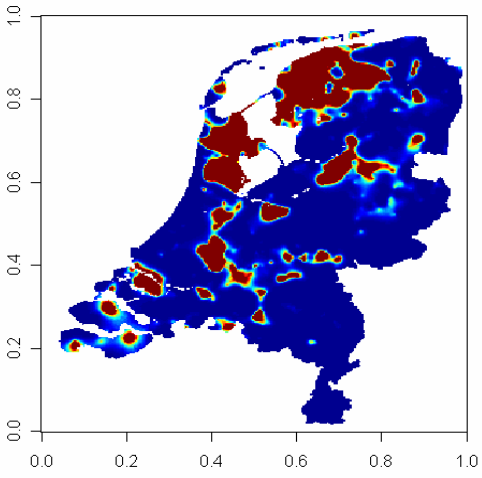
(a) Akima Linear Interpolation



(b) Inverse Distance Weighted Interpolation

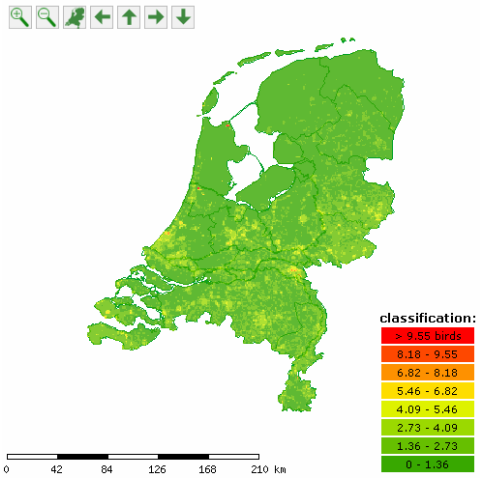


(c) Locally Weighted Regression

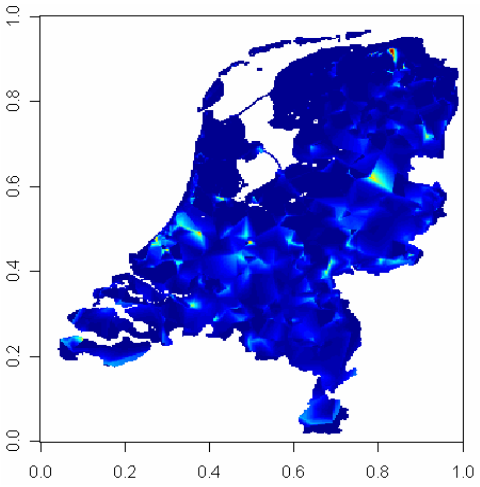


(d) Ordinary Kriging

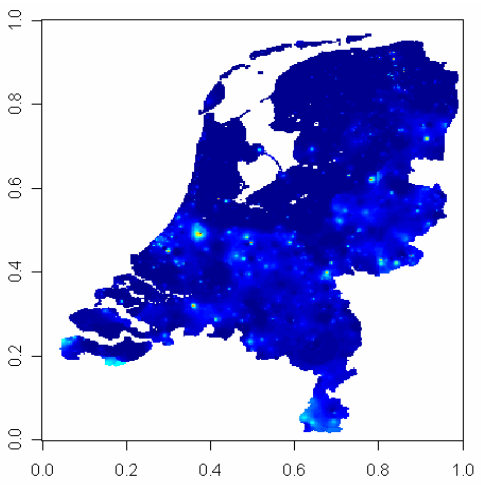
Prediction Maps (Houtduif)



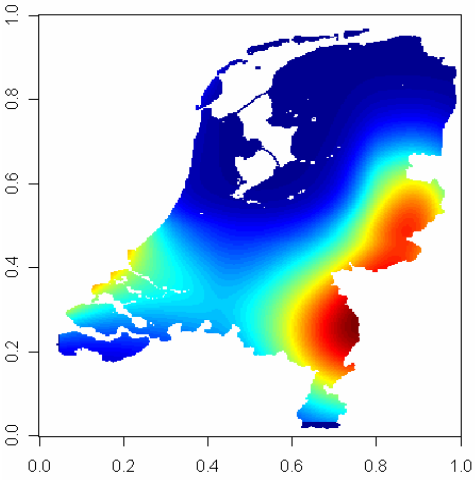
BAMBAS



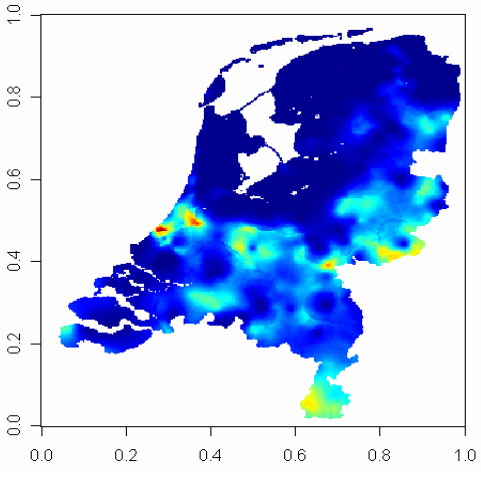
(a) Akima Linear Interpolation



(b) Inverse Distance Weighted Interpolation

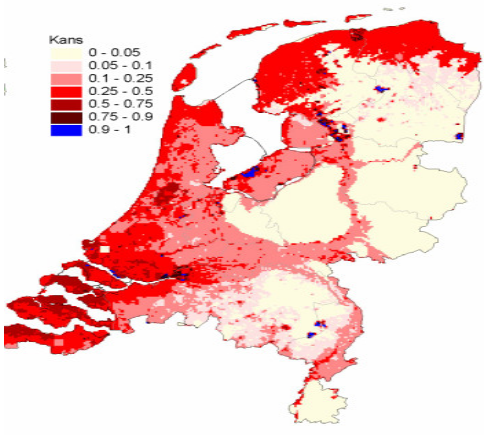


(c) Locally Weighted Regression

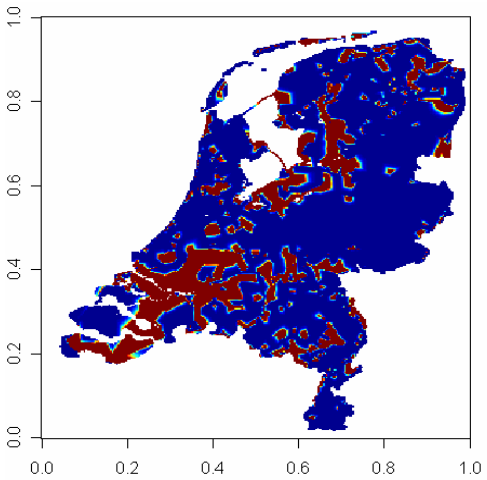


(d) Ordinary Kriging

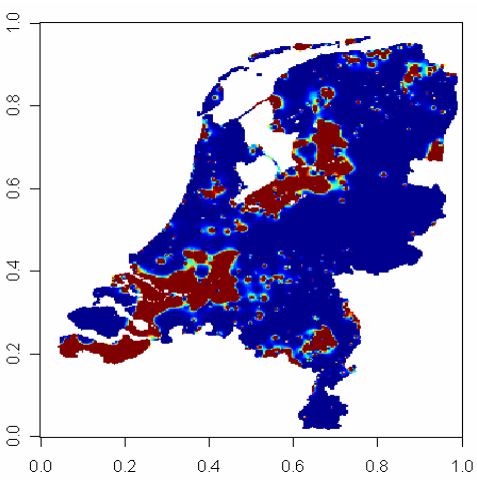
Prediction Maps (Blauwborst)



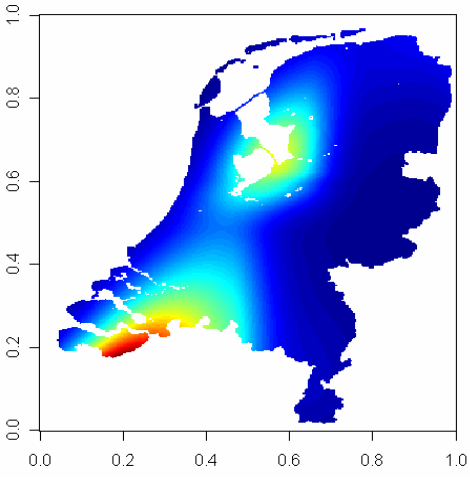
BAMBAS



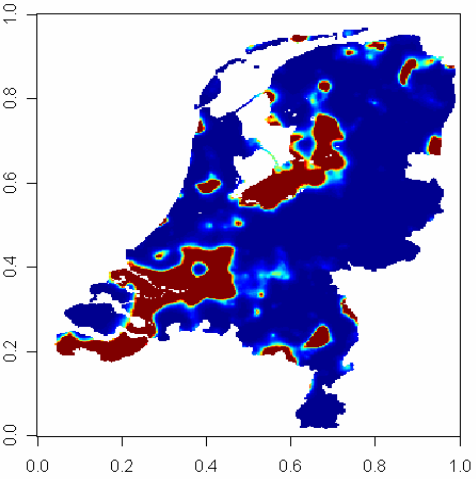
(a) Akima Linear Interpolation



(b) Inverse Distance Weighted Interpolation

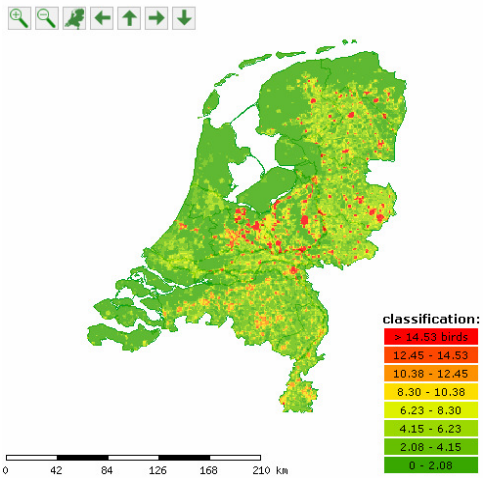


(c) Locally Weighted Regression

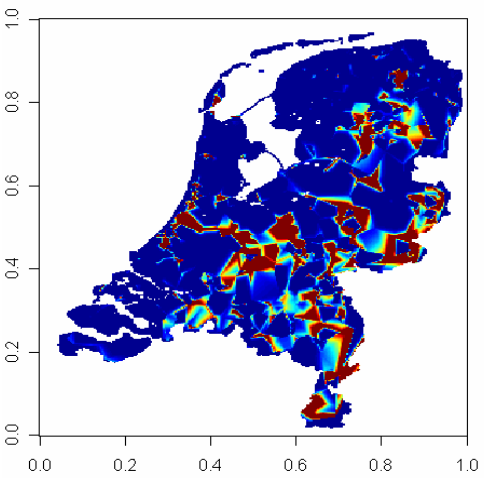


(d) Ordinary Kriging

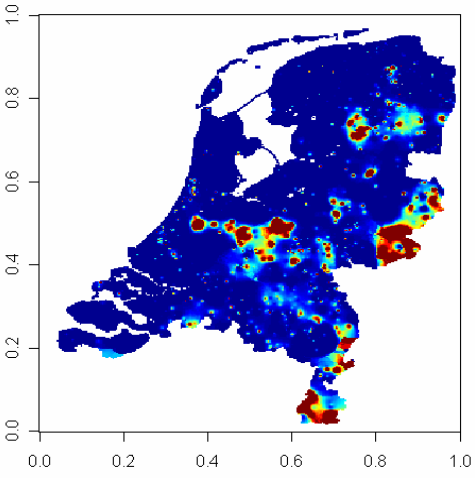
Prediction Maps (Spreeuw)



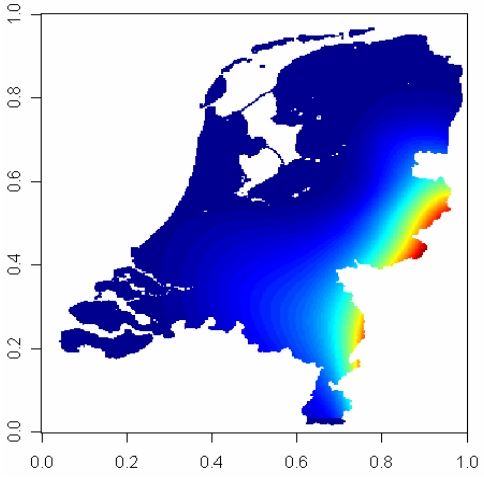
BAMBAS



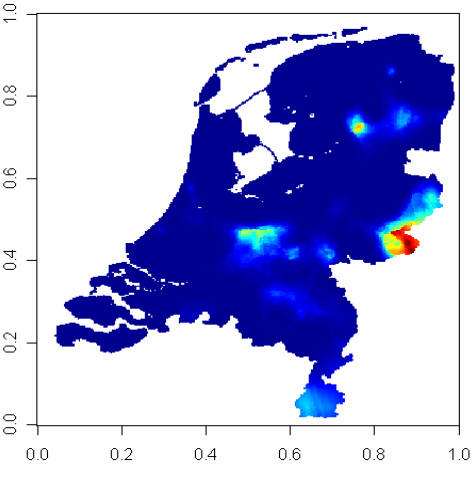
(a) Akima Linear Interpolation



(b) Inverse Distance Weighted Interpolation



(c) Locally Weighted Regression



(d) Ordinary Kriging