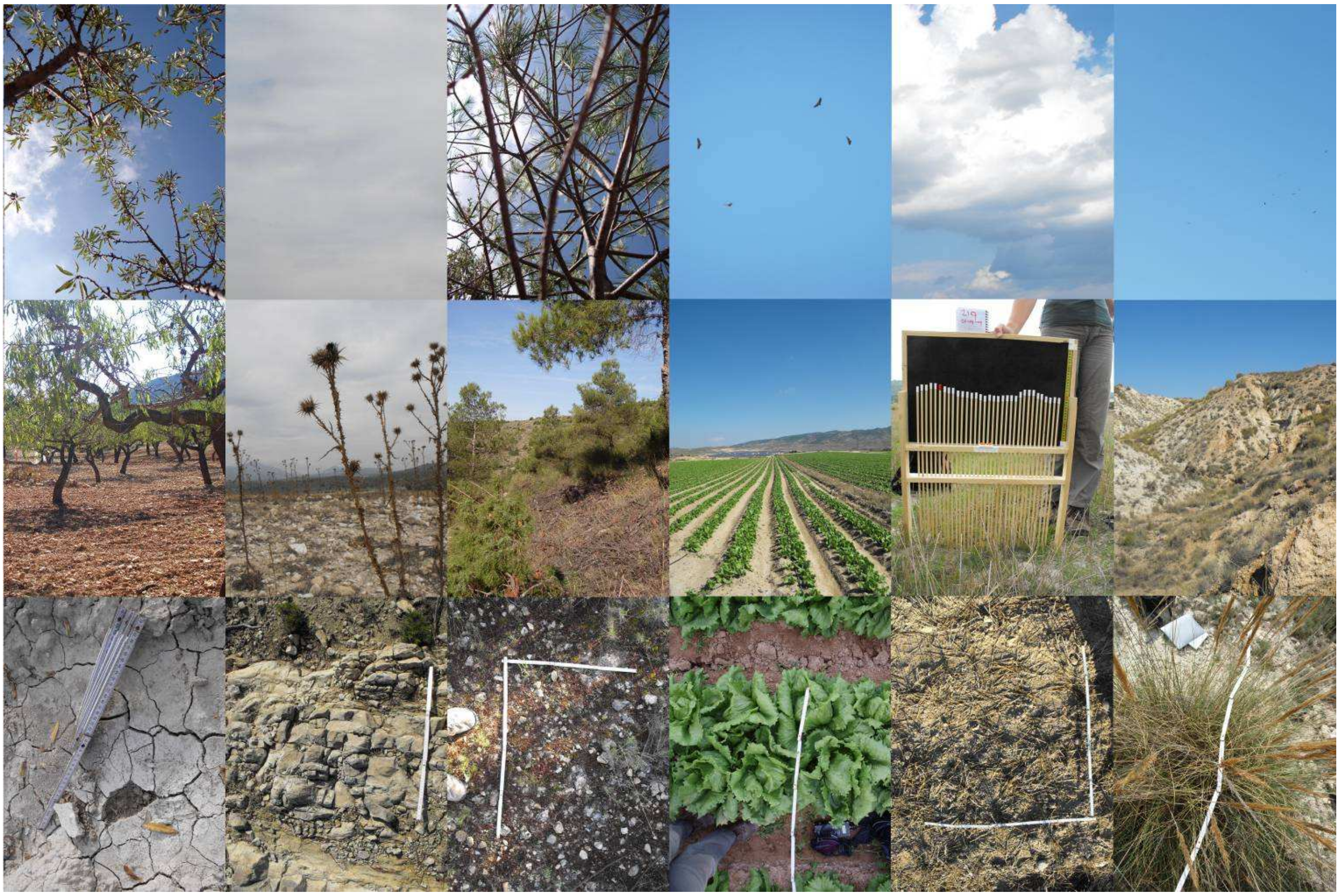
Land Degradation & Development Group

Statistical analysis of erosion parameters for the sediment yield model "LISEM" in the Torrealvilla catchment, SE Spain



MSc Thesis by H.J. Wijbrans
November 2010

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Statistical analysis of erosion parameters for the sediment yield model “LISEM” in the Torrealvilla catchment, SE Spain

Minor Master thesis Land Degradation and Development Group submitted in partial fulfillment of the degree of Master of Science in Hydrology and Water Quality at Wageningen University, the Netherlands.

MSc Thesis by	**H.J. Wijbrans, BSc**
Program:	MSc Hydrology and Water Quality (MHW)
Reg nr:	8702 1097 5010
Course:	LDD 80324
	Master thesis Land Degradation and Development
Date:	November 2010

Supervisors:
J.E.M. Baartman, MSc
Dr. S.D. Keesstra

Examinator:
Prof.dr.ir. L. Stroosnijder

Wageningen University, Land Degradation and Development Group

Abstract

This research is an analysis of the dependency of the erosion parameters for LISEM (Limburg Soil Erosion Model) on the different landscape features (land use, geology and soil). This dependency is based on auxiliary maps, for the Torrealvilla catchment in South East Spain. Found dependencies are used to create a relevant interpolation of the parameters, which can be used as input maps of LISEM (Jetten, 2002). Also, knowledge on the relation between the erosion parameters and landscape features can improve future fieldwork efficiency.

Field data is gathered according to Hessel (2002). Statistical tests (ANOVA and Kruskall-Wallis) are used to asses variances and dependencies between the landscape feature classes and the LISEM parameters. A prediction is made for the erosion parameters using linear regression, Ordinary Kriging or Regression Kriging, depending on significance of the relation between the parameter and auxiliary maps and spatial dependency.

Predicted maps of parameters are validated using a separate dataset. No correlation is found between the landscape feature maps and the measured LISEM parameters. Results of validation is that modelled parameter maps have little or no correlation with the validation dataset. With a lowest efficiency value of 0.01 and a highest efficiency of 0.33, based on the coefficient of determination (R^2). It can not be concluded that no correlation is present in the landscape, because the options that the data is not representative and that the used methods do not work in this situation are both a possible cause for the lack of correlation.

Table of contents

1. Introduction

Because fieldwork is expensive and takes a lot of time, efficiency in fieldwork is important. Often, fieldwork is a limiting factor to environmental research due to limitations in time and money. This can cause limited field data and result in poor field representation. For these reasons it is relevant to make fieldwork as efficient as possible and is it useful to try to improve the effectiveness of data collection.

The goal of this research is to improve field sampling efficiency by finding a relation between landscape characteristics (geology, soil and land use) and erosion parameters of the model LISEM (Limburg Soil Erosion Model). In most literature the methods of data analysis are described in a very compact way (De Roo et al., 1996, 1996a, 1996b; Jetten et al*.,* 1996; Hessel, 2003; Boer et al., 2005). No literature was found on the statistical analysis of the parameters of LISEM. This may mean that the data were assumed to be relevant and no data analysis has been done.

This research tries to relate model parameters to landscape features based on auxiliary maps. This is done by gathering field data, which is analysed for correlations with the landscape features using statistics. The hypothesis of this research is that the LISEM parameters have a correlation with the landscape features, depending on the nature of the parameters (section 1.1). If a correlation is present between model parameters and landscape features, that knowledge can be used in future data collection to make field sampling more efficient.

To obtain a dataset for LISEM to look at these correlations, fieldwork is done in the Torreavilla catchment in South East Spain. The measured parameters are: Surface cover by vegetation (PER), leaf area index (LAI),crop height (CH), random roughness (RR), stone fraction (STFRC), median grain size of the soil (D50), cohesion (COH) of the soil, initial soil moisture content (θi) and saturated soil moisture content (θs).

The sampling scheme of the fieldwork was set up in such a way that the correlation between the LISEM parameters and the landscape features could be investigated with the data. (Lamberink, 2009). As LISEM is a spatially explicit model based on a GIS, a spatial map for each different parameter is needed. These spatial maps are based on field measured point data, which need to be interpolated over the catchment area If a correlation is found, it can be used to make maps of the parameters for the catchment, and therefore improve the input maps for LISEM (Limburg Soil Erosion Model), which will make LISEM output more relevant.

1.1 Research Questions

To analyse the relation of the LISEM parameters with the auxiliary maps, the following research objectives are defined:

- Obtain a significant calibration and validation data set to be used in LISEM for the Torrealvilla catchment in SE Spain by collecting data through fieldwork and lab work.
- Analyse data set for significant correlation between the different parameters and three available auxiliary maps (land use/geological/soil).
- Analyse data set for spatial dependency of the parameters.
- Give a prediction for the parameter values for the Torrealvilla catchment in the form of a map based on the correlation and spatial dependency of the parameter.
- Validate the predicted parameters, by comparing the predicted values with the values of the measured validation dataset.

To reach these objectives the following research questions have been defined:

- Can a significant correlation between the LISEM parameters and the land use-, soil- and geological maps of the Torrealvilla catchment be found?
- Are the LISEM parameters spatially dependant?

- In what way can the data points of the different parameters be interpolated to a relevant map of that parameter?

At the start of the research an expectation of correlation of different parameters with landscape features was made based on observations in the field and prior knowledge (Table 1.1). In the field, clear visual correlations are found in parameters crop height and leaf area index. Other correlations are not obvious from field observations, but some hypotheses are defined. Also a devision of the parameters is made by Hessel et al. (2002) into the following categories: vegetation and land use related, soil surface related, erosion and deposition related and infiltration. These categories can also be used to formulate a hypothesis

Surface cover by vegetation (PER), leaf area index (LAI) and crop height (CH) are expected to be dependent on the plant type and cover of an area and therefore have a correlation with land use. Random roughness (RR) is influenced by ploughing of the soil and the type of material and is because of that expected to have a correlation with land use as well as geology. Stone fraction (STFRC) and texture of the soil (D50) are expected to have a correlation with geology, because those factors are dependant on the parent material of the soil. Cohesion (COH) of the soil and soil moisture contents (θi and θs) depend on the storage capacity of the soil which could be related to geology, so these parameters are expected to correlate with that.
The depth of the soil is expected to correlate with the soil map. The crust is expected to depend on the geology of the area and on the land use, because crust is removed by ploughing.

Table 1.1: Expected relation (grey fields) of the parameters to the landscape features at the start of the study

Parameter	Code	Unit	Geology	Soil	Land use
Vegetation and land use related					
Fraction surface cover by vegetation and litter	PER	%			
Leaf Area Index	LAI	-			
Crop height	CH	cm			
Soil surface related					
Random roughness	RR	cm			
Stone fraction cover on surface	STFRC	-			
Crust	CRUST	-			
Erosion and deposition related					
Cohesion of the soil	COH	kPa			
Median of texture of suspended matter	D50	µm			
Infiltration related (Green & Ampt layer 1 option)					
Initial moisture content	THETAI	Vol%			
Saturated moisture content	THETAS	Vol%			
Soil depth	SOILDEPTH	cm			

1.2 Context of the research

This research is a minor Master thesis at the chair group Land Degradation and Development at the Wageningen University. The subject of this thesis fits within the PhD research of Jantiene Baartman MSc.
The research has collaborated with the minor Master thesis of Kirsten Lamberink who has set up a sampling scheme for the field work in the research area, of which the data is used in this research. Earlier research on the modelling of erosion with LISEM in the Torrealvilla catchment in South East Spain has been done by Maarten Lammers (2009) in the form of a master thesis, which was a project within the research of Jantiene Baartman as well.

In this reseach, a data set will be obtained for LISEM through fieldwork in the Torrealvilla catchment in South East Spain and lab work afterwards to process samples taken in the field. The fieldwork will be done according to the field work set up made by Lamberink (2009).
Afterwards the data set will be analysed on different aspects of the statistical significance of the dataset and the dependency between the different parameters and the landscape features (geology/ land use/ soil) described in this thesis.

1.3 Area description

The research area of this study is the Torrealvilla catchment (Rambla de Torrealvilla), which is located in the South East of Spain (Fig 1.1). The Torrealvilla is a tributary of the Guadalentin river upstream of Lorca in the province of Murcia in South East Spain. The catchment has an area of about 250 km^2, and an altitude varying from 350 m above sea level at the outlet of the Torrealvilla to 1530 m above sea level in the northern and eastern mountains (Fig 1.2).

The catchment is situated on the northern horst of the Guadalentin depression and consists of a series of plateau levels with marls and Quaternary gravels (Hooke et al*.,* 2000; Bull et al*.,* 1999), and of mountain ranges in the northern and eastern reaches of the area.
In the northern part of the area the geology consists mainly of limestones and marls in the mountains (the Sierra Cambrón in the north and the Sierra Espuña in the northeast) and glacis areas with gentle slopes on the lower parts of these ranges. The northwestern part of the catchment contains mostly limestone and marl hills. The northern and southern part of the catchment is divided by a calcarenite ridge in the centre of the catchment which is south-west, northeast oriented (Fig 1.2, 1.3).
The southern part of the area contains mainly marls and conglomerates with a mountain range containing mainly limestones and calcarenites in the southeast of the catchment (the Sierra de la Tercia). In this part of the catchment the Torrealvilla is deeply incised into the marls, creating channels with steep and high side slopes.

In the catchment most soils are Cambisols on the flat areas, and Lithosols on the steeper slopes. Near the tributary (Rambla) Fluvisols are found (Fig 1.4). The area has mostly shallow soils, with deeper soils in the valleys.

The research area has a Mediterranean semi-arid climate with dry summers and rainfall concentrated in autumn and spring. The average temperature of the area is around 17 ℃ (Hooke and Mant, 2000). The Guadalentin basin has an annual rainfall varying between less than 300 mm to more than 500 mm per year (De Vente et al*.,* 2008). The smaller rainfall events produce runoff only on local slopes, while runoff only reaches the channels during larger rainfall events. Even these larger flows may not continue through the entire catchment area, due to infiltration in the channel floor and evaporation. (Hooke and Mant 2002; Bull et al., 1999). Very large floods have a small likelihood to occur in the catchment area, due to the localized scale of most rainfall events. When they do occur, they can cause flash floods that last for two to three hours (Bull et al., 1999). The discharge peaks of these events are very short and intense and can cause a lot of damage and carry a lot of sediment (Bull et al*.,*1999, Hooke and Mant 2002).

Land use in the Torrealvilla basin is farm land where cereals are grown, and olive- and almond orchards (Fig. 1.5). On the flatter part of the catchment irrigated land with lettuce and broccoli and non-irrigated cereals are present, while the orchards are located on the steeper slopes. On the steepest slopes matorral- (shrubland) and forest areas are present, these are also planted on previously bare areas to prevent erosion. Two locations are present where the rock conglomerate is mined for industry.

In the northern part of the catchment area vineyards are present. Also new forms of land use like solar panels are present in the area. Although still small in total size, these last two land use types of land use have been rapidly increasing over the last few years. Both vineyards and solar panels are not included in the present research.

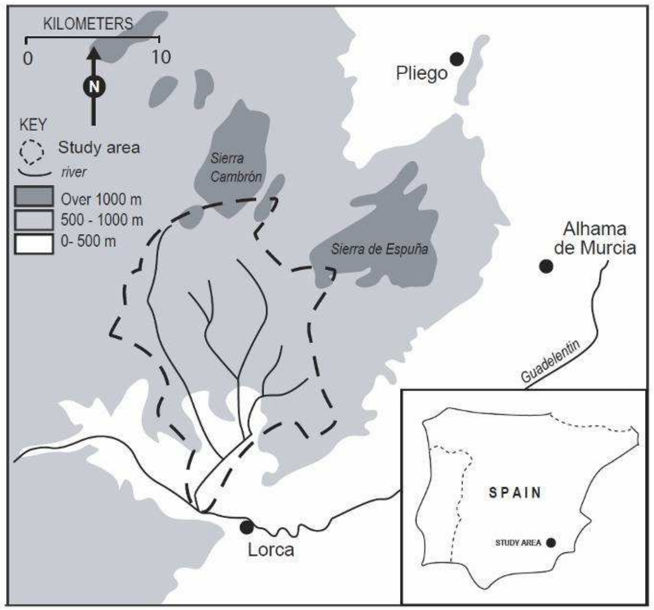


Fig 1.1: Location of the fieldwork area, after Bull et al 1999

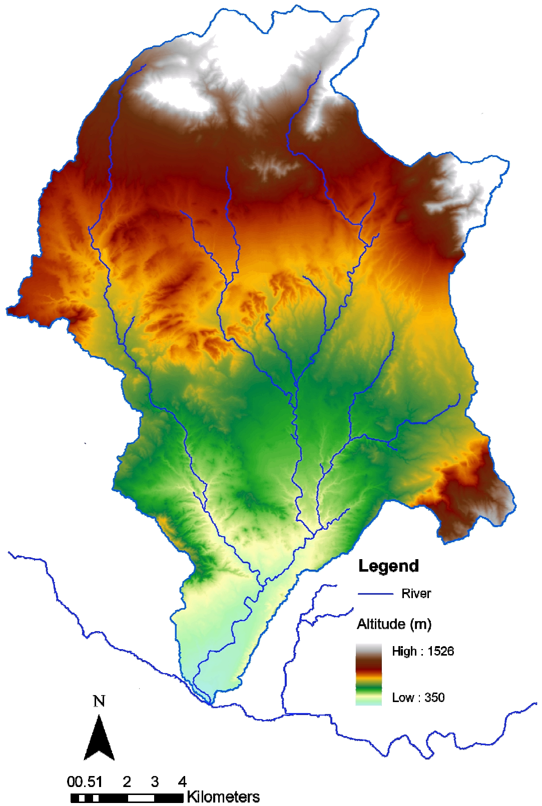


Fig 1.2: Digital elevation model of the Torrealvilla catchment

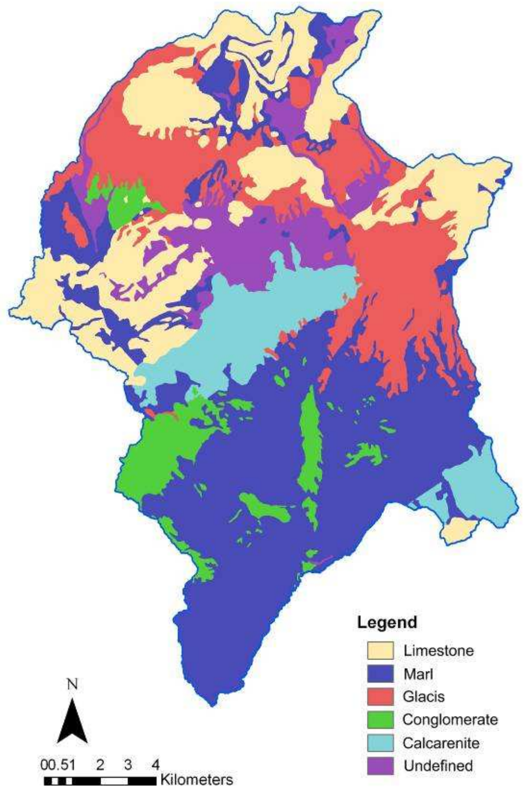


Fig 1.3: Geological map of the Torrealvilla catchment (IGME, 1981), adjusted by Lamberink (2009)

*Fig 1.4: Soil map (*LUCDEME, 1988*) of the Torrealvilla catchment , adjusted by Lamberink (2009)*

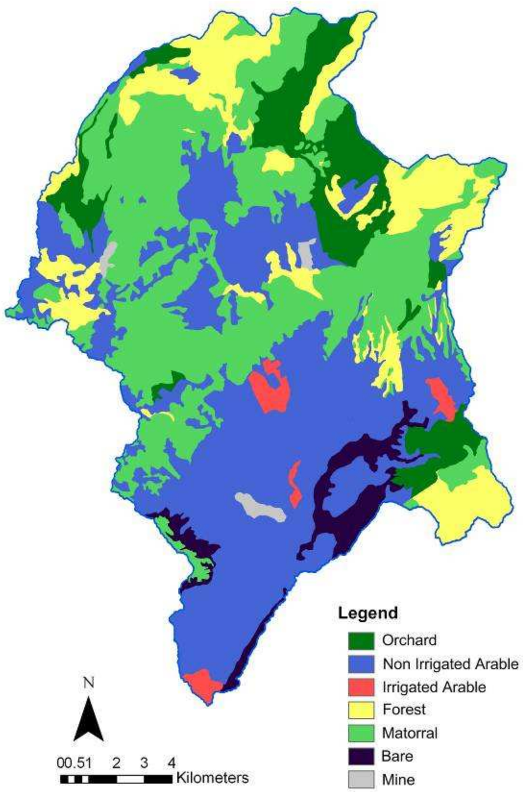


Fig 1.5: Land use map (CORINE Land Cover 2000) of the Torrealvilla catchment , adjusted by Lamberink (2009)

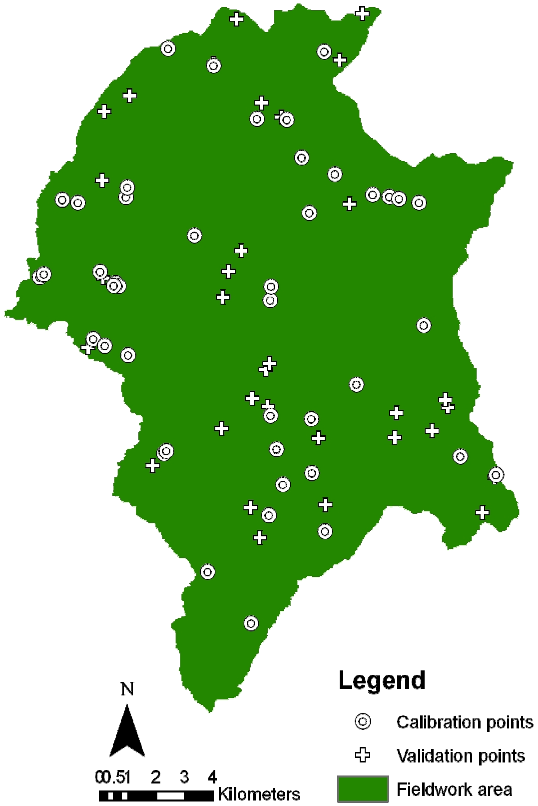


Fig 2.1: Locations of the data points from the fieldwork, September 2009

2. Materials and methods

2.1 LISEM

This research looks at the relation of the LISEM (Limburg Soil Erosion Model) parameters with the landscape features (soil, geology and land use). LISEM is a single event physically based hydrological and soil erosion model. LISEM simulates a catchment during and immediately after a rainfall event and predicts where in the catchment erosion and sedimentation occurs (de Roo et al*.,*1996).
The model is incorporated in a GIS, and accordingly most input data for LISEM are in the form of spatial maps. This makes modelling of larger catchments easier as spatial differences in the catchment are taken into account. The spatial and temporal variability of a rainfall event can also be taken into account through a combination of a rain gauge map with the time series of the individual points. (de Roo et al*.,* 1996a; Jetten, 2002).
Because LISEM is a physically based model, a lot of parameters are required for the calibration of the model, compared with other models that use for example regression equations. Most of these parameters need to be measured in the field, a few (like Mannings n) are taken from literature. LISEM divides the required parameters are into categories: vegetation and land use related, soil surface related, erosion and deposition related and infiltration related (Hessel et al*.,* 2002*)*.

As LISEM is a spatially explicit model based on a GIS, a spatial map for each different parameter is needed. These spatial maps are based on field measured point data, which need to be interpolated over the catchment area. When a parameter has a correlation with one of the auxiliary maps, a spatial prediction of the parameter can be made by taking an average value for each landscape unit of the auxiliary map. When the parameter is related to more than one landscape feature, linear regression will be used to combine correlation with the significant auxiliary maps. Also the role of spatial dependency in the correlation of data points will be taken into account, as well as the possibility to combine regression with Kriging. If a significant correlation is found, the auxiliary maps can be used to extrapolate the parameter data to spatial maps, that can be used as input for LISEM.

2.2 Fieldwork

The fieldwork sampling design was developed by Lamberink (2009) with the research questions of this project in mind. Input parameters were measured for LISEM in the fieldwork area (Fig 2.1).

For the calibration dataset, the fieldwork area was divided into 43 subareas. This was done according to the different combinations of landscape features (geology, soils and land use) present in the area, based on the available geological, soil and land use maps. The areas which had the same combination in soil, land use and geology, were taken together to form one subarea. That means that some of the defined subareas consist of a few unconnected areas.
For the calibration dataset, one representative location was selected during the fieldwork in each of these subareas, for the measurements of the parameters in that area. During the fieldwork, the original maps were checked in the landscape and adjusted where necessary.

For the validation dataset, 32 measurement points were sampled in total. Two points were sampled in each landscape feature class. For validation, the locations of the sampling points were selected randomly. In the case that the landscape feature class of the location was different than expected, another data point for that class was taken. When possible, the deviating data point was used as a data point for an other class that was not yet taken in the field.

For each measuring location (both calibration and validation) the following parameters were measured: Aspect and slope, plant cover, stone fraction, crusting of topsoil, crop height, leaf area index , random roughness, cohesion of the topsoil, soil depth, median grain size and initial- and saturated soil moisture content. Used methods to measure these parameters are explained in section 2.3.

State of the area during measurements

The data points were taken in the period of September 23 to October 19 2009.
Besides the data points used for this study, also rainfall data was recorded and discharge data was measured in the field during the fieldwork period. These data are used by Jantiene Baartman and are not part of this research. Some of the measurements can be influenced by weather conditions. Rainfall occurred on the 21st and the 28th of September and on the 13th of October, which resulted in discharge of the Prado and the Torrealvilla on the 28th of September and in discharge of only the Prado on the 13th of October, as well as wet soils and soil surfaces for a few days after the rain.
Cereals were already harvested at the start of the fieldwork, and at the end of the fieldwork period these fields where being ploughed. The irrigated crops were fully grown when measured; the last two days of our fieldwork harvesting of these crops began.

2.3 Measuring methods

Measurement of the LISEM input parameters in the field and processing of the data in the lab were done according to Hessel (2002). Following an explanation of the parameters and the method of measuring and processing is given.

Coordinates and elevation

The coordinates and elevation of each location was recorded using a Garmin eTrex Legend GPS. The projection used was "European Datum-1950; UTM Zone 30N"

Aspect and slope

At each location the aspect and the slope of the landscape was recorded with a compass and a clinometer respectively.

Plant cover (PER)

The fraction of plant cover [%] of the surface was estimated on each measurement location. This parameter is the fraction of cover of plants or plant materials on the ground and is used in LISEM to calculate the interception. In the estimation, leaves and other plant material lying on the ground were taken into account, as well as the low plants. Higher plants like trees where left out of this estimation, because these are taken into account with the leaf area index (see below).

Stone Fraction

The stone fraction [-] is the fraction of the surface that is covered with stones and is estimated in the field, similar to the plant cover. In LISEM, for the fraction covered with stones, no splash erosion is calculated (Hessel, 2002).

Crust

Crusting of the topsoil is categorized in 4 categories from zero to three, reference for these categories are shown in Fig 2.2.

0 = no crust
1 = a little crusting is visible, it is not possible to separate pieces of crust, without breaking them apart
2 = clear crusting, pieces of crust can be lifted from the ground
3 = very clear crusting, crust is thicker than 5 mm

Fig 2.2: Reference for categories Crust (left to right: category 0 to 3)

Crop Height (CH)

The crop height [cm] is estimated by measuring the height of the different plants in the area and estimating a representative height for each plant species. With the height of the plants, the amount of plants counted in an representative area and the area of the plant from above, a weighted average of the crop height can be calculated for the area (eq 2.1). The same area as used for the leaf area index is taken as representative for this parameter (see section Leaf Area Index). The crop height is used in LISEM to calculate through fall kinetic energy for splash erosion Hessel (2002).

$$CH_{Average} = \frac{\sum (CH_{plant} * n * Area_{plant})}{\sum (n * Area_{plant})} \qquad \text{eq. 2.1}$$

Where:

$CH_{Average}$ = Estimated average crop height over a representative area [cm]
CH_{plant} = Height of a plant species [cm]
n = Number of plants in the representative area [-]
$Area_{plant}$ = Area of a plant species from above [m^2]

Leaf Area Index

The Leaf Area Index (LAI) [-] is the fraction of the combined area of plant leaves per square meter. This fraction can be larger than one in the case of multiple layers of leaves. This factor is estimated in the field as an average over an area of representative size. LISEM uses the LAI to calculate water storage (Hessel, 2002).
On each measuring location the amount of different plants are counted in this selected area. The heights of the measured plants are divided into categories for each plant species, to reduce the amount of plants for which the leaves have to be counted.

For each category of plant-height, the number of leaves or the number of branches is counted, or estimated in the case of large plants. Also a sample of leaves, branches, fruit and/or flowers are photographed. From these photographs an average area is derived.
With the following formula the leaf area index was calculated.

$$LAI = \frac{\sum (Area_{leaf} * n_{leaf} + Area_{flower} * n_{flower}) * n_{plant}}{Area_{rep}}$$ eq. 2.2

Where:

LAI	= Leaf Area Index [-]
$Area_{leaf}$	= average area of a leave [m^2]
n_{leaf}	= number of leaves per plant [-]
$Area_{flower}$	= average area of a flower [m^2]
$n_{flovers}$	= number of flowers per plant [-]
n_{plant}	= number of plants in the representative area [-]
$Area_{rep}$	= representative Area [m^2]

Random Roughness

The Random Roughness (RR) [cm] is the standard deviation of the height of the surface on a small scale, and is used in LISEM to calculate the overland flow (Hessel ,2002). This was measured on a line of 60 cm using a pin meter (Fig 2.3), with the pins at 1.5 cm distance from each other. The pin meter works in such a way, that the profile of the surface is reproduced by the pins. A photo was made from these profiles on each field location, and the standard deviation of the height of the pins was determined from this photo using “pmpproj” (Kilpelainen, no date) according to Wagner et al. (1991) and Jester et al 2005.

Fig 2.3: Pin meter for measurements of Random Roughness

Fig 2.4: Pocket vane tester for cohesion measurements

Cohesion

The cohesion [kPa] is the rotational shear stress of the soil surface (Hessel, 2002). Cohesion is measured with a Pocket vane tester, by Eijkelkamp Agrisearch Equipment, the Netherlands (Fig 2.4) on wet soil. In LISEM, cohesion is used to calculate the erosion caused by overland flow.

Saturated soil moisture content

The saturated soil moisture content (θ_S) [Vol%] is used in LISEM to calculate the saturated conductivity of the soil (Hessel, 2002). It is measured as the ratio of weight in an oven dried soil sample and the weight of that sample when saturated and is calculated with eq. 2.3. Samples were dried at 105ºC for at least 24h. Samples were taken with 100cm^3 sampling rings.

$$\theta_s = \frac{M_{saturated} - M_{dry}}{V \times \rho_{water}}$$ eq. 2.3

Where:

θ_S	= saturated soil moisture content [Vol%]
$M_{saturated}$	= mass of the soil sample when saturated [kg]
M_{dry}	= mass of the soil sample when dry [kg]
V	= volume of the soil sample [m^3]
ρ_{water}	= water density [kg/m^3]

Initial soil moisture content

The initial soil moisture content (θ_i) [Vol%] is measured using two different methods. One method uses the same sample rings as the saturated moisture content, but here the ratio between the initial weight of the sample and the dry weight of the soil sample was calculated (eq. 2.4).
The second method of measuring the initial moisture content was using a Trime TDR Probe (Time Domain Reflectometry Probe), by IMKO, Germany which measures the initial moisture content directly, using electrodes (Hessel, 2002).
The initial moisture content is used in LISEM to calculate the, storage of water in the soil.
The initial soil moisture content is the volume percent of moisture in the soil at the moment of measuring.

$$\theta_i = \frac{M_{initial} - M_{dry}}{V \times \rho_{water}}$$ eq. 2.4

Where:

θ_i	= initial soil moisture content [Vol%]
$M_{initial}$	= initial mass of the soil sample [kg]
M_{dry}	= mass of the soil sample when dry [kg]
V	= volume of the soil sample [m^3]
ρ_{water}	= water density [kg/m^3]

The TDR data was used for the initial soil moisture content data, because not all ring samples were completely dry when measured.

Median grain size

The median grain size (D50) [μm] is used in LISEM to calculate the volumetric transport capacity, which relates to overland flow (Hessel, 2002). The median grain size of the mineral particles is determined from a soil sample using a laser particle sizer (A22-c-version; Fritsch GmbH). Which is based on the principle that a certain sized particle diffracts light at a specific angle. The laser particle sizer measures the light diffraction of the sample, from which the grain sizes can be calculated (Vu website, no date; Konert and vanderBerghe, 1997; Beuselinck et al., 1998). Before the sample is analysed, the organic matter is removed from the sample using a 30 % HCL and a 30% H_2O_2 solution.

Soil Depth

The depth of the soil [cm] is measured with a gauge auger of 30 cm. The depth is measured three times at each measured site.

2.4 Statistical analysis

The objective of this research is to analyse the dependency of the LISEM parameters on the different landscape units, after which a relevant interpolation of the parameters can be created for the catchment using this dependency. For the statistical analysis of the data dependency, the programs SPSS and R-statistics were used. To make a prediction for the parameters, a linear regression model was made for each parameter, with the landscape feature classes (geology, soil and land use) as dependent factors. The landscape features that were included in the model were chosen based on the variance between the landscape feature classes for the parameter using the ANOVA test, or the Kruskall-Wallis test for the non-parametric data. Because according to Field (2009), parametric tests are more accurate than non-parametric tests, the parametric tests where used when possible. When it was not possible to use parametric tests because the assumptions are not met, the non-parametric tests are used.

To decide whether to use an ANOVA or a Kruskall-Wallis test, first the data need to be checked for the assumptions of the ANOVA (see section 2.4.1).

For the parameters that showed a relevant spatial dependency (based on their semivariogram), a combination of the regression model and Universal Kriging was performed. For the parameters that showed no significant correlation with any of the landscape features, Ordindary Kriging was used for interpolation of the data.

2.4.1 Assumptions

For the statistical analysis the program SPSS was used. To know whether to use parametric (ANOVA) or non-parametric (Kruskall-Wallis) tests, the assumptions of these tests have to be checked.
The assumptions for the parametric statistical tests, such as ANOVA, are (Field, 2009):

- *Normality:* The sampling distribution is normally distributed. It is assumed that the sampled data has the same distribution as the total distribution of the variable. Therefore it is assumed that, if the sampled data has a normal distribution, the total data distribution will be normal as well.
- *Homogeneity of variance:* which means that the variances should be the same throughout the data.
- *Interval:* The data should be measured at an interval which picks up the variation in the data that you are interested in.
- *Independence:* the individual data points should not depend on each other.

The interval and independence of the data are assumed to be correct, because these assumptions were taken into account in the data collection and sampling design, and can not be checked further. To check whether the assumptions for normality and homogeneity of variance are met the following tests are done on the data. The results of these tests determines which subsequent test should be used to analyse the variance between the groups.

In this research, not all the parameter data has a normal distribution. In some situations it is possible to transform the non-normal data to into normal data, in order to be able to perform the parametric tests (Field, 2009), but that is not possible for the data in this research. Transformation is only possible when all the data are deviating from normality in a similar way, because all the data needs to be transformed in the same way, in order to be able to compare the data. Because, in this case, the data was deviating from normality in different way's for the different parameters, it was not useful to transform the data.

Skewness & Kurtosis

The values for the skewness and kurtosis of a distribution are an indication of the normality of the distribution, and can be used to make a decision whether to use parametric or non-parametric tests. The z-scores of the skewness and kurtosis are 0 at normal distributions. Positive values of skewness indicate a pile-up of scores on the left part of the distribution, while negative values indicate a pile-up on the right. Positive values of kurtosis indicate a pointy and heavy tailed distribution, while negative values indicate flat and light tailed distributions. Values for skewness and kurtosis can be transformed to z-scores (eq. 2.5 and 2.6) to be able to compare the skew and kurtosis in different samples. (Field, 2009).

The transformation of values to z-scores are as follows (Field, 2009):

$$z_{skewness} = S - M / SE_{skewness} \qquad \text{eq. 2.5}$$

$$z_{kurtosis} = K - M / SE_{kurtosis} \qquad \text{eq. 2.6}$$

Where:

z	= z score for skewness or kurtosis
S	= unadjusted value for skewness
K	= unadjusted value for kurtosis
SE	= standard error
M	= mean of the z-score, which is zero.

Kolmogorov-Smirnov

The K-S test is a second way to find out whether a dataset has a normal distributions, it compares the scores of the data to a normally distributed set of scores, that has the same mean and standard deviation as the data. The null hypothesis of this test is that the data differs from the normal distribution. If the test is not significant at $p=<0.05$ (when $p>0.05$), the null hypothesis is valid and the distribution is not significantly different from the normal distribution. If the test is significant ($p<0.05$), that means the null hypothesis is incorrect and the distribution has a significant deviation from the normal distribution (Field, 2009). The test statistic gives the largest difference of the model (here the normal distribution) to the data as a fraction (Davis, 1986).

Levenes test

Levenes test tests for the assumption of homogeneity of variance, with the null hypothesis that the variances of the groups are equal, by calculating the absolute difference between each score and the mean of the group from which it came. If the result of the test is significant at $p =< 0.05$, it can be concluded that the null hypothesis is incorrect and the variances of the groups are not equal. If the test is insignificant with $p > 0.05$, the null hypothesis is correct and the variances of the groups can assumed to be equal (Field, 2009).

2.4.2 Regression

If a relation between a parameter and a landscape feature exists, the measurements of the parameter for different landscape feature classes are significantly different from each other. So to check whether a landscape feature has a correlation with a parameter, the variance between the parameter values in the different classes of a landscape feature can be analysed.
To check if there is a significant difference between the different landscape feature classes for a parameters, an Analysis of Variance is done for the parameters that meet the assumptions described in the previous paragraph. The other parameters are investigated using the Kruskall-Wallis test.

ANOVA

The Analysis of Variance (ANOVA) test compares the variance between more than two groups of data. It is used to test the hypothesis that the different groups of data are identical (Davis 1986). In the one-way analysis, the variance of the distribution within the different categories are compared with the variance between the different categories (Davis 1986). So the test is a measure of how much of the variance in the data can be explained by the different categories. When the variation within the group is significantly smaller than the variation between the groups, the groups are significantly different from each other. When the variation within the group is larger than the variation between the groups, the groups are not significantly different from each other.
The test statistic of the ANOVA is the F ratio (eq. 2.7), which compares the amount of systematic variance in the data (the variance between the different groups) to the unsystematic variance (the variance within the groups). In other words, The F-ratio is the ratio of the model to its error (Field, 2009). If the assumption of homogeneity of variance is broken, Welch 's and Brown-Forsythe's F can be used, because they correct for this assumption (Field, 2009).

$$F = \frac{MSM}{MSR}$$ eq. 2.7

Where
MSM = mean squares of the model; the variance among the groups of samples
MSR = residual mean squares; the variance within the groups of samples

For this research, the ANOVA is performed for the data, sorted by landscape feature. If the test has a significant outcome, the values for that parameter differ per landscape feature class, for the landscape feature the data was sorted by, and a correlation is found between the landscape feature and the parameter.

Kruskall Wallis

The Kruskall Wallis test is a non-parametric test, testing the correlation of several samples and is the non-parametric alternative to the ANOVA. (Davis 1986, Field, 2009). In this test, the observations are ranked, from smallest to largest value of the parameter. The rank-numbers are summed for each category, after which the test statistic H can be calculated (eq. 2.8).
The null hypothesis for this test is that all populations have identical distributions.
An assumption in the Kruskall Wallis test is that all values have been taken randomly, and that the samples are independent from one another (Davis 1986).

From the sum of ranks the H-statistic can be calculated according to:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$ eq. 2.8

Where
R_i = sum of ranks for group i
n_i = number of data points in group i
N = total sample size
k = number of groups

The test statistic H is approximately distributed as a chi-squared distribution with k-1 degrees of freedom, and is, like ANOVA's F-ratio, the ratio of the model to its error (Davis 1986, Field 2009).

Linear regression

The landscape features that have a significant variance within the different categories can be used to make a prediction of the parameter. An approximation of the parameters can be made with a linear regression model.

For the initial analysis of correlation between the landscape features and the measured parameters a linear regression between these two is calculated in SPSS using the following formula (Davis 1986, Field 2009):

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0); \quad q_0(s_0) \equiv 1,$$ eq. 2.9

Where
$\hat{z}(s_0)$ = predicted value at location s_0
$\hat{\beta}_k$ = estimated linear model coefficient of the k'th parameter predictor
$q_k(s_0)$ = values of the k'th parameter predictor value at location s_0
p = number of predictors of the auxiliary values

The assumptions for this prediction are:

- The error term has a normal distribution with a mean of zero
- The value of the error term is independent of the values of the variables and of the values of the other error terms.
- The variance of the error term is constant across the cases and independent of the variables (Field, 2009).

The coefficients (β) are calculated using the OLS (Ordinary Least Squares) in SPSS and R-statistics, using the following formula

$$\beta_i = \frac{\sum (x_i - x_{avg})(y_i - y_{avg})}{\sum (x_i - x_{avg})^2} \qquad \text{eq. 2.10}$$

Where:

β_i	= estimated linear model coefficient
x_i	= value of the x parameter
x_{avg}	= average value of the x parameter
y_i	= value of the y parameter
y_{avg}	= average value of the y parameter

The regression analysis in SPSS gives the value of the coefficients of the model as well as the R^2 and the adjusted R^2 (For explanation of R^2 see paragraph 2.4).

The linear regression analysis is calculated for the various combinations of landscape features (i.e. a combination of soil and land use and geological units or only one or two of the three depending on the ANOVA or Kruskall-Wallis tests), to analyse the dependency of each of the landscape features. The landscape features that are used in the prediction model for the parameters are based on the results of the ANOVA and on the significance of the correlation based on the R^2 of the linear regression.

Because the predictors in this case are categorical, dummy variables with the value 0 or 1 are used to be able to use linear regression. The amount of dummy variables is the amount of categories in a predictor value, minus one. In the case of a dependency, for example, only on land use, eq. 2.11 has the following from:

$$\hat{z}(s_0) = \beta_0 + \beta_1 * LU_1 + \beta_2 * LU_2 + \beta_3 * LU_3 + \beta_4 * LU_4 + \beta_5 * LU_5 + \beta_6 * LU_6 \qquad \text{eq. 2.11}$$

Where:

$\hat{z}(s_0)$	= predicted value at location s_0
β	= estimated linear model coefficient
LU_n	= dummy value 1 or 0 for land use n

With all LU_k's having the value of 0, except for the land use type which is present at $z(s_0)$, which has value 1.

Because seven land use types are present in the research area, six are used for the linear regression model. For a model based on a combination of landscape features, a similar formula is given. In this case each landscape feature has one dummy with the value one, and the rest with the value zero.

2.4.3 Kriging

Because the data has a spatial configuration it is useful to assess its spatial dependency, and look at the possibility to incorporate this in the prediction for the parameters.
In case of no significant correlation between the different landscape features and a parameter, there might be a significant spatial dependency that can be used to create a relevant map for the parameter. Ordinary Kriging can be used to create an interpolation in this case.
In case of a significant correlation with one or more of the landscape features and a spatial dependency, a combination of regression and kriging can be made when a significant spatial dependency is (still) present in the residuals of the data. This dependency is used in the prediction for the parameter using Regression Kriging. The Kriging is done with the program R-statistics (R Development Core Team, 2010). Used scripts are given in Appendix C.

Ordinary Kriging

To asses the spatial dependency of the data, the semivariogram of the data is estimated according to the following formula (Davis, 1986):

$$\gamma_h = \frac{\sum_{i}^{n-h} (X_i - X_{i+h})^2}{2n}$$ eq. 2.12

Where:

γ_h	= semivariance for distance h
X_i	= measured value at location i
X_{i+h}	= measured value at distance h from i
n	= the number of points

This formula calculates the average of the squared difference between two values at distance h. With this estimated semivariogram, the value of a parameter can be calculated at non-visited locations using the known locations around it.

For Ordinary Kriging the following formula is used:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i \cdot z(s_i),$$ eq. 2.13

Where:

$\hat{z}(s_0)$	= predicted value at location s_0
λ_i	= kriging weights, depending on the spatial correlation
$z(s_i)$	= known value at location s_i
n	= number of data points

Regression Kriging

The combination of regression and kriging is done according to Hengl (2004; 2007). The value of a parameter is estimated by summing the drift of the residuals with the predicted drift (eq 2.14). Drift is the deviation from the prediction for each x. In this case, the drift is fitted with a linear regression model and the residuals are interpolated using Universal Kriging, according to the following formula, which is a combination of equations 2.13 and 2.9 (Hengl, 2004):

$$\begin{aligned}\hat{z}(s_0) &= \hat{m}(s_0) + \hat{e}(s_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} \lambda_i \cdot e(s_i)\end{aligned}$$ eq. 2.14

With $q_0(s_0) = 1$

Where:

$\hat{z}(s_0)$	= predicted value at location s_0
$\hat{m}(s_0)$	= predicted drift
$\hat{e}(s_0)$	= predicted residual
$\hat{\beta}_k$	= estimated linear model coefficient of the k'th parameter predictor
$q_k(s_0)$	= values of the k'th parameter value at location s_0
p	= number of predictors of the auxiliary values
λ_i	= kriging weights, depending on the spatial correlation
$e(s_i)$	= residual at location s_i
n	= number of datapoints

In the case that both regression and kriging give no significant results, the best representation of the parameter is the mean of the observations.

2.5 Validation

The separate validation dataset that was measured in the field, is used to validate the results of the regression and kriging analysis. The measured values of the validation dataset is compared with the predicted values on the same location using the coefficient of determination or efficiency (Field, 2009). This value compares the observed validation data with the predicted values for the parameter and gives an indication of the efficiency of the used model.

$$R^2 = EF = 1 - \frac{\sum_{i=1}^{NumObs} (OBS_i - MOD_i)^2}{\sum (OBS_i - MOBS)^2}$$

Where:

OBS_i	= observed value of parameter on location i
MOD_i	= Modeled value of parameter on location i
MOBS	= Mean of observed values

The R^2 gives the fraction of variance in the parameter of the sampled data, which is accounted for by the predictor. It has a range of 0 to 1 for significant values, where the value 1 means that 100 percent of the variance in the dataset can be explained by the model and a value of 0 means that all the variance in the dataset can be explained by the mean. Is the value of the efficiency is below zero, it means that the mean of the data is a better model for the data than the model.

The adjusted R^2 gives the fraction of variance in the parameter, which is accounted for by the predictor, over the total area. The adjusted R^2 is a relevant indication of the significance of the correlation, because it is estimation for the significance of the correlation for the entire area instead of only the sampled locations. SPSS derives the adjusted R^2 using Wherry's equation (Field, 2009).

3. Results

3.1 Data

The data as gathered in the field and processed according to Chapter 2 (Material and methods) are shown as a histogram in figure 3.1 for the parameters soil depth and leaf area index, the rest of the parameters are shown in appendix A. It shows from the histograms that not all parameters are normally distributed. To answer the question which landscape features are correlated with which parameters, the variance between the different classes of the landscape features have to be analysed. This can be done with various statistical tests which is shown in the following paragraphs. A few outliers of the data are found while exploring the dataset for normality. The outliers that had a known reason to differ from the dataset where excluded from the data.

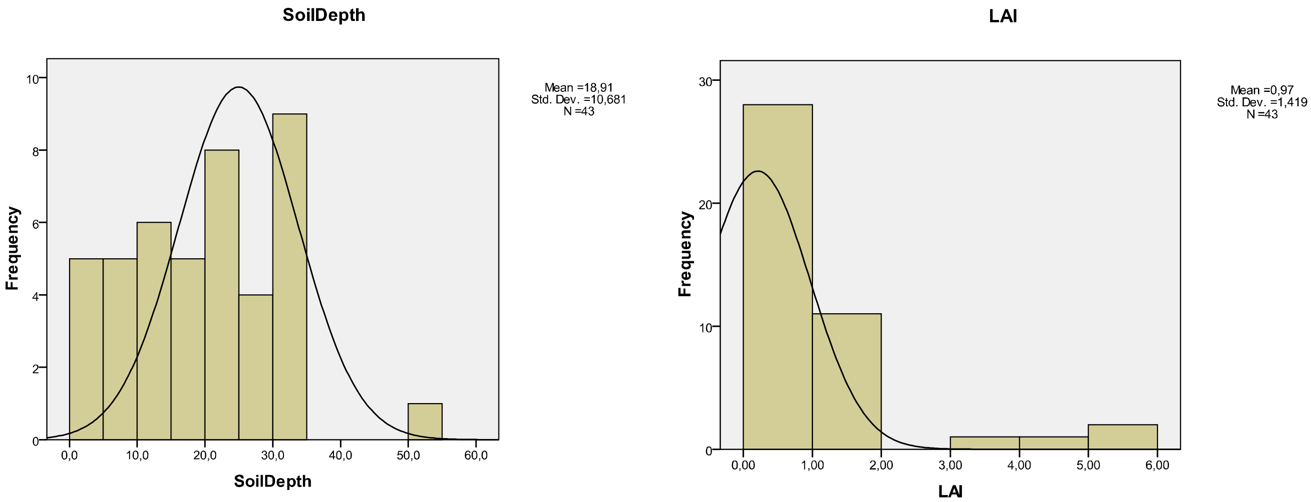


Fig 3.1: Histograms of the calibration data for the parameters soil depth and leaf area index

3.2 Checking assumptions

To decide whether to use parametric or non parametric tests for the correlation analysis, the assumptions for parametric tests are considered.

3.2.1 Normality

The normality of the data is checked per parameter by calculating the z-scores of skewness and kurtosis of the distribution and by performing Kolmogorow-Smirnov test. If both tests result in a significant results at $p \leq 0.05$, the data are assumed to be normal, if that is not the case non-parametric tests are performed on that parameter.

Skewness and kurtosis

Results of skewness and kurtosis are shown in table 3.4 and Appendix B. Values of the z-scores for both the kurtosis and the skewness of the distribution are below the absolute value of 1.96 for the parameters: crust, Soildepth, initial soil moisture (Theta I), cohesion (COH) and surface plant cover (PER). For these parameters the values of the z-scores for skewness and kurtosis are significant at $p<0.05$, which means the distribution can be considered normal. The other parameters have higher absolute z-scores for the skewness, the kurtosis or both, which means it can not be assumed that these parameters have normal distributions.

Kolmogorov-Smirnov
The results for the Kolmorgorov-Smirnov test (table 3.4 and appendix B table B.2) is that the surface plant cover (PER), cohesion (COH), initial soil moisture (Theta I), soil depth, and saturated soil moisture (Theta S) have non significant values ($p \geq 0.05$) so their distributions can assumed to be normal. The other parameters have a significance lower than 0.05, which means these can not assumed to be normal.
The results of this test do not entirely coincide with the results based on skewness and kurtosis. The distribution of the crust has normal values for skewness and kurtosis, but non-normal values for the Kolmogorov-Smirnov test. The parameter saturated soil moisture (Theta S) did not have a normal distribution according to the Skewness and the Kurtosis, but did have a normal distribution according to the Kolmogorov-Smirnov test. In these cases non-parametric tests are used for the parameters.

3.2.2 Homogeneity of variance

Levene's test is done to check the homogeneity of variance in the distribution. The test can not be performed for land use, because there are not enough data points for each land use type to make enough pairs of spread and levels. Therefore it is assumed that, if the data is homogeneous for both geology and soil, the data has a homogeneous distribution for land use as well.

The results of this test (appendix B, table B.3) show that the parameters crust, initial soil moisture (ThetaI), saturated soil moisture (ThetaS), median grain size (D50) and random roughness (RR) have a significant results (at $p \leq 0.05$) for Levene's test when the distribution is sorted by either Soil, Geology, or both, and have therefore not an equal variance over the whole distribution. Because the assumption for normality is met in the case of the parameter Theta I, an adjusted F statistic can be used to compensate for the lack of homogeneity in the variance, while still carrying out the ANOVA. In the following table (table 3.4) a summary of the assumption checks for parametric tests are shown, with the appropriate subsequent test for the parameter.

Table 3.4: Summary of checked assumptions and the appropriate test to analyse the variance in the data; when an assumption is met, grey; assumption is broken, white.

Parameter	Normality			Equality of variance	Appropriate test
	Kurtosis	Skewness	K-S test	Levene's test	
PER					Oneway ANOVA
COH					Oneway ANOVA
Soil Depth					Oneway ANOVA
Theta I					Oneway ANOVA, Welch and Brown-Forsythe's F
Crust					Kruskal-Wallis test
LAI					Kruskal-Wallis test
CH					Kruskal-Wallis test
STFRC					Kruskal-Wallis test
D50					Kruskal-Wallis test
RR					Kruskal-Wallis test
Theta S					Kruskal-Wallis test

3.3 Regression

In the following part the variation between different landscape feature classes is evaluated for each parameter, using the ANOVA test or the Kruskall-Wallis test.

A oneway-ANOVA is done for the parameters: Plant cover (PER), cohesion (COH), Soil Depth and initial soil moisture (ThetaI), with adjusted test statistic (F) for the initial soil moisture (ThetaI). As can be seen from table 3.5 not all landscape features have a significant variance between the feature classes for the different parameters. The parameter 'plant cover' (PER) only has a significant variance between landscape feature classes for the different land uses at $p \leq 0.05$. The cohesion has at $p \leq 0.05$ a significant variance between landscape feature classes for the different soils, the Soil Depth between the landscape feature classes of both the soils and the land use types. For the initial soil moisture (Theta I), no significant variation between the landscape feature classes are found at $p \leq 0.05$.

For the parameters with a non normal distribution, the Kruskall-Wallis test is done. Results are given in Table 3.6. Stone fraction has a significant result of the test for all three landscape features at $p \leq 0.05$. Saturated soil moisture (ThetaS) has a significant variance between landscape feature classes of both geology and land use at $p \leq 0.05$. Leaf area index (LAI), crop height (CH) and random roughness (RR) all three have a significant variance between the landscape feature classes for the different land uses. The median grain size (D50) had an insignificant variance between all the feature classes at $p \leq 0.05$. A summary of these findings is shown in table 3.7

Table3.5: SPSS output ANOVA, grey values are significant variance (at p ≤ 0.05).

		df between groups	df within groups	F	sig.
PER	Soil	6	35	0.737	0.485
	Geology			0.8	0.557
	Land Use			5.611	0
COH	Soil	6	35	4.03	0.026
	Geology			0.218	0.953
	Land Use			0.363	0.701
Soil Depth	Soil	6	35	26.121	0
	Geology			2.422	0.054
	Land Use			5.097	0.001
Thetal	Soil: Welch	6	36	2.041	0.157
	Soil: Brown-Forsythe			1.405	0.261
	Geology: Welch			0.475	0.731
	Geology: Brown-Forsythe			0.557	0.79
	Land Use: Welch			2.475	0.095
	Land Use: Brown-Forsythe			1.532	0.278

Table 3.6: SPSS output for the Kruskall-Wallis test per landscape feature and parameter, grey values are the significant output (at p ≤ 0.05).

		chi-square	df	asymptotic sig	monte carlo significance
LAI	Soil	2.661	2	0.264	0.270
	Geology	4.253	5	0.514	0.535
	Land use	38.709	6	0.000	0.000
CH	Soil	0.71	2	0.701	0.713
	Geology	6.944	5	0.225	0.224
	Land use	38.378	6	0.000	0.000
D50	Soil	5.154	2	0.076	0.072
	Geology	6.86	5	0.231	0.229
	Land use	7.291	6	0.295	0.307
RR	Soil	0.779	2	0.677	0.684
	Geology	1.457	5	0.918	0.926
	Land use	12.882	6	0.045	0.032
Stfrc	Soil	15.585	2	0.000	0.000
	Geology	23.637	5	0.000	0.000
	Land use	16.27	6	0.012	0.005
ThetaS	Soil	4.249	2	0.119	0.117
	Geology	12.14	5	0.033	0.023
	Land use	13.156	6	0.041	0.022
Crust	Soil	6.535	2	0.038	0.036
	Geology	10.42	5	0.064	0.058
	Land use	12.666	6	0.049	0.036

3.4 Correlation

3.4.1 Landscape features

For the landscape feature/parameter combinations that have a significant variation between the means, the correlation is calculated. To analyse the presence of this correlation, the coefficient of determination (R^2) is calculated for each of the different landscape features as can be seen in table 3.7. The values of this coefficient shows that the landscape feature/parameter combinations that have a low significance in the analysis of variance tests, also have a low correlation. The correlations between the parameters and the landscape features that do have a significant result for the ANOVA and Kruskall-Wallis tests mostly have a significant correlation as well.
The arbitrary value of a correlation coefficient of 0.4 will be considered significant and will be analysed further for significance in the prediction of the parameter. In the case that this value is too low to give a significant correlation, the regression model coefficients will be low and the contribution of the parameter to the model will be low.

Table 3.7: Summary of tables 3.5 and 3.6, grey indicates significant test results. And the correlation between the landscape features and the parameters based on the Adjusted R^2

Parameter	Significant variance at p ≤ 0.05 Based on ANOVA (table 3.5) or Kruskall Wallis (table 3.6)			Correlation based on Adjusted R^2		
	Soil	Land Use	Geology	Soil	Land Use	Geology
Crust				0.16	0.15	0.21
Theta S				-0.03	0.20	0.22
Theta I				0.01	0.13	-0.06
PER				-0.01	0.40	-0.03
LAI				-0.03	0.69	-0.05
CH				-0.03	0.85	-0.04
RR				0.07	0.37	-0.07
STFRC				0.29	0.20	0.44
Soil Depth				0.55	0.38	0.15
COH				0.03	-0.03	-0.03
D50				0.05	-0.02	0.06

3.4.2 Spatial correlation

The semivariograms of most parameters show some spatial correlation (appendix D), but mostly on such a small spatial scale that kriging will have no relevant results (Fig 3.2), because the range of the semivariogram is smaller than the distance between the data points. Spatial dependency is assumed to be relevant if the estimated semivariogram have a range larger than the distance to the surrounding points, which is taken as 2000 meter in this case. In the cases of the parameters soil depth, cohesion and median grain size, the ranges of the semivariograms are significant. (see Fig 3.3 for an example (cohesion), for others see appendix D). In these cases kriging is included in the prediction of the parameter map. The linear regression residuals have mostly a smaller spatial correlation than the semivariograms based on the parameter data.

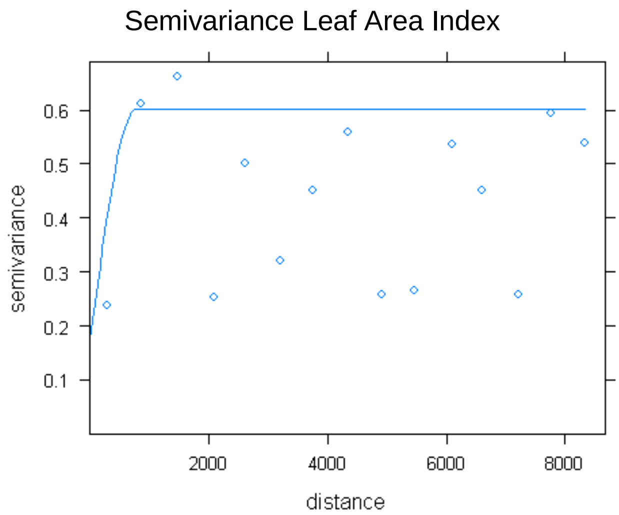


Fig 3.2: Semivariance of parameter Leaf Area Index, range of semivariance is too small to be significant

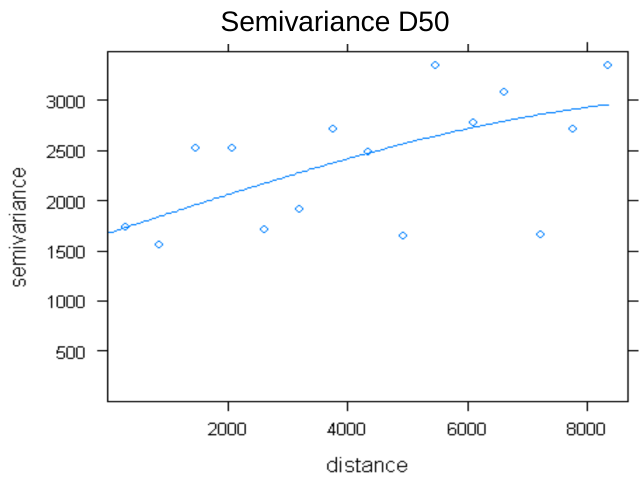


Fig 3.3 Semivariance of parameter D50, rang of semivariance is significant enough to be included in the method.

3.5 Regression and Kriging Results

In the following paragraph, the results of the regression and the kriging are shown. To validate the obtained models for the parameters, a validation dataset is obtained through fieldwork according to paragraph 2.2. These validation points can be used to check the accuracy of the prediction by comparing the predicted value at the data locations with the measured value
The parameters median grain size (D50), leaf area index (LAI) and Soil depth are used as an example, the results of the other parameters can be found in appendix D.

Data of the parameters crust, and the saturated and initial soil moisture content, has no significant spatial correlation, and no significant correlation with the landscape features based on the calibration dataset, so in these cases the average of the parameter measurements is the best predictor.

The cohesion and median grain size has a significant spatial correlation and no significant correlation with the landscape features, so Ordinary Kriging is performed on these parameters (see Fig 3.4 to 3.6 for examples, for the other model results: see appendix D). Figure 3.4a shows the results of Ordinary Kriging for the parameter median grain size (D50), based on the semivariogram shown in Figure 3.3. The comparison with the validation dataset and the modelled parameter prediction (Fig 3.4b) shows that the correlation between the validation dataset and the model results is very low. There is no significant correlation between the model and the validation dataset (Fig 3.4). The coefficient of determination (R^2) of the comparison of the measured values with the predicted values is 0.01, which means that the Ordinary Kriging model based on the calibration points does not predict the whole catchment area sufficiently and the mean of the validation data is an equally good predictor for the parameter values.

The parameters plant cover (PER), leaf area index (LAI) , crop height (CH), random roughness (RR) and stone fraction have no significant spatial correlation, for these cases only regression is performed. The parameters median grain size (D50), Random roughness (RR)and plant cover (PER) have a R^2 between 0.01 and 0.07. The parameters leaf area index, crop height and stone fraction had relatively high correlations between 0.2 and 0.3 (table 3.4). But these correlations are still not significant. In Fig 3.5b the predicted map of the parameter leaf area index shows values ranging from -0.7 to 5.3. The negative values are an anomaly of the regression-method. This method tries to find the best fit for the data and does not take the range of the parameter into account. Also, the parameter leaf area index has a different range of values of the measurements (approx 0 - 16) compared the range of values that the model predicted (approx 0-4). This is probably due to a damping of the extreme values in the prediction.

The parameter soil depth is the only parameter where both a spatial correlation and a correlation with the landscape features (soil and land use) is found in the analysis. To compare the methods and the effect of the different approaches, an Ordinaring Kringing and a regression is done as well as the Regression Kriging (Figure 3.6). In table 3.4 it can be seen that the parameter soil depth has a more significant result for the regression method than for the Regression Kriging. Both the Regression Kriging and Ordinary Kriging have a non significant results. This can mean that the estimated semivariogram did not represent the data enough to increase the relevance of the prediction. The Regression method had a non-significant outcome as well.

Table 3.8: Model efficiency based on validation data per parameter

Parameter	Model based on	Model efficiency (R^2)
PER	Regression; Land use	0.06
LAI	Regression; Land use	0.33
CH	Regression; Land use	0.25
RR	Regression; Land use	0.07
STFRC	Regression; Geology	0.33
COH	Ordinary Kriging	0.01
D50	Ordinary Kriging	0.01
SD	Regression Kriging (regression: Soil and Land use)	0.23
SD	Ordinary Kriging	0.01
SD	Regression: Soil and Land use	0.30

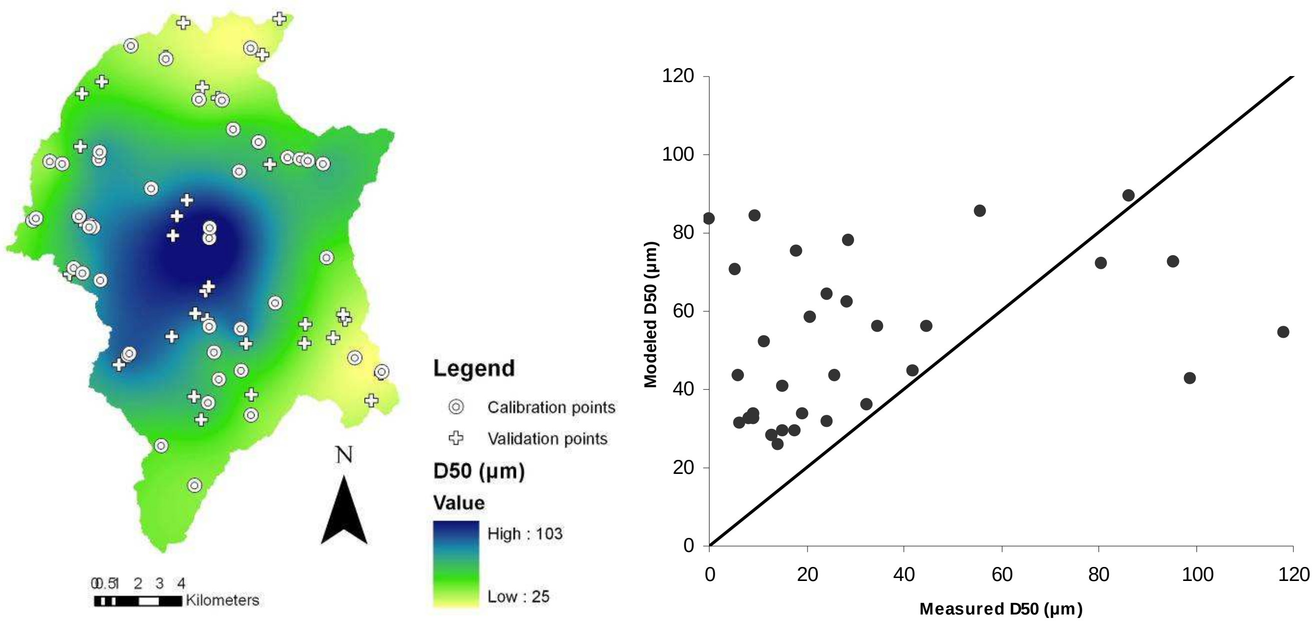


Fig 3.4: Output for D50; a) The Interpolated map based on calibration data, with the validation and calibration data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line)

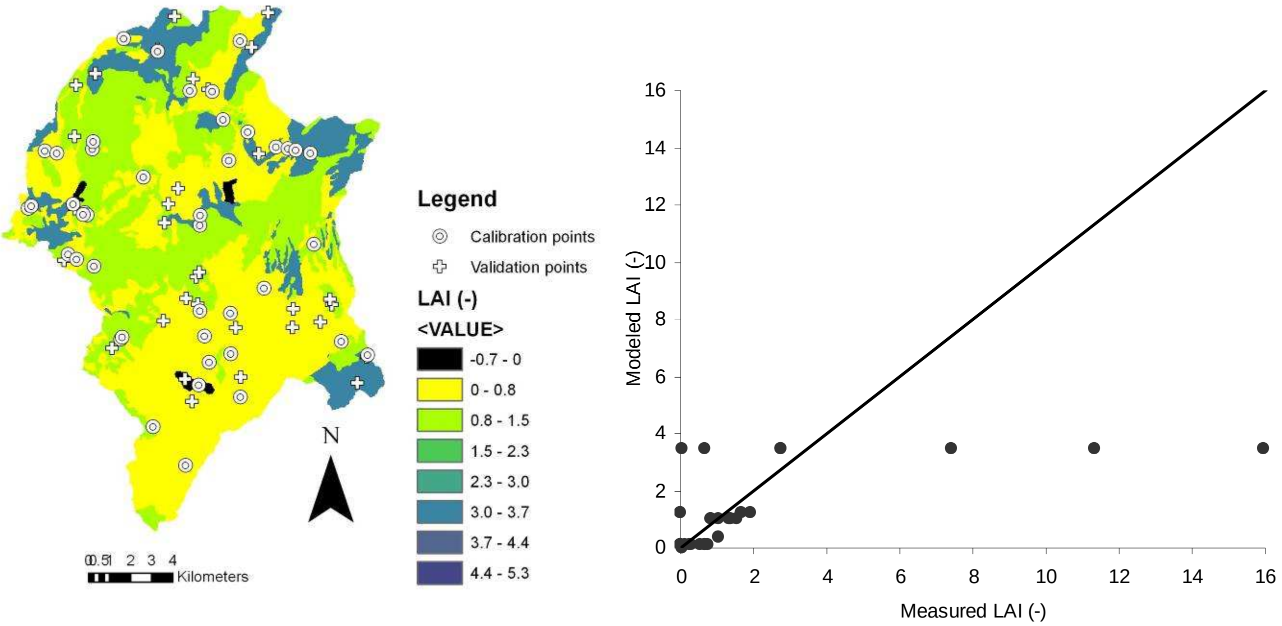


Fig 3.5: Output for LAI; a) The Interpolated map based on calibration data, with the validation and calibration data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line)

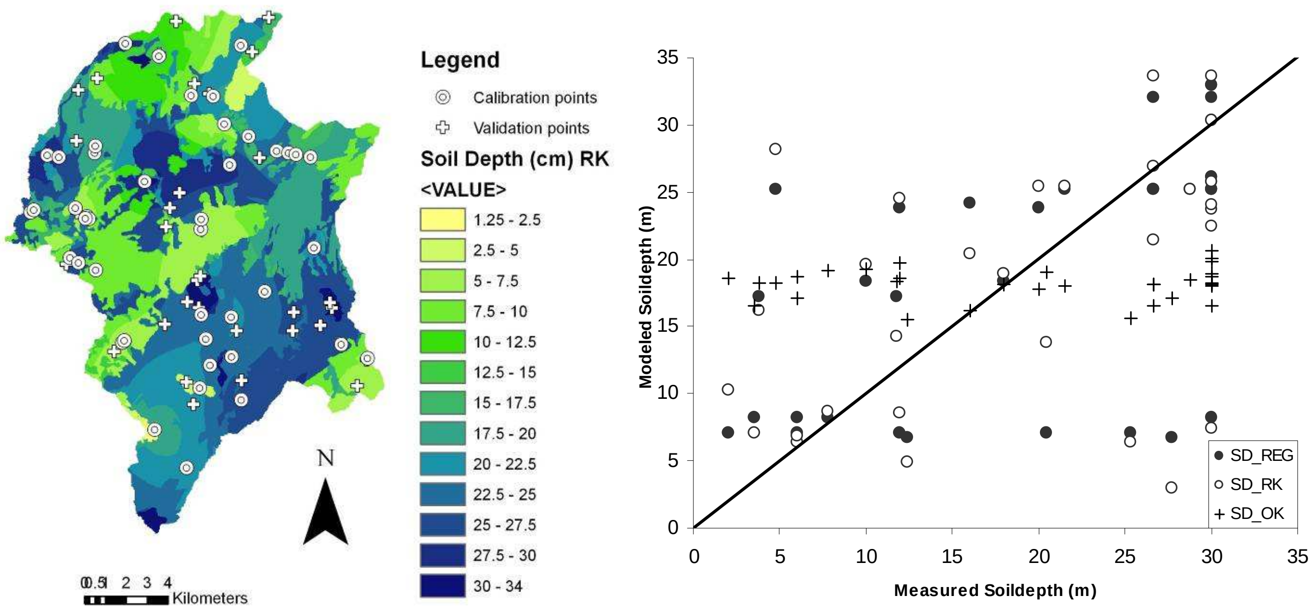


Fig 3.6: Output for Soil Depth; a) The Interpolated map based on calibration data, with the validation and calibration data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line)

4. Discussion

4.1 Results

In chapter 3 it is shown that the correlation of the modelled parameter values with the measured validation dataset is very low, which is not what was expected when formulating the hypothesis for this research. In most literature the methods of the data analysis are described in a very compact way (De Roo et al., 1996, 1996a, 1996b; Jetten et al*.,* 1996; Hessel, 2003; Boer et al., 2005). No literature is found on the statistical analysis of the parameters of LISEM. This may mean that the data were assumed to be relevant and no data analysis has been done. Jetten et al. (1996) state on this topic: "*To avoid unnecessary complexity we assumed that the sampling always lead to the correct prediction of the mean of a field, whatever the number of samples*".
In this study it is found that the correlation between the landscape features and the parameters are not as straightforward as it may look on first sight. Also some of the measuring techniques may depend too much on estimations to be exact (paragraph 4.1.2). It is therefore useful to have insight in these insecurities to know how relevant the data or model results are for further research using the data.

The question is why no correlation was found, when it was expected at the start of this research. The reason that no correlation is found between the measured validation dataset and the modelled parameter values can be due to the following causes: (1) the data is not representative, (2) the method is not correct, (3) there is no correlation between the parameters and the landscape features.

The field measurements of the parameters have considerable insecurities in the measuring methods. This can be a possible reason for not finding a correlation when expected. Another reason why the data might not be representative can be that the number of data points in the dataset are not enough to prevent a systematic error between the two datasets. When a dataset is large enough, it can be assumed that the distribution of the data is the same as the distribution of the parameter value in the field (Field, 2009). On the other hand, when the dataset is not large enough, the dataset may not represent the distribution of the parameter in the field and this can cause a systematic error between the two datasets. This will result in an biased prediction for the parameter when extrapolating the data to the whole area using regression and Kriging. Consequently one of the reasons why the dataset could not be representative could be that the variability in the landscape was larger than we were able to measure with the amount of data points we gathered. Because we measured only one data point per landscape unit, the small scale variability was not measured. Only the different combinations of landscape features were be compared. As a result, it is unknown how heterogeneous the parameters are within the combination of landscape features and on a small scale. Also because only one data point is taken for each combination of landscape features, nothing can be said about the representatively of that data point. During fieldwork we choose the data point that seemed to be representative on sight, but no validation of this can be done because there was only one data point per landscape feature combination. Because we don't know what the small scale variability of the parameters is, it is hard to say whether the variability of the parameter values is due to general variability or due to a correlation with the landscape features.

Most researches where LISEM is used as a model, have considerably smaller catchment areas than in this research (De Roo et al., 1996, 1996a, 1996b; Jetten et al*.,* 1996; Stolte, 2003; Hessel, 2003; Boer et al., 2005). In this research, it may be the case that not enough data points where taken, in order to map the entire variability of the parameters. If a smaller area would have been studied, the data points would be taken closer to each other, and the variability of the parameter values might have been mapped more significantly.

Also the different methods used for selecting the validation and calibration points may have caused a difference in dataset characteristics between the calibration and validation datasets that is due to the used methods. The method for selecting the calibration points was based on selecting the most representative location in the field for a land unit, whereas the selection

method of the validation dataset was based on a random set of locations. The reason we chose these techniques is that we wanted all the different combinations of landscape features to be represented in the data, so we could compare them. While the validation data needed to be random to make sure the data is not biased. This difference in sampling methods may have caused a difference in the distribution of the two datasets. A possibility for avoiding this problem is to choose one method for both datasets. However, both methods have advantages and disadvantages for the two different datasets that were needed and it was not useful to select one of the methods for both datasets.
During fieldwork a clear correlation between the crop height and the land use was observed. This correlation was not found in the regression method and the prediction of the parameters. This shows that a correlation that was present in the landscape did not show up in the prediction. Which increases the likelihood that either the methods or the data are not representative.

The second option: the method does not work for this situation, is a possible reason for the fact that no correlation was found, and is also influenced by the data. If the data set is not large enough, or representative enough, that effects the method. The number of data points can be too low to create a significant interpolation. Another reason why the used methods for creating a prediction do not work for the situation, can be because the initial correlation between the parameter data and the landscape feature maps are not high enough. The correlation between the landscape features and the parameters were low when considering the calibration dataset. Using this data for the interpolation, the low correlation is extrapolated in the interpolation, resulting in a prediction that had a low correlation to start with and also had a low correlation with the validation dataset.
Another variability that influences the significance of the prediction is the map where the regression is based on. The landscape feature maps of areas were adjusted in the field when necessary (Lamberink 2009), so errors from this side are as minimal as possible, but can still exist. Also the category of the validation points validation points where adjusted to the actual situation, if the landscape feature did not match the map, which reduces the error caused by possible errors of the map.

Because the options that the data is not representative, and that the used methods do not work in this case, are a possible cause for the lack of correlation, it cannot be concluded that no correlation is present. But it can be concluded that for this particular dataset and these methods used, no correlation was present.

4.2 Methods

4.2.1 Landscape conditions

The measurements of the parameters crust and initial soil moisture are sensitive to precipitation, because the value of these parameters are affected by the wetness of the soil. Because the weather varied during the fieldwork (section 2.2), it is probable that this has influenced the measurements. Crop height is varying in time as well due to growth and harvesting of the crops. These variations have not affected our measurements as such, because the measurements were taken in a short time period, so crop growth within the dataset can be neglected. When modelling over a large time period, this does play a role, so the question is how representative the measured crop height is for the long term crop height of agricultural areas. This should be accounted for in different model runs. The measurements were not influenced by harvesting, because the harvesting was either already done in the fieldwork period, or was started after our fieldwork.

4.2.2 Parameters

Soil depth

To measure the soil depth a gauge with a maximum depth of 30 cm was used. This was chosen instead of auguring or digging pits to measure the soil depth because this method saves a lot of time. The disadvantage of using a gauge is that a maximum of 30 cm can be used. At 17 (of 75) data points the maximum soil depth of 30 cm was measured. The exact depth of the soil is therefore not known in these places. This means that in 20 % of the data points, the soil depth may be underestimated in different quantities, which results in an bias. This bias will also influence the accuracy of the predicted soil depth map, because the values for the validation dataset were measured with the same method, this bias will not show in the results.

Leaf Area Index

The leaf area index measurements are quite subjective, this is due to the fact that the measuring technique consists of three steps, which all have a measure of subjectivity in them.

The first step, selecting a representative area and counting the number of each plant species is an estimation that depends of the insights of the researcher, often a location varies a lot in plant cover over a small area. Large plants, like trees, that have a large leaf area index relative to other plants and the choice whether to include, for example, one or two trees in a sample can make a significant in the eventual leaf area index.

The second step, counting the leaves of a representative plant, is subjective as well, because the plant that is taken as representative is a choice as well as the estimation of the number of leaves in the case of bigger plants, especially in the case of (pine) trees.

The photographing and processing of the area of the individual leaves is the least subjective step in the process, but still has a measure of subjectivity in the selection of the representative leafs to photograph.

Due to the successive steps which all include a form of estimation, the measuring errors made at one point in the analysis, can be extrapolated to the next step and, can therefore increase with each step. This is what probably caused the outliers of the leaf area index (Fig 3.4), where values from 4 to 15 where measured. These values mean that the fraction of leaves per area is in those cases 400 to 1500 percent which is a lot for South East Spain. Domingo et al. (1999) found values for leaf area index ranging from 0.45 to 1.05 in South East Spain, which is significantly lower than the values measured in this research. These outlying values are all values that are in a forest where the value of the trees are relatively high compared the values of the other land use types.

Crop height, Crust, PER and Stone Fraction

In the crop height measurements, the selected representative area is of large influence on the eventual representative crop height, as is the case with the leaf area index. The selection of this area is subjective.

The parameters Crust, plant cover and Stone Fraction are estimated in the field and are therefore subjective as well.

Random roughness

In the areas where tillage was present (arable areas and orchards), the direction of the random roughness board had significant influence on the measured random roughness. According to Hessel (2002 it is inadvisable to measure the random roughness in tilled areas, due to a possible overestimation of the roughness in the case of runoff in the direction of the tilling. These methods were used for the tilled areas, because the most extensive tilling was done in irrigated areas, which were on flat areas, it was assumed that the random roughness was still significant and because no alternative measuring method for the random roughness was present. These measurements for random roughness are biased, because the tilling of the area is not random.

Initial and saturated soil moisture content

The samples for soil moisture contents were taken using a sample ring of 100 cc. This is quite a small volume, and may not be large enough to be representative for the area, also due to the occurrence of stones in the sample. According to Hessel (2002) stones in the sample should be avoided because the soil moisture content is over the percentage of soil, in which rock is not counted. However, samples without stones where quite impossible in the fieldwork area and it can be argued that because it was impossible to select soil samples without stones, it is more representative to include the stones in the sample.

During the lab measurements of the initial soil moisture content the samples were not completely dry during the measurement of the dry weight which results in an error. This is corrected by using the measured initial weight instead of the measured dry weight, in the case that the initial weight was smaller than the measured dry weight. This corrected the error a bit, but then the assumption arises that the samples were dry when taken in the field. Because the samples for which this was the case were taken when it had been dry for a couple of days, this is considered a valid assumption.

Grain size
According to Konert (1997) and Murray (2000) limestone samples don't have a particle size, due to the softness of the rock and the large erosion rate. This causes the grain sizes to aggregate and disaggregate during processing, which makes the size of the particles arbitrary. Because the samples are limestone for the largest part, the limestone part of the sample was a significant part of the grain size and these particles are still taken as the particle sizes.

4.2.3 Statistical Analysis

During the data analysis the statistical assumptions were met as often as possible. In this research the choice was made to use both normal and non-normal tests for the data, depending on whether or not the parametric tests where appropriate considering the assumptions of these tests. Because according to Field (2009) parametric tests are more accurate than non-parametric tests, the parametric tests where used when possible. Another option was to transform the data, but that option was not applicable to this dataset because the different parameters were deviating from normality in different ways.
When it was not possible to use parametric tests because the assumptions were not met, the non-parametric tests were used. The results of these tests were treated in the same way during the analysis of the data. This was an acceptable method, because the parameters did not need to be compared with each other, only with the landscape feature maps.
For some assumptions, the method could not be adjusted when assumptions were not met, these assumptions were met as well as possible when gathering the data. The assumption of independence of the data could not be confirmed. The assumption that the interval of the parameter values in the dataset is similar to the interval of the parameter values in the field could not be verified. Because the method for measuring the soil depth an gauge was used, that assumption was not met for this parameter.
The statistical tests used in this research, assume that data is selected randomly (Field 2009), this assumption could not be met for de calibration dataset, because the sampling locations were selected on representativeness. Two methods to test for normality were performed; checking the skewness and kurtosis and the Kolmogorov-Smirnov test. These tests did not have consistent results on the normality of different parameters, due to differences in methods. In the analysis it was assumed that the data was not normal when one of the tests gave a non-normal results.

4.1.5 Interpolation

The spatial dependency of the data was used when the model (Appendix R) gave a significant semivariogram (a semivariogram which has a range of over 2000 meters or more), in order for a significant amount of data points to be in range. That way the prediction is based on enough data points. The predicted semivariograms that were used (Fig 3.2, appendix D) had a low correlation with the data. That low correlation influenced the prediction of the parameters, which resulted in a prediction that had a low significance. As can be seen in table 3.4 the parameters on which ordinary kriging was used, had a very low model efficiency.

Some predictions of parameters gave values that were not in the physical range of the parameter (for example, a negative value, or a value above 100 in a percentage) (Appendix D). Also a damping of the extreme values can be seen when looking at the predictions for the parameters. For example the leaf area index (LAI).

5. Conclusions and recommendations

Based on the regression analysis done in this study, no correlation is found between the landscape feature maps and the measured LISEM parameters. Most parameters have a low correlation with the landscape feature maps based on the calibration dataset. After regression analysis and interpolation, the predicted map for the parameters compared with the validation dataset, results in no correlation.

The calibration dataset is used to create an interpolation of the parameter for the whole catchment area using either linear regression, Ordinary Kriging or Regression Kriging, dependent on the presence of a correlation between the parameter and the landscape feature map and on the spatial correlation. The low correlation of the calibration data with the landscape features, and low spatial correlation caused the prediction of the different parameters to be insignificant. No significant correlation is found between the interpolated prediction and the validation dataset. The highest correlation of the measured validation data and the interpolated prediction, were the four parameters; leaf area index, crop height, stone fraction and soil depth. These parameters have a correlation of 20 to 30 percent of the variance explained by the prediction, which is not a significant correlation.

The three possible causes for this lack of correlation in the results are: (1) the data is not representative, (2) the method is not correct, (3) there is no correlation between the parameters and the landscape features.

It is possible that a reason for the lack in correlation is based on the large variability of the landscape and the parameters, which would mean that the data is not representative. This is the case for both the calibration and the validation dataset. For the parameters that were easy to estimate by sight (for instance crop height), we saw in the field that on a small scale there was variation in the parameter within a unit. Besides the variation within a unit, we a correlation could be seen between crop height and land use in the field. This could mean that there is a small scale variation within the landscape units, but also a large scale correlation in the landscape. Because we had only one data point per landscape unit, this small scale variation, within a landscape unit, adds to the variation between the different landscape feature combinations and makes it harder to find a correlation, if present.

The second option is that the method does not work for this situation. It is possible that that is the case; it could be that it is caused by the number of data points that were available. When starting with a low correlation, then performing an interpolation and validating the results with a dataset that also had a low starting correlation with the landscape features, the prediction will have a low correlation with the landscape as well. The result is that no correlation can be found because the variability of the data and low correlation adds up in each step of this method, and results in a dataset that can not be correlated. The parameters that had a prediction based on kriging had a low correlation with the validation data set because the predicted semivariogram, on which the prediction of the parameter was based, did not have a significant enough correlation with the dataset to give a significant prediction for the parameter.

It can not be concluded that no correlation is present between the landscape features and the parameters, because it is possible that the lack of found correlation can be due to the representativeness of the data or the applicability of the methods used. It is useful to have some insight in these insecurities, results are for further research using the data.

To find out whether there is correlation between landscape features and the parameters, a representative dataset for the area is needed. That was possibly not the case in this research, which is one of the reasons why it can not be concluded that no correlation is present between the parameters and the landscape features, even though no correlation was found. In future research a study for the heterogeneity on a small scale might be useful, to make sure a dataset is representative and the range of possible values for a parameter is met. This can be done by taking samples on shorter distances from each other and within a landscape unit as well. More data points per surface area are needed.

For future fieldwork it is advisable to measure the random roughness in two directions when the area is tilled; one in line with the tilling and one perpendicular to the tilling and adjust the random roughness according to the slope and aspect of the site. However, the direction in which an area is tilled may vary over time.
The measuring method of the leaf area index, gave a lot of outliers. It was often unclear whether one or two trees per plot were representative for the area. In the case of a larger plot area, this would be less of a problem, because an average is more easily made.

Another possibility for future research in this area is to exclude the landscape feature soil, because it showed in the fieldwork that the soil was similar throughout the landscape and probably the least varying landscape feature. When excluded more data points could be taken in the other landscape feature and a better view of the variation of the parameters can be obtained in this way.

6. Literature

Beuselinck L., Govers G., Poesen J, Degraer G., Froyen L., 1998, *Grain-size analysis by laser diffractometry: comparison with the sieve-pipette method*. Catena 32, 193–208

Bivand R.S., Pebesma E.J., Gomez-Rubio V., 2008 *Applied spatial data analysis with R.* Springer, NY.http://www.asdar-book.org/.

Boer, M.M., Puidefabregas, J., 2005. *Redistribution Processes Create Multiple Challenges for the Prediction of Storm Runoff and Soil Loss from Semiarid Hillslopes.* Geophysical Research Abstracts, Vol. 7, 06065, 2005

Bull, L.J., Kirkby, M.J., Shannon, J. and Hooke, J.M., 1999. *The impact of rainstorms on floods in ephemeral channels in southeast Spain*. Catena 38, pp. 191-209.

CORINE Land Cover, 2000, European Environment Agency, copenhagen, denmark

Davis J.C., 1986. *Statistics and Data Analysis in Geology*. Second Edition, New York: John Wiley & sons.

De Roo APJ, 1996. *The LISEM project: an introduction*. Hydrological Processes 10: 1021-1025

De Roo APJ, Wesseling CG, Ritsema CJ, 1996a. *LISEM: A single-event physically based hydrological and soil erosion model for drainage basins. I: Theory, input and output.* Hydrological Processes 10: 1107-1117

De Roo APJ, Offermans RJE, Cremers NHDT, 1996b. *LISEM: A single-event, physically based hydrological and soil erosion model for drainage basins. II: Sensitivity analysis, validation and application.* Hydrological Processes 10: 1119-1126

De Vente J, Solé A (2008) Information review of land degradation in the Guadalentin basin (SE-Spain) for the DESIRE HIS

Domingo, F., Villagarcía, L., Brenner, A. J., Puigdefábregas J. (1999). *Evapotranspiration model for semi-arid shrub-lands tested against data from SE Spain.* Agricultural and Forest Meteorology Volume 95, Issue 2, 2 June 1999, Pages 67-84

Field, A., (2009) *Discovering Statistics using SPSS.* Third Edition, London SAGE Publications Ltd.

Hengl, T., Heuvelink, G.B.M, Stein, A., 2004. *A generic framework for spatial prediction of soil variables based on regression-kriging.* Geoderma 120 (2004) 75–93

Hengl, T., Heuvelink, G.B.M, Rossiter, D.G., 2007. *About regression-kriging: From equations to case studies.* Computers & Geosciences 33 (2007) 1301–1315

Hessel, R., van Dijck, S. and van den Elsen, E., 2002. *LISEM project field measurements manual. Version 1.2*. Alterra, Green World Research, Wageningen, Netherlands.

Hessel R, Jetten V, Baoyuan L, Yan Z, Stolte J (2003) *Calibration of the LISEM model for a small Loess Plateau catchment*. Catena 54: 235-254

Hooke JM, Mant JM, 2000. *Geomorphological impacts of a flood event on ephemeral channels in SE Spain*. Geomorphology 34: 163-180

Hooke, J.M. and Mant, J.M., 2002. *Floodwater use and management strategies in valleys of southeast Spain*. Land Degradation & Development 13, pp. 165-175.

IGME, 1981. *Mapa Geologico de Espana, 1:50.000, Sheet Lorca (953)*. Servicio de Publicaciones Ministerio de Industria, Madrid.

Jester W., Klik A., 2005: *Soil surface roughness measurement - Methods, applicability, and surface representation.* CATENA 64 (2-3): 174-192

Jetten, V., Boiffin, J., De Roo, A., 1996. *Defining monitoring strategies for runoff and erosion studies in Agricultural catchments: A simulation approach.* European Journal of Soil Science, 47,579-592.

Jetten, V., 2002. *LISEM user manual, version 2.x. Draft version January 2002.* Utrecht Centre for Environment and Landscape Dynamics, Utrecht University, Netherlands.

Keitt, T.H., Bivand R., Pebesma E., Rowlingson, B., (2010)*. rgdal: Bindings for the Geospatial Data Abstraction Library.* R package version 0.6-28. http://CRAN.R-project.org/package=rgdal

Kilpelainen, J., no date. *prmprj program, beta version,* Agricultural research centre, Finland.

Konert, M., van der Berghe, J., 1997. *Comparison of laser grain size analysis with pipette and sieve analysis: a solution for the underestimation of the clay fraction,* Sedimentology 44, 523-535

LUCDEME, 1988*. Hoja 953 - Lorca, Mapa de Suelos; 1:100 000*. Ministerio de Agricultura, Pesca y Alimentacion, Murcia.

Lamberink, K., 2009. *Setting up and carrying out a sampling design to run LISEM for the Torrealvilla catchment in South-east Spain.* Master Thesis, Wageningen University, Land Degradation and Development Group, the Netherlands, unpublished.

Lammers, M., 2009. *Preparing the implementation of LISEM for the Prado catchment in SE Spain. Data collection and LISEM calibration review.* Master Thesis, Wageningen University, Land Degradation and Development Group, the Netherlands, unpublished.

Pebesma, E.J., 2004*. Multivariable geostatistics in S: the gstat package.* Computers & Geosciences, 30: 683-691.

Pebesma, E.J., Bivand,R.S., 2005. *Classes and methods for spatial data in R.* R News 5 (2), http://cran.r-project.org/doc/Rnews/.

R Development Core Team, 2010. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sarkar D., 2010*. lattice: Lattice Graphics.* R package version 0.18-5. http://r-forge.r-project.org/projects/lattice/

Stolte, J., van Venrooij, B. Zhang, G., Trouwborst, K.O., Liu, G, Ritsema, C.J., Hessel, R. , 2003. *Land-use induced spatial heterogeneity of soil hydraulic properties on the Loess Plateau in China.* Catena 54, 59–75.

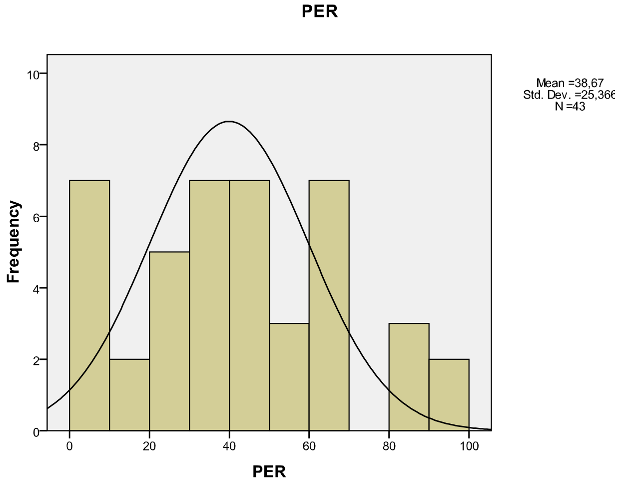
VU-website (no date), website Faculteit der Aard en Levenswetenschappen, Grain Size Analysis Laboratory, Available from: http://www.falw.vu.nl/nl/onderzoek/laboratoria/labs-aardwetenschappen/fysisch-laboratorium/sediment-analysis-laboratory/grain-size-analysis.asp [site viewed: 21-06-2010]

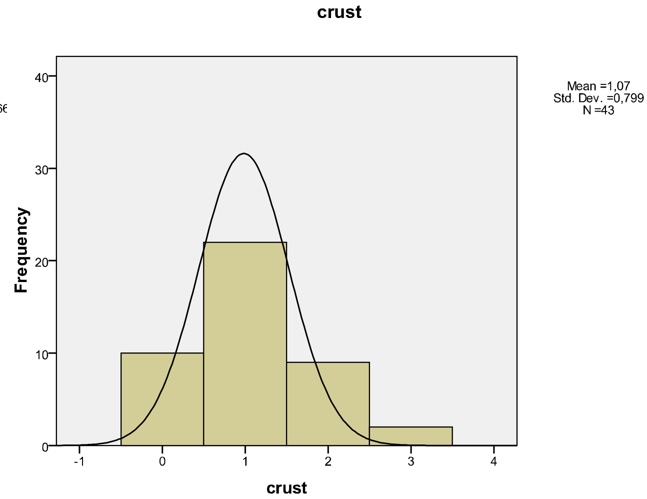
Wagner L.E., Yu Y., 1991: *Digitization of profile meter photographs*. Transactions of the American Society of Agricultural Engineers, 34 (2): 412-416

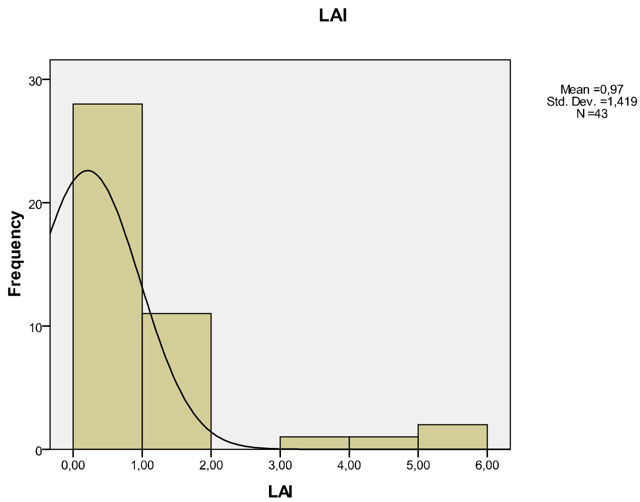
7. Appendices

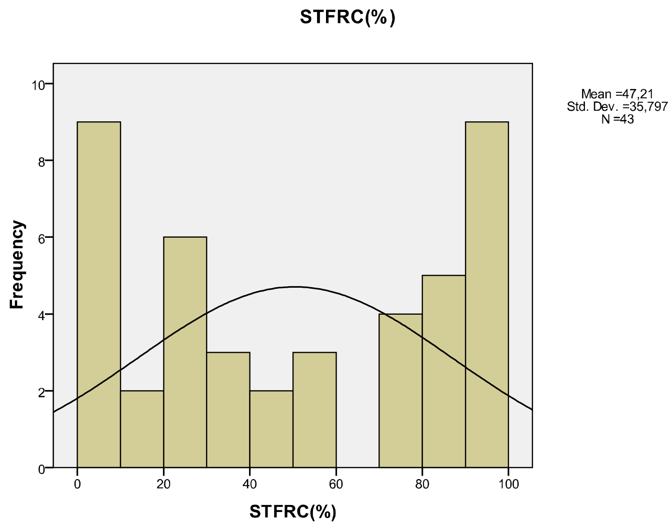
A: Data

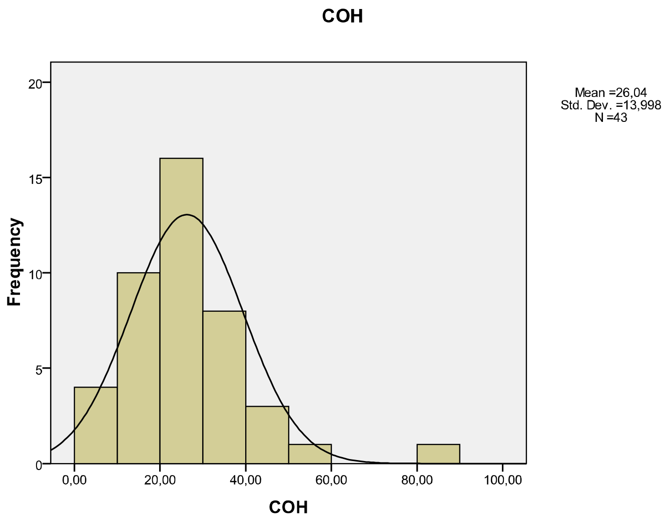
Histograms of the calibration data per parameter:

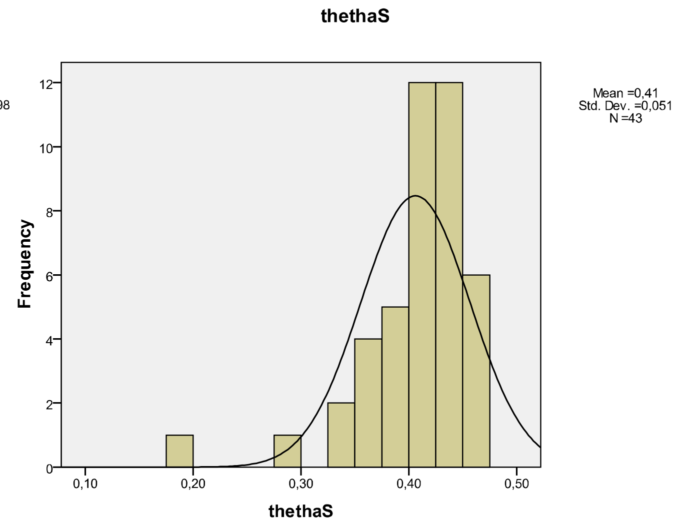


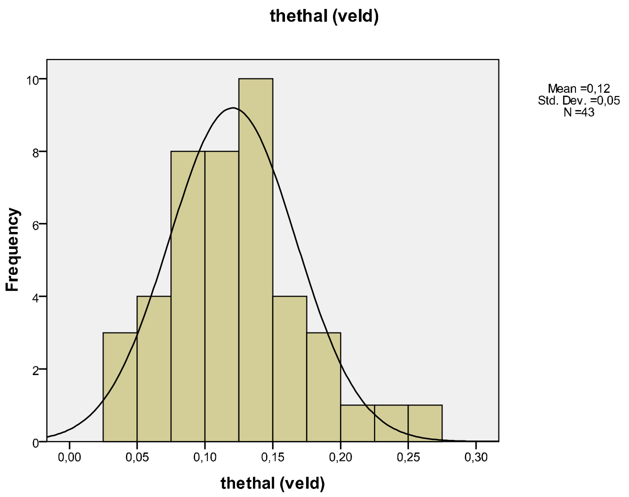


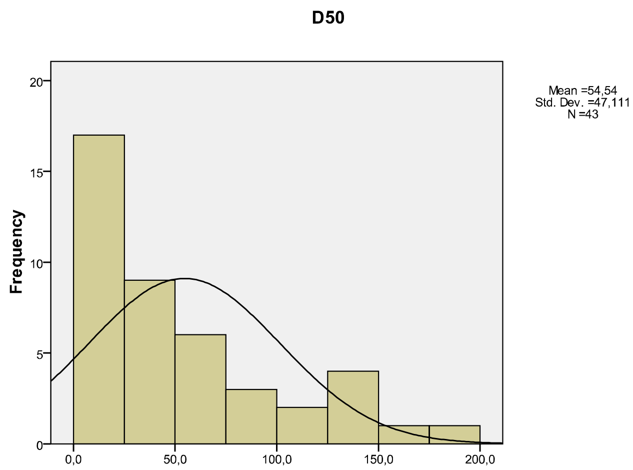


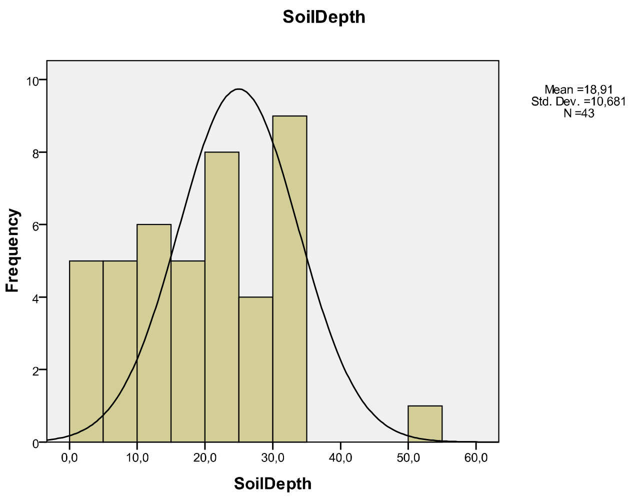


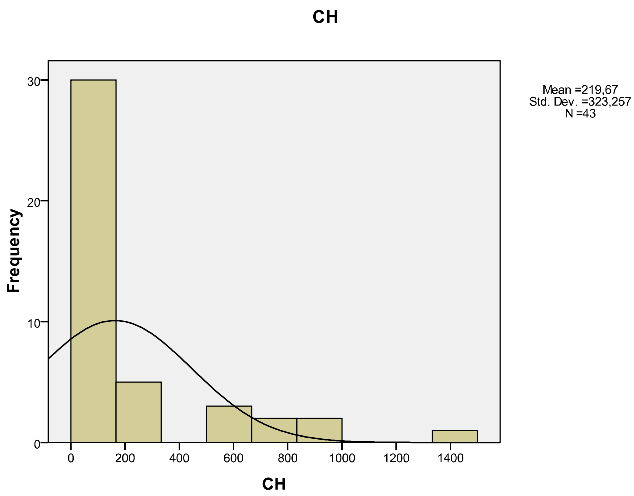





thethal (veld)
Mean =0,12
Std. Dev. =0,05
N =43
Frequency
0 2 4 6 8 10
0,00 0,05 0,10 0,15 0,20 0,25 0,30
thethal (veld)


D50
Mean =54,54
Std. Dev. =47,111
N =43
Frequency
0 5 10 15 20
0,0 50,0 100,0 150,0 200,0


SoilDepth
Mean =18,91
Std. Dev. =10,681
N =43
Frequency
0 2 4 6 8 10
0,0 10,0 20,0 30,0 40,0 50,0 60,0
SoilDepth


CH
Mean =219,67
Std. Dev. =323,257
N =43
Frequency
0 10 20 30
0 200 400 600 800 1000 1200 1400
CH

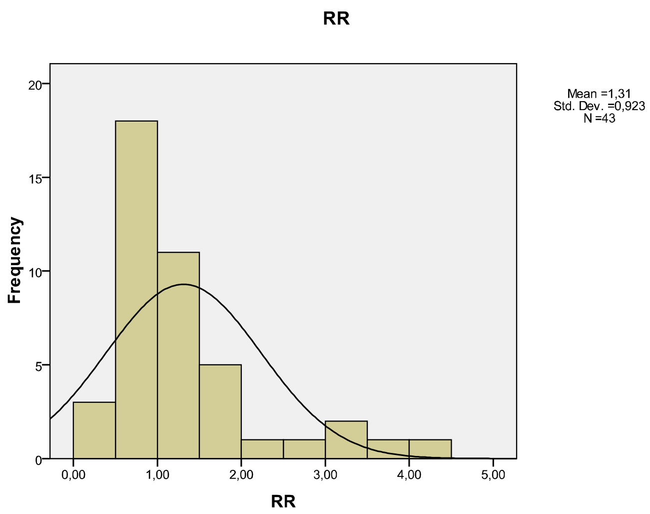

RR
Mean =1,31
Std. Dev. =0,923
N =43
Frequency
0 5 10 15 20
0,00 1,00 2,00 3,00 4,00 5,00
RR

Table A.1: Calibration dataset

Site	code	Land use	Geology	Soil	date	X	Y	PER	crust	LAI	CH	RR	STFRC	COH	θS	θI	D50	SoilDepth
								%	-	-	cm	cm	-	kPa	Vol%	Vol%	μm	cm
201	322	nonirrigaded arable	marl	fluvisol	24-9-2009	612588	4175095	30	0	0.00	21	1.30	0.05	18.3	0.39	0.13	31	19.5
202	145	mattoral	conglomerate	cambisol	24-9-2009	611055	4176914	50	1	0.49	60	1.32	0.8	55.6	0.42	0.14	24	6.3
203	142	nonirrigaded arable	conglomerate	cambisol	24-9-2009	614729	4180398	40	1	0.00	20	1.61	0.5	24.5	0.46	0.10	34	22.7
204	122	nonirrigaded arable	marl	cambisol	24-9-2009	616311	4183513	25	2	0.00	5	0.54	0.01	20.9	0.43	0.10	41	50.0
205	135	mattoral	glacis	cambisol	24-9-2009	618684	4185606	60	2	0.67	51	1.72	0.25	40.5	0.46	0.14	20	22.3
206	121	orchard	marl	cambisol	25-9-2009	615169	4195275	5	1	0.59	300	3.21	0.01	46.1	0.33	0.17	7.1	19.3
207	161	orchard	undefined	cambisol	25-9-2009	613847	4192858	5	1	0.71	500	2.58	0.75	10.0	0.41	0.14	23	14.5
208	214	forest	limestone	lithosol	25-9-2009	614375	4191528	30	1	1.86	601	1.24	0.95	80.1	0.45	0.18	62	7.0
209	215	mattoral	limestone	lithosol	25-9-2009	612793	4192896	35	1	1.11	70	1.42	0.95	29.4	0.45	0.13	23	7.4
210	224	forest	marl	lithosol	25-9-2009	611262	4194772	25	1	1.56	668	1.32	0.7	2.0	0.42	0.16	35	21.0
211	131	orchard	glacis	cambisol	25-9-2009	615543	4190935	5	1	0.03	230	1.98	0.9	26.8	0.44	0.06	91	18.3
212	321	orchard	marl	fluvisol	26-9-2009	605916	4190040	5	3	0.44	290	3.99	0.01	10.6	0.41	0.08	3.3	31.7
213	361	orchard	undefined	fluvisol	26-9-2009	606472	4189935	4	2	0.59	300	0.79	0.4	28.4	0.41	0.07	18	30.0
214	112	nonirrigaded arable	limestone	cambisol	26-9-2009	605121	4187307	60	2	0.00	15	0.99	0.2	23.5	0.43	0.09	11	28.0
215	114	forest	limestone	cambisol	26-9-2009	605270	4187417	50	1	1.73	565	0.70	0.9	42.5	0.45	0.16	66	18.3
216	116	abandoned/bare	limestone	cambisol	26-9-2009	607005	4185127	20	1	0.11	110	0.89	0.8	21.9	0.38	0.11	56	23.3
217	166	abandoned/grain	undefined	cambisol	26-9-2009	607412	4184892	60	1	0.00	43	1.13	0.07	27.8	0.42	0.13	24	25.3
218	215	mattoral	limestone	lithosol	26-9-2009	607801	4187099	40	1	1.12	141	1.50	0.95	30.7	0.35	0.08	51	5.8
219	162	nonirrigaded arable	undefined	cambisol	26-9-2009	607903	4186990	80	2	0.00	10	1.57	0.05	16.0	0.45	0.05	47	25.0
220	216	bare	limestone	lithosol	26-9-2009	607735	4187013	40	2	0.08	83	0.66	0.3	23.5	0.41	0.12	128	11.3
221	247	mine	conglomerate	lithosol	30-9-2009	613214	4178902	0	0	0.00	0	1.26	0.85	7.8	0.19	0.11	117	0.0
222	255	mattoral	calcarenite	lithosol	1-10-2009	619981	4180983	40	0	1.29	122	0.81	0.8	14.7	0.36	0.24	11	13.7
223	217	mine/commercial	limestone	lithosol	2-10-2009	607255	4187507	10	0	0.00	15	0.59	0.95	16.7	0.39	0.25	78	1.5
224	146	bare	conglomerate	cambisol	2-10-2009	608170	4190138	30	0	0.07	73	0.76	0.5	19.3	0.44	0.20	67	21.0
225	235	mattoral	glacis	lithosol	2-10-2009	608207	4190475	60	0	1.54	110	0.88	0.95	8.8	0.43	0.15	144	9.0
226	126	bare	marl	cambisol	3-10-2009	615201	4178336	25	1	0.30	68	0.69	0.1	23.2	0.29	0.10	28	30.0
227	123	irrigated arable	marl	cambisol	3-10-2009	613703	4179990	40	1	0.74	15	0.92	0	32.0	0.33	0.20	19	30.0
228	342	nonirrigaded arable	conglomerate	fluvisol	3-10-2009	614722	4182307	1	1	0.00	0	0.61	0.7	6.4	0.39	0.08	167	20.0
229	242	nonirrigaded arable	conglomerate	lithosol	5-10-2009	609508	4181084	30	1	0.00	0	0.80	0.9	25.5	0.40	0.10	131	18.5
230	245	mattoral	conglomerate	lithosol	5-10-2009	609589	4181171	30	1	0.76	95	0.87	0.99	23.9	0.43	0.14	66	0.7
231	225	mattoral	marl	lithosol	5-10-2009	608238	4184554	30	2	0.67	122	0.35	0.1	30.7	0.41	0.10	85	12.0
232	164	forest	undefined	cambisol	6-10-2009	616881	4190228	90	0	3.93	829	0.57	0.5	31.4	0.44	0.19	30	11.5
233	111	orchard	limestone	cambisol	6-10-2009	617460	4190153	80	2	0.03	250	3.19	0.2	39.6	0.37	0.14	38	30.0
234	134	forest	glacis	cambisol	6-10-2009	617812	4190074	60	0	4.11	1405	0.88	0.75	19.0	0.46	0.14	14	13.0
235	124	forest	marl	cambisol	6-10-2009	618515	4189935	65	1	5.94	967	0.45	0.2	38.6	0.41	0.16	132	20.3
236	152	nonirrigaded arable	calcarenite	cambisol	7-10-2009	613260	4186502	50	1	0.00	20	1.30	0.2	24.8	0.40	0.04	189	20.8
237	255	mattoral	calcarenite	lithosol	7-10-2009	613284	4186974	60	0	1.48	101	2.08	0.8	38.4	0.42	0.05	103	4.0
238	254	forest	calcarenite	lithosol	7-10-2009	621230	4180332	99	0	5.44	987	1.09	0.3	18.6	0.45	0.07	18	3.7
239	323	irrigated arable	marl	fluvisol	8-10-2009	613275	4182418	40	3	1.15	20	4.44	0.01	13.1	0.38	0.12	24	30.0
240	132	nonirrigaded arable	glacis	cambisol	9-10-2009	609653	4195380	14	1	0.00	2	1.40	0.4	24.8	0.43	0.04	20	30.0
241	157	mine/commercial	undefined	cambisol	10-10-2009	614642	4189580	20	1	0.00	1	0.34	0.25	33.0	0.44	0.08	28	26.7
242	165	mattoral	undefined	cambisol	12-10-2009	610581	4188785	40	1	1.11	92	0.62	0.3	21.6	0.43	0.09	20	30.0
243	323	irrigated arable	marl	fluvisol	13-10-2009	613491	4181242	80	2	1.87	70	1.03	0	28.4	0.36	0.12	21	30.0

Table A.2: Validation dataset

site	code	Land use	Geology	Soil	date	X	Y	PER	crust	LAI	CH	RR	STFRC	COH	COH	θS	θI	D50
								%	-	-	cm	cm	-	kPa	kPa	Vol%	Vol%	μm
L1b	161	orchard almond	undefined	cambisol	10-10-2009	612957	4193447	37	2	0.85	350	1.71	0.25	28.8	28.8	0.46	0.07	9.3
L2a	322	non irrigated/bare	marl	fluvisol	8-10-2009	613102	4184047	90	2	0.00	7	0.51	0.05	17.3	17.3	0.44	0.12	9.5
L3b	323	irrigated arable	marl	fluvisol	8-10-2009	613177	4128734	40	3	1.92	20	3.72	0.03	26.8	26.8	0.49	0.14	5.3
L3a	123	irrigated arable	marl	cambisol	17-10-2009	619545	4182725	65	3	1.66	20	2.87	0.01	33.7	33.7	0.37	0.21	8.2
L4a	214	forest	limestone	lithosol	9-10-2009	612076	4196418	60	0	2.75	978	1.15	0.9	24.8	24.8	0.46	0.10	18
G1b	211	orchard	limestone	lithosol	12-10-2009	616513	4196603	5	0	0.04	300	1.08	0.95	18.6	18.6	0.45	0.06	6.2
L4b	214	forest	limestone	lithosol	15-10-2009	615721	4194965	90	0	15.96	984	0.92	0.55	22.9	22.9	0.52	0.12	13
S2a	235	mattorral	glacis	lithosol	9-10-2009	607398	4193153	35	1	1.38	217	1.41	0.8	26.5	26.5	0.50	0.08	42
S2b	234	forest	glacis	lithosol	2-10-2009	607318	4190719	60	0	1.04	226	0.24	0.95	16.0	16.0	0.48	0.13	118
L5b	245	mattorral	conglomerate	lithosol	5-10-2009	611547	4181954	20	1	1.35	93	0.47	0.8	24.2	24.2	0.36	0.08	96
L6a	126	bare	marl	cambisol	13-10-2009	618990	4181886	40	2	0.66	56	1.87	0.2	22.2	22.2	0.43	0.11	9
L6b	246	bare	conglomerate	lithosol	13-10-2009	615212	4179265	60	0	0.28	62	1.37	0.6	19.6	19.6	0.38	0.08	26
L1a	111	orchard almond	limestone	cambisol	2-10-2009	607349	4187281	5	0	0.03	250	0.98	0.15	8.4	8.4	0.42	0.22	24
L7a	147	mine	conglomerate	cambisol	8-10-2009	612571	4179182	5	2	0.03	150	1.48	0.1	21.2	21.2	0.37	0.07	21
L7b	247	mine	conglomerate	lithosol	8-10-2009	612242	4188233	30	1	0.26	70	1.27	0.8	35.0	35.0	0.40	0.07	29
G1a	215	mattorral	limestone	lithosol	12-10-2009	606830	4184824	60	1	1.53	84	2.01	0.9	28.4	28.4	0.45	0.08	45
G2a	121	abandoned orchard	marl	cambisol	7-10-2009	617714	4182512	90	3	0.12	249	2.29	0.01	21.6	21.6	0.40	0.12	6.1
G2b	124	forest	marl	cambisol	9-10-2009	611268	4194852	40	2	2.77	117	1.06	0.5	15.4	15.4	0.43	0.09	24
G3a	136	bare	glacis	cambisol	7-10-2009	619449	4182985	10	3	0.00	0	0.32	0.1	21.2	21.2	0.41	0.08	19
G3b	234	forest	glacis	lithosol	9-10-2009	608307	4193706	70	0	11.32	493	0.71	0.95	19.9	19.9	0.43	0.08	99
L2b	142	Non irrigated arable	marl/conglomerate	cambisol	3-10-2009	614978	4181627	3	1	0.00	0	1.14	0.15	16.3	16.3	0.45	0.11	28
G4b	145	mattorral	conglomerate	cambisol	3-10-2009	612905	4178115	20	2	0.55	143	1.03	0.4	16.3	16.3	0.42	0.18	11
G4a	245	mattorral	conglomerate	lithosol	16-10-2009	609109	4180643	30	1	0.84	88	0.82	0.9	36.0	36.0	0.42	0.05	81
G5b	254	forest	calcarenite	lithosol	1-10-2009	620751	4178996	100	0	7.40	982	1.07	0.6	4.6	4.6	0.41	0.15	15
G5a	152	non irrigated arable	calcarenite	cambisol	12-10-2009	611594	4186600	25	1	0.00	20	0.52	0.1	22.9	22.9	0.40	0.06	86
G6a	166	bare	undefined	cambisol	12-10-2009	611782	4187507	55	1	0.68	54	1.39	0.4	20.6	20.6	0.45	0.09	0
G6b	261	orchard	undefined	lithosol	10-10-2009	613670	4192955	30	2	1.02	200	1.46	0.4	28.4	28.4	0.44	0.05	33
S1a	123	irrigated arable	marl	cambisol	7-10-2009	617665	4181639	40	3	0.74	13	4.90	0	19.6	19.6	0.39	0.24	15
S1b	166	bare/abandoned	undefined	cambisol	6-10-2009	616074	4189897	95	1	0.02	74	0.45	0.05	23.5	23.5	0.50	0.12	35
L5a	155	mattorral	calcarenite	cambisol	7-10-2009	621199	4180249	50	2	0.65	60	0.43	0.01	47.1	47.1	0.46	0.18	14
S3a	322	non irrigated arable	marl	fluvisol	8-10-2009	613251	4184253	0	0	0.00	0	1.35	0.05	4.8	4.8	0.45	0.03	56
S3b	322	non irrigated arable	marl	fluvisol	10-10-2009	612625	4183017	90	0	0.00	20	0.61	0.05	37.6	37.6	0.45	0.14	18

Table A.3: Summary of statisics calibration and validation dataset

Parameter	Dataset	Min	Max	Mean	Standard Deviation
PER	Calibration	0	99	39	25
	Validation	0	100	45	30
crust	Calibration	0.0	3.0	1.1	0.8
	Validation	0.0	3.0	1.3	1.1
LAI	Calibration	0.00	5.94	0.97	1.42
	Validation	0.00	15.96	1.75	3.47
CH	Calibration	0	1405	220	323
	Validation	0	984	199	281
RR	Calibration	0.34	4.44	1.31	0.92
	Validation	0.24	4.90	1.33	0.99
STFRC	Calibration	0.00	0.99	0.47	0.36
	Validation	0.00	0.95	0.40	0.36
COH	Calibration	2.0	80.1	26.0	14.0
	Validation	4.6	47.1	22.8	9.0
θS	Calibration	0.19	0.46	0.41	0.05
	Validation	0.36	0.52	0.43	0.04
θI (veld)	Calibration	0.04	0.25	0.12	0.05
	Validation	0.03	0.24	0.11	0.05
D50	Calibration	3.3	189.4	54.5	47.1
	Validation	0.0	118.2	31.8	31.0
SoilDepth	Calibration	0.0	50.0	18.9	10.7
	Validation	2.0	30.0	19.4	10.2

B: Statistics

Table B.1: SPSS output for the Skewness and Kurtosis of the parameter distributions, grey coloured values are significant at p ≤ 0.05.

	PER	crust	LAI	CH	RR	STFRC	COH	θS	θI (veld)	D50	SoilDepth
Skewness	0.472	0.46	2	2	2	0.076	0.375	-1	0.615	1	-0.34
Std. Error of Skewness	0.365	0.361	0.361	0.361	0.361	0.361	0.365	0.365	0.361	0.361	0.365
Z-score Skewness	1.29	1.27	6.20	5.73	5.28	0.21	1.03	-2.91	1.70	3.53	-0.92
Kurtosis	-0.19	-0.01	5	4	4	-2	0.337	1	0.349	0.766	-1
Std. Error of Kurtosis	0.717	0.709	0.709	0.709	0.709	0.709	0.717	0.717	0.709	0.709	0.717
Z-score Kurtosis	-0.26	-0.02	6.96	5.68	5.07	-2.29	0.47	1.41	0.49	1.08	-1.48

TableB.2: SPSS output of the Kolmogorov-Smirnov test, grey coloured values can be considered normal (at p ≤ 0.05).

Kolmogorov-Smirnov	PER	crust	LAI	CH	RR	STFRC	COH	θS	θI (veld)	D50	SoilDepth
Statistic	.128	.295	.243	.313	.186	.154	.082	.131	.092	.236	.128
df	39	39	39	39	39	39	39	39	39	39	39
Sig.	.108	.000	.000	.000	.002	.020	,200	.090	,200*	.000	.104

TableB.3: SPSS output of Levene's test, grey coloured values are the significance values which give proof for a homogeneous distribution (at p ≤ 0.05).

		Levene statistic	df1	df2	sig
Soildepth	Soil	.048	2	39	.953
	Geology	1.203	5	36	.327
PER	Soil	.558	2	39	.577
	Geology	1.617	5	36	.180
Crust	Soil	3.222	2	40	.050
	Geology	1.517	5	37	.208
LAI	Soil	.631	2	40	.537
	Geology	1.011	5	37	.425
CH	Soil	1.197	2	40	.313
	Geology	2.173	5	37	.078
RR	Soil	12.182	2	40	.000
	Geology	3.450	5	37	.012
STFRC	Soil	.264	2	40	.769
	Geology	2.430	5	37	.053
COH	Soil	.102	2	39	.904
	Geology	.467	5	36	.798
ThetaS	Soil	1.493	2	39	.237
	Geology	2.604	5	36	.041
Thetal	Soil	1.348	2	40	.271
	Geology	2.946	5	37	.025
D50	Soil	.542	2	40	.586
	Geology	3.146	5	37	.018

C: R- Script

This appendix gives an overview and explanation of the r-script that was used to make the prediction maps for the parameters. The packages used for this script were: sp (Pebesma et al., 2005; Bivand et al., 2008), lattice (Sarkar, 2010), gstat (Pebesma, 2004; Pebesma et al., 2005), and rgdal (Keitt et al., 2010).

```
[1]    library(sp)
[2]    library(lattice)
[3]    library(gstat)
[4]    library(rgdal)
[5]    data <- read.delim("cali_SD_0319.txt")
[6]    coordinates(data)=~X+Y
```

This part of the script loads the data:
Line 1 – 4 loads the packages needed for this R-script.
Line 5 loads the table containing the field data, spatial data and dummy-variables for the landscape features.
Line 6 defines the columns called ‘X’ and ‘Y’ from the data table as spatial coordinates

```
[7]           g = gstat(id = "SoilDepth",
[8]      formula = SoilDepth~lu1+lu2+lu3+lu4+lu5+lu6+
[9]            soil1+soil2, data = data)
[10]          vg = variogram(g)
[11]          vgm = vgm(75,"Sph",6000,50)
[12]          windows(width = 5, height = 4)
[13]          plot(vg,vgm)
[14]          vgmf = fit.variogram(vg,vgm)
[15]          windows(width = 5, height = 4)
[16]          plot(vg,vgmf)
[17]          savePlot(filename = "name", type = "png")
```

This part of the script estimates the semivariogram:

Line 7-9 creates a ‘gstat’ object, which is an object that holds all the information necessary for univariate or multivariate geostatistical prediction (Pebesma et. al, 2005). This object is used by the gstat package for the spatial correlation and gives it the name ‘g’:
id = the identifier of the variable as defined in the data.
Formula = formula that defines the dependent variable as a linear model of independent variables; In this example the dependent variable is called ‘SoilDepth’. When ordinary kriging is used the formula is SoilDepth ~1. When regression or regression kriging is used, the formula is the sum of all the variables minus one, on which ‘SoilDepth’ is linearly dependent, in this case “`lu1+lu2+lu3+lu4+lu5+lu6+soil1+soil2`” . These are the dummy variables for the landscape feartures. (Pebesma et .al, 2005)
Line 10 calculates the semivariance of the data using the gstat object ‘g’.
Line 11 defines an initial semivariogram estimation, estimated by the user on sight
Line 12 and 13 plot the calculated semivariance with the initial estimation of the semivariogram.
Line 14 uses the calculated semivariance and the initial estimation of the semivariogram to calculate the estimated semivariogram, this semivariogram is used in line 31-33.
Line 15 and 16 plot the calculated semivariance with the estimated semivariogram.
Line 17 saves the plot.

```
[18]        gebied = readGDAL("lu1_40m.asc")
[19]        gebied$lu1 = gebied$band1
[21]        gebied$band1 = NULL
[22]        gebied$lu2 = readGDAL("lu2_40m.asc")$band1
[23]        gebied$lu3 = readGDAL("lu3_40m.asc")$band1
[24]        gebied$lu4 = readGDAL("lu4_40m.asc")$band1
[25]        gebied$lu5 = readGDAL("lu5_40m.asc")$band1
[26]        gebied$lu6 = readGDAL("lu6_40m.asc")$band1
[27]        gebied$lu7 = readGDAL("lu7_40m.asc")$band1
[28]        gebied$soil1 = readGDAL("soil1_40m.asc")$band1
[29]        gebied$soil2 = readGDAL("soil2_40m.asc")$band1
[30]        gebied$soil3 = readGDAL("soil3_40m.asc")$band1
```

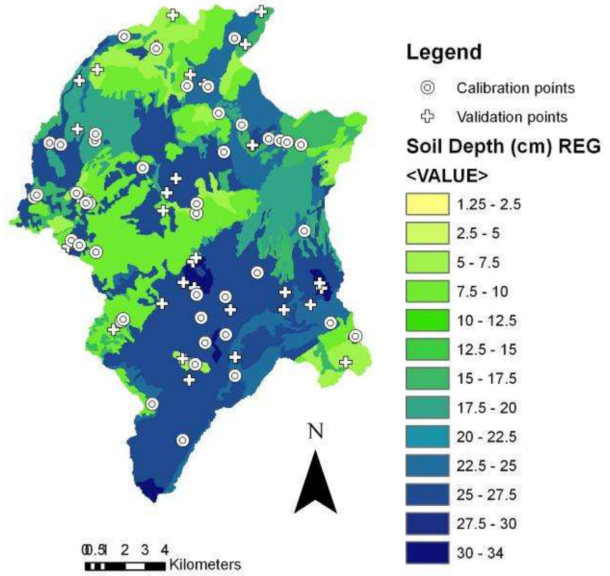
This part of the script loads the predictor maps into a array that is used for the linear regression later in the script (line 35). The predictor maps are the landscape features maps of the research area. The classes of the maps are divided over the predictor maps, where each predictor map represents one class. The predictor maps are 1 for the area that belongs to the landscape feature class, and 0 for the rest f the research area.

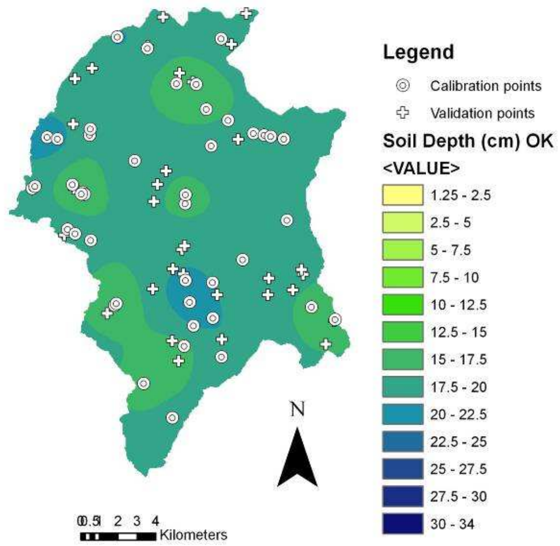
Line 18 defines the array with the first predictor map
Line 19-20 renames the layer in the array from ‘band1’ to ‘lu1’
Line 21-30 defines the rest of the predictor maps as layers in the array.

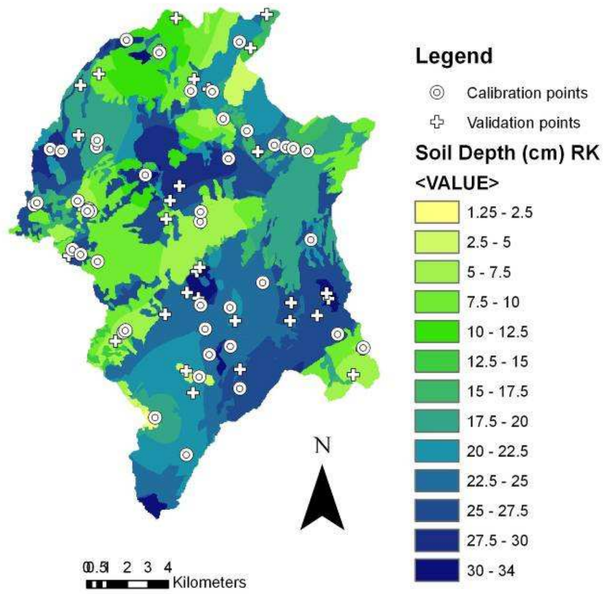
```
[31]       gSD = gstat(id = "SoilDepth", formula =
[32]       SoilDepth~lu1+lu2+lu3+lu4+lu5+lu6+soil1+soil2,
[33]       data = data, model = vgmf)
[34]
[35]       krig = predict.gstat(gSD, newdata = gebied, BLUE = FALSE,
[36]  debug.level = -1)
[37]       windows(width = 5, height = 5)
[38]  spplot(krig, zcol = "SoilDepth.pred", col.regions =
[39]       bpy.colors())
[40]       savePlot(filename = "name", type = "png")
[41]       write.asciigrid(krig, "name.asc")
```

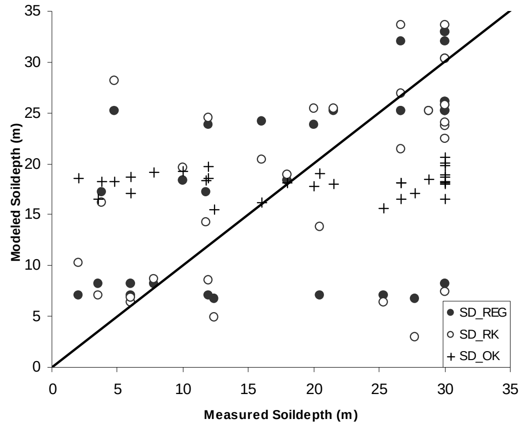
In this part of the scrip the actual kriging or regression takes place.
Line 31-33 creates a ‘gstat’ object, which is used in the kriging statement ‘predict.gestat’. This time the estimated semivariogram data is added to the object (‘model = vgmf’) which enables kriging.
Line 35-39 calculates the kriging and/or incorporates the calculated regression(of line 31-33) to spatial data.
newdata= refers to the landscape feature maps, used for regression. When using ordinary kriging, the array ‘gebied’ is used only for the area over which the calculation takes place. When regression is used, the array is used to assign the calculated regression results to the appropriate areas.
BLUE = true or false. When ‘true’, the statements returns only the trend estimates and linear regression is performed. When ‘false’ the statement returns kriging predictions.
debuglevel = prints process counter during prediction. Is used to see progress during calculations.
Line 37-39 plots the calculation.
Line 40-41 saves the calculation as a picture and ASCII-file.

D: Results









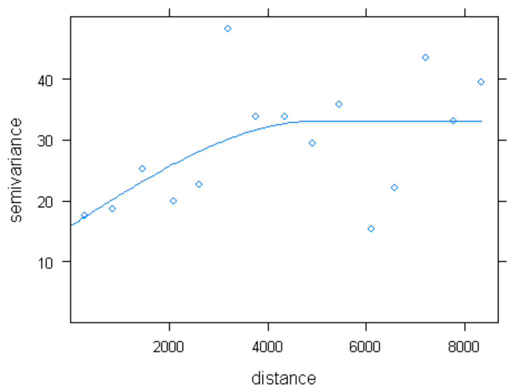
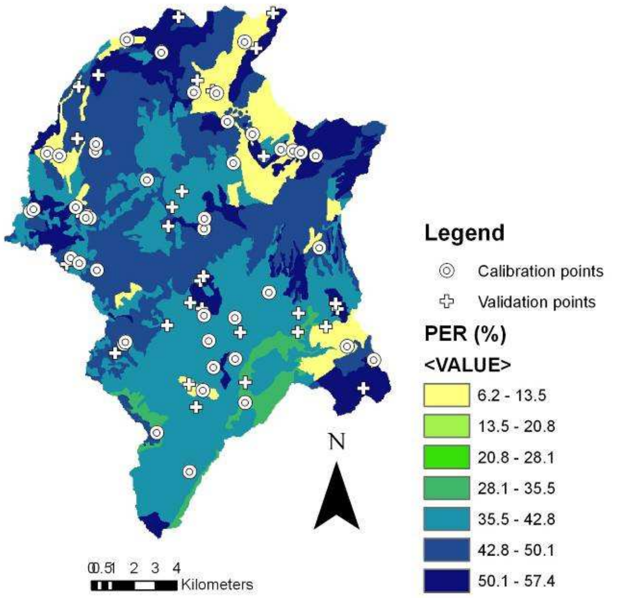


Fig D.1: Output for Soil Depth; a) The Interpolated map based on Regression, with calibration data and the validation data locations, b) The Interpolated map based Ordinary Kriging, c) The Interpolated map based Regression Kriging, d) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the regression method, ordinary kriging method and regression kriging method e) Semivariogram



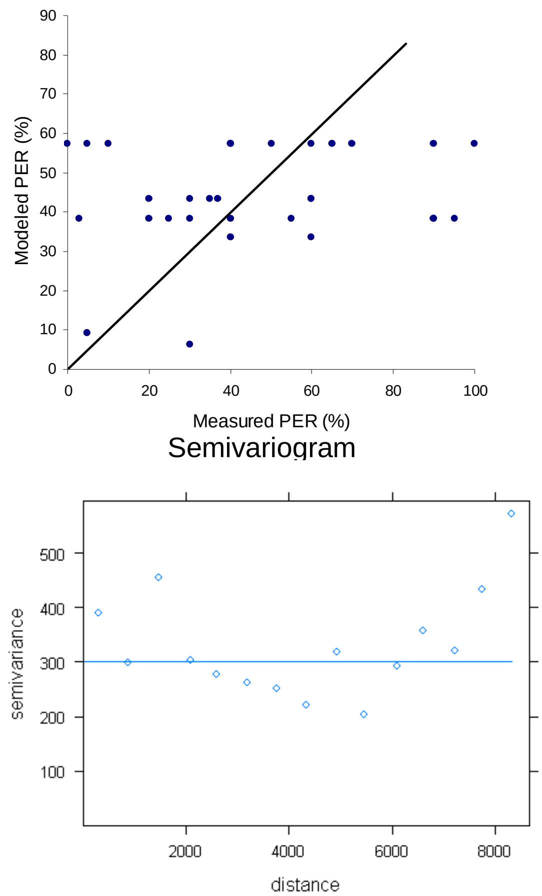
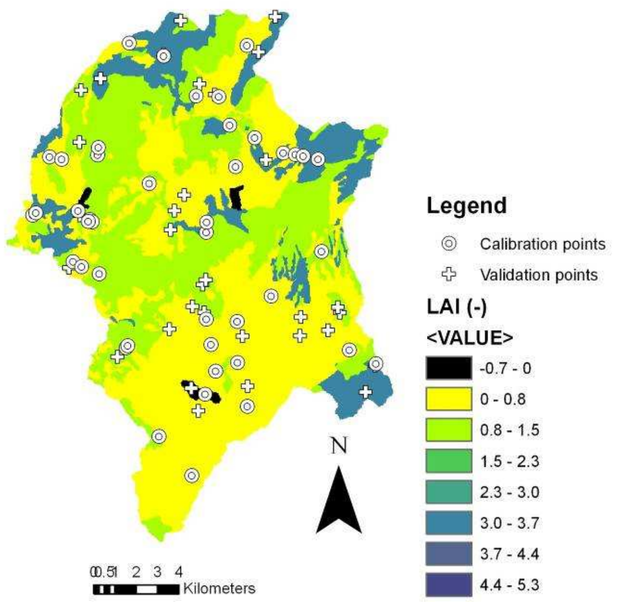


Fig D.2: Output for Plant cover; a) The Interpolated map based regression, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the regression method, c) Semivariogram.



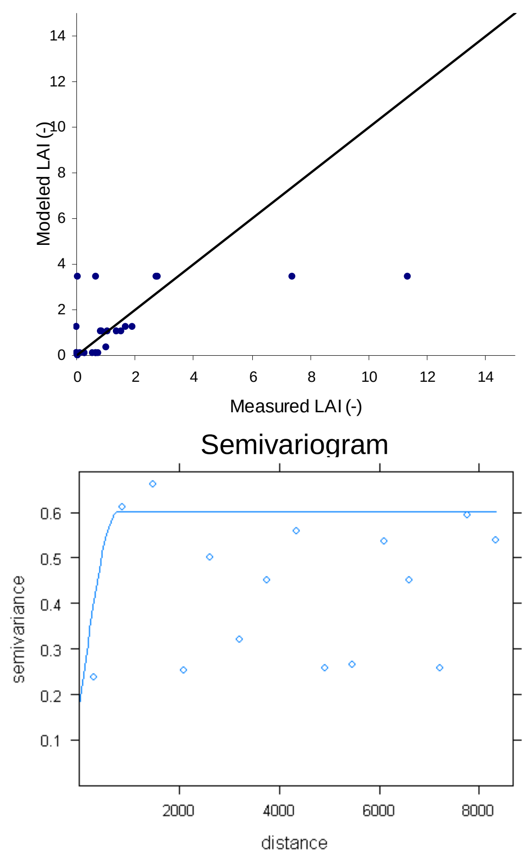
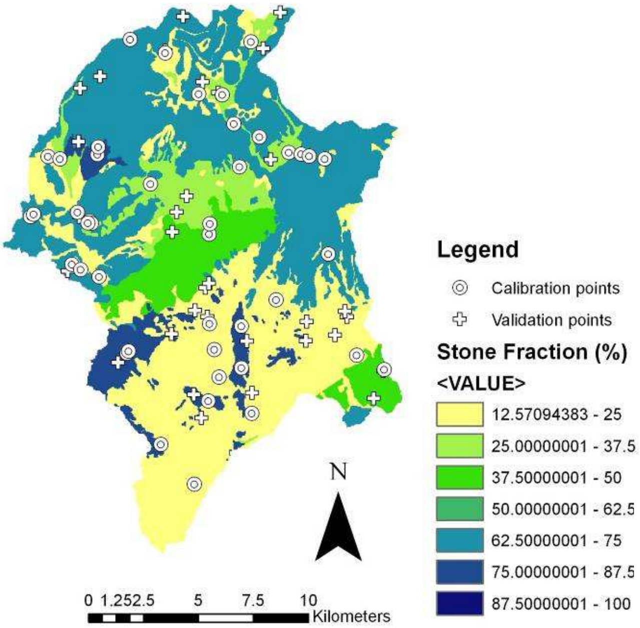


Fig D.3: Output for Leaf area Index; a) The Interpolated map based regression, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the regression method, c) Semivariogram.



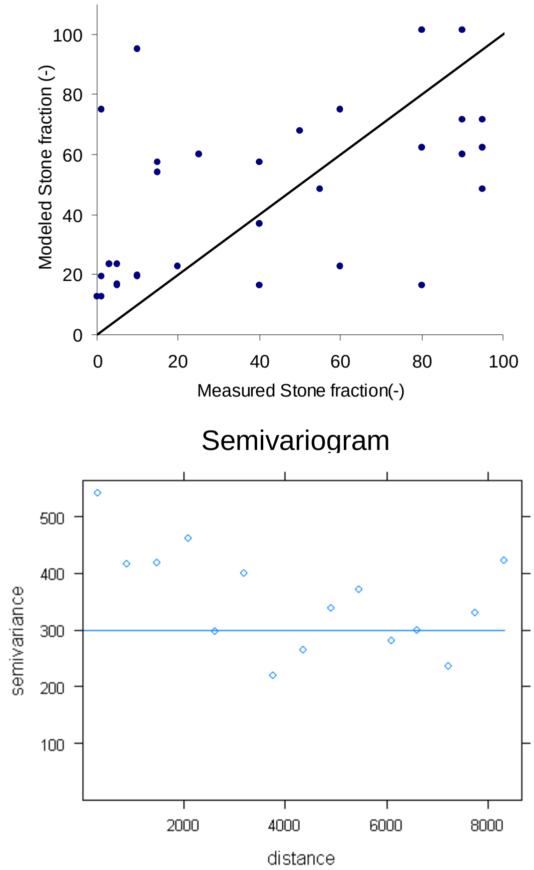
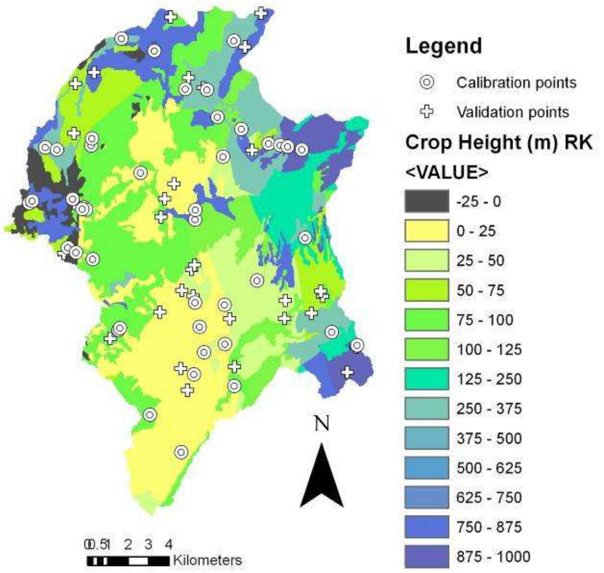


Fig D.4: Output for Stone fraction; a) The Interpolated map based regression, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the regression method, c) Semivariogram.



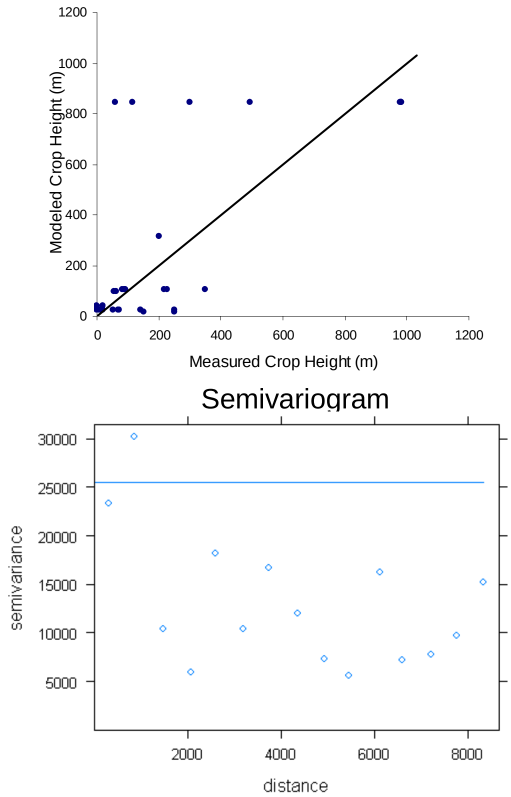
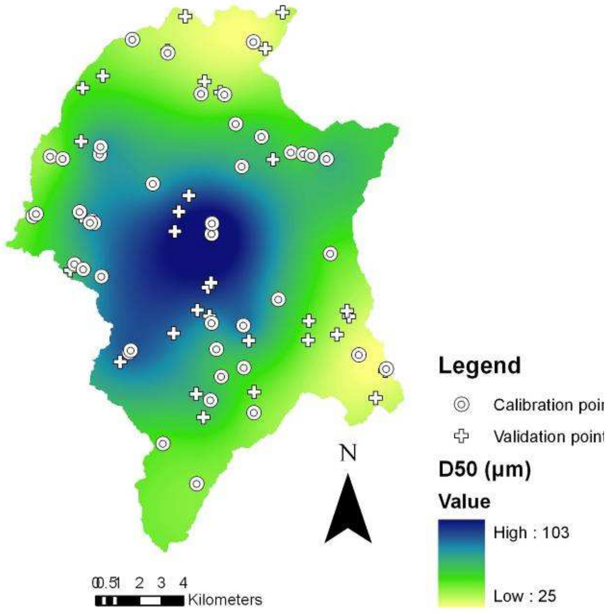
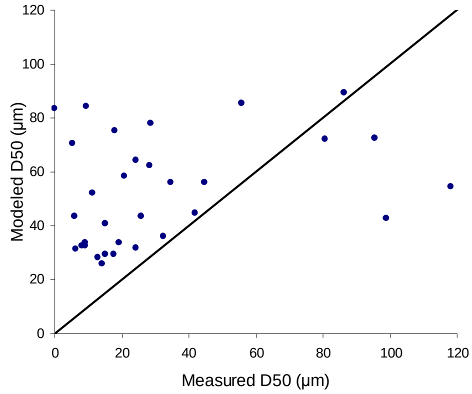


Fig D.5: Output for Crop Height; a) The Interpolated map based regression, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the regression method, c) Semivariogram.





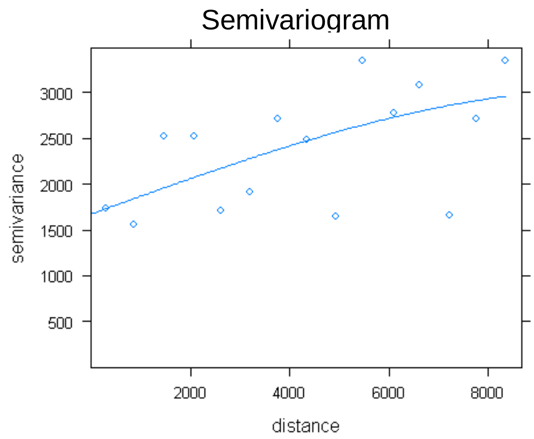
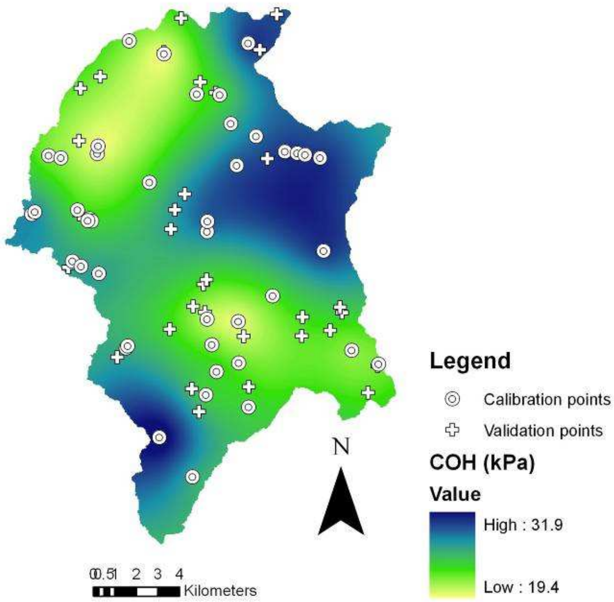
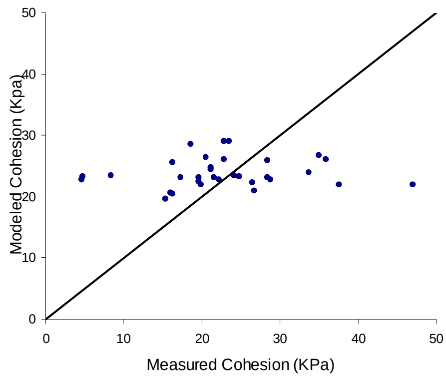


Fig D.6: Output for Grain size; a) The Interpolated map based Ordinary Kriging, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the Ordinary Kriging method, c) Semivariogram.





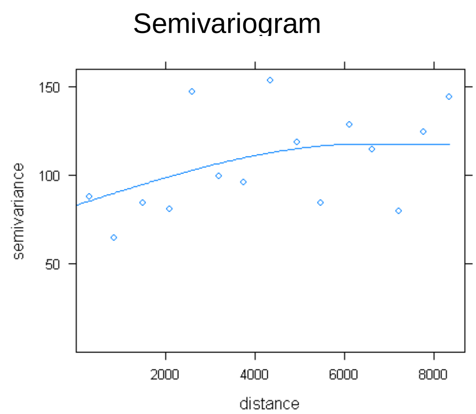
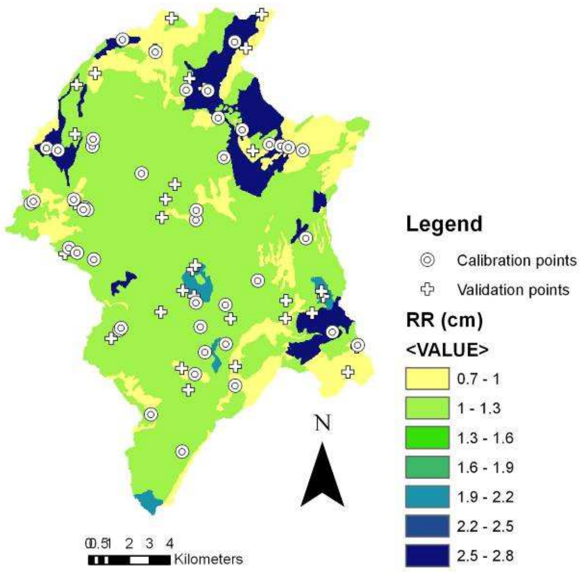
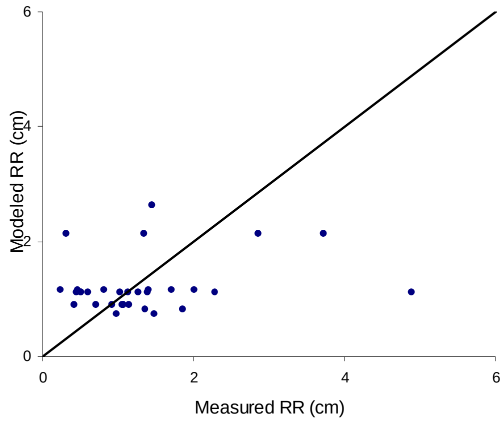


Fig D.7: Output for Cohesion; The Interpolated map based Ordinary Kriging, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the Ordinary Kriging method, c) Semivariogram.





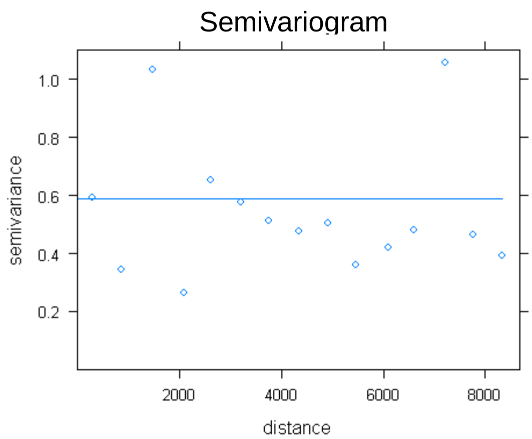
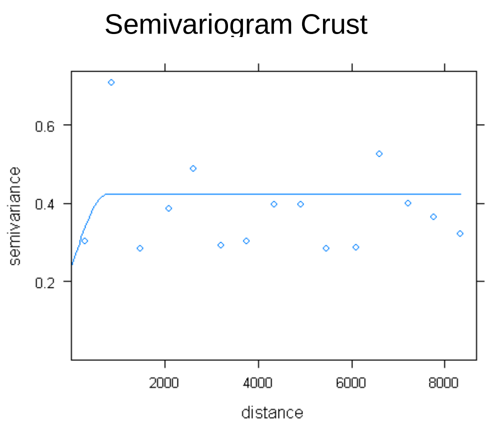
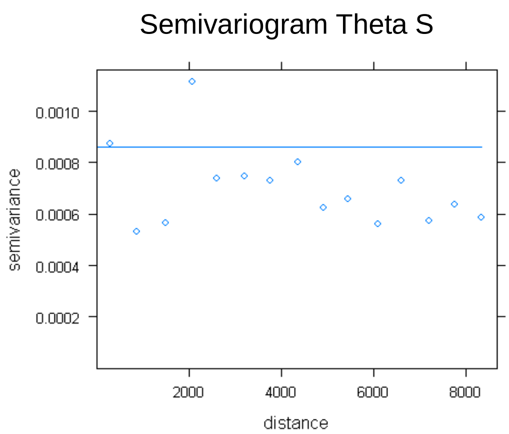


Fig D.8: Output for Plant cover; a) The Interpolated map based regression, with calibration data and the validation data locations, b) Scatterplot of the modeled versus the measured data (dots) with the 1:1 line (line) of the regression method, c) Semivariogram.





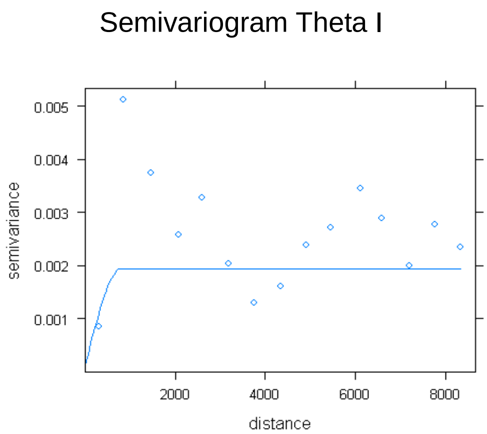


Fig D.9; a) Semivariogram od parameter crust, b) Semivariogram parameter Theta S, c) Semivariogram parameter Theta I