Université catholique de Louvain

Faculty of bioscience engineering

LBRTI2101A Data Science in bioscience engineering
Partim A : spatial and temporal data

Spatial analysis of pedologycal data of an agricultural land in Hesbaye

Students : Lemaire Romain - 50581700
Parys Louis - 72561700
Thonnard Julien - 06441800
Van Eetvelt Mattias - 16601800

Professor : Bogaert Patrik

Teaching assistant : Toussaint François

December 2022

Contents

1 Introduction

The purpose of the work is to analyze a dataset showing spatial dependencies. It is a dataset of soil data from an agricultural plot in Hesbaye. The crop present on the plot during the study was beet. Part of this plot has been cultivated "forever", another small part has been refilled and the rest was pasture until a more or less recent past as seen on the figure 1. The plot is located in the lower half of a northwest-facing slope, with an average slope of 2.2%. The overall dataset consists of about twenty variables observed at 176 nodes of a regular mesh.

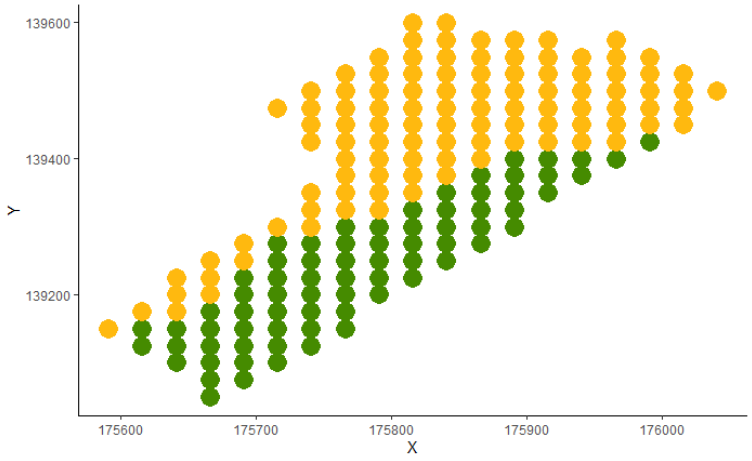


Figure 1: Division of the parcel

This report will be divided into six main parts. The first section will provide an introduction to the data provided for the project. The second part will conduct a deeper analysis of the data, including its distribution and correlations. The third section will focus on identifying the spatial dependencies of the various data through the use of variograms. The fourth section will make predictions on the data using two methods: the inverse distance weighting method (IDW) and kriging. The fifth section will again focus on predicting the data using another method: cokriging. Finally, the last part of the report will deal with the validation of the variograms via their confidence intervals.

1.1 Description of the dataset

The given dataset contains 176 observations of 6 variables. From the six variables, two are geographical variables and the other four are physico-chemical variables describing the soil's properties.

- x : Represents the horizontal coordinate in Lambert72.
- y : Represents the vertical coordinate in Lambert72.
- A : Represents the clay textural fraction in percent [%] .
- pH_{KCl} : Represents the potential of hydrogen of a soil solution measured in a solution of potassium chloride.
- CEC : Represents the cation exchange capacity, which is the total capacity of a soil to hold exchangeable cations. [meq/100g]
- $IB.samples$: Represents an empirical sensitivity crusting index.

The crusting index is calculated using the formula :

$$IB = \frac{1.5 \times LF + 0.75 \times LG}{A + 10 \times OM} \tag{1}$$

where:

LF = Proportion of fine silt [%]
LG = Proportion of coarse silt [%]
OM = Proportion of organic matter [%]
A = Proportion of clay [%]

1.2 Literature review on key concepts involved in the analysis

Cation exchange capacity (CEC) is a measure of a soil's ability to hold and exchange positively charged ions, such as calcium, magnesium, potassium, and sodium, which are essential for plant growth. The CEC of a soil is expressed in milliequivalents per 100 grams of soil (meq/100 g) and determines the amount of these nutrients that can be held and made available to plants. The crusting index is a measure of a soil's tendency to form a hard, compact surface layer when wetted and dried. Soils with a high crusting index can be difficult for plants to emerge from and for water to infiltrate.

CEC and Fertility
Soils with a high CEC tend to be more fertile because they have a greater capacity to hold and exchange cations, which can supply nutrients to plants and maintain a proper balance of nutrients in the soil. The CEC of a soil is influenced by factors such as the type and amount of clay and organic matter present, as well as the pH of the soil. Clay minerals and organic matter both have a high affinity for cations, and therefore contribute significantly to the CEC of a soil[1].

CEC and KCl pH
The pH of the soil can affect the CEC because it can influence the distribution and abundance of cations in the soil. At a lower pH, the soil may have a higher concentration of hydrogen ions (H+), which can displace other cations that are held onto the soil particles. This can lower the CEC of the soil. At a higher pH, the soil may have a lower concentration of hydrogen ions, allowing other cations to be held onto the soil particles more readily. This can increase the CEC of the soil.

CEC and Crusting Index
There is a link between the CEC and crusting index of a soil: soils with a high CEC tend to have a lower crusting index, and vice versa. The presence of a high amount of cations in a soil can help prevent the formation of a hard, compact surface layer when the soil is wetted and dried. Soils with a low CEC, on the other hand, are more prone to crusting because they have fewer cations to hold the soil particles together [2].

To conclude, the CEC and crusting index of a soil are important factors to consider when assessing the fertility and potential productivity of a soil. The CEC can be used to determine the appropriate amount of fertilizers and other soil amendments to optimize plant growth.

2 Exploratory data analysis

In this section, the exploration of the given dataset is presented. This part is important in order to notice some observations that will help later for the result analysis.

2.1 Summary

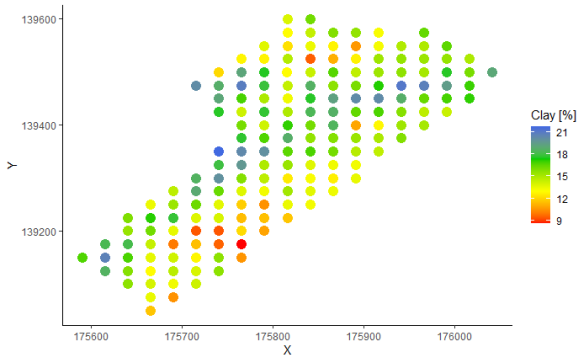
First information can be given via the dataset summary. This gives an idea of the range of measurements. A first useful information is that the table indicates that the variable *IB.samples* is poorly sampled and lacks measurements. This is useful information to know for further analysis.

```
       x                y                A              pH_KCL          CEC             IB.samples
 Min.   :175591   Min.   :139049   Min.   : 8.61   Min.   :5.550   Min.   : 6.594   Min.   :1.110
 1st Qu.:175741   1st Qu.:139249   1st Qu.:13.44   1st Qu.:7.110   1st Qu.: 8.820   1st Qu.:1.480
 Median :175816   Median :139374   Median :15.26   Median :7.365   Median :10.379   Median :1.710
 Mean   :175814   Mean   :139359   Mean   :15.24   Mean   :7.327   Mean   :10.318   Mean   :1.811
 3rd Qu.:175891   3rd Qu.:139474   3rd Qu.:16.84   3rd Qu.:7.580   3rd Qu.:11.539   3rd Qu.:2.150
 Max.   :176041   Max.   :139599   Max.   :21.69   Max.   :7.910   Max.   :14.828   Max.   :2.790
                                                                                   NA's   :76
```

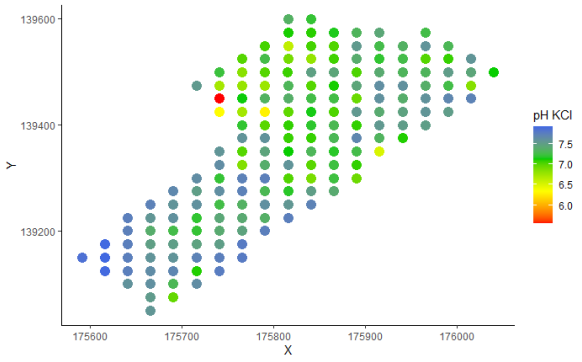
Figure 2: Summary of the different variables

2.2 Data mapping

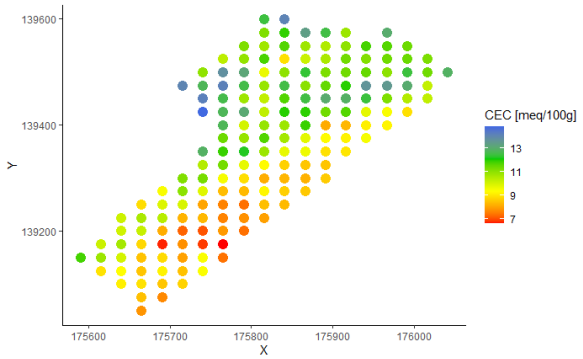
The mapping of the data allows to observe some links between the variables. It can be seen that the CEC and the clay both have similarities. When the clay values are high, the CEC values are also high and *vice-versa*. This was to be expected since the clay contributes to the CEC of the soil.
For the pH KCl, one point has a very low value. It is expected to be an outlier in the later analysis.
Finally, it is observed that the beat index was not sampled at all coordinates compared to the others variables. This data mapping also enables to observe the difference of soil uses between the northern and the southern sides of the parcel.



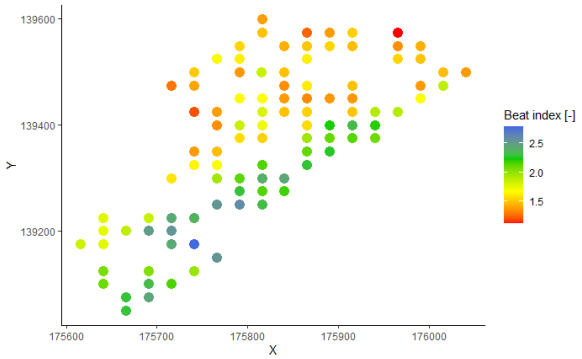
(a) Clay



(b) pH KCl



(c) CEC



(d) Beat index

Figure 3: Map of the four different variables

2.3 Examination of the variables's distributions via histograms

As seen on figure 4 and 5 the distribution of both the variable clay and pH KCl can be described has being close to a normal distribution. It is therefore unnecessary to apply any transformation to either of these variables.

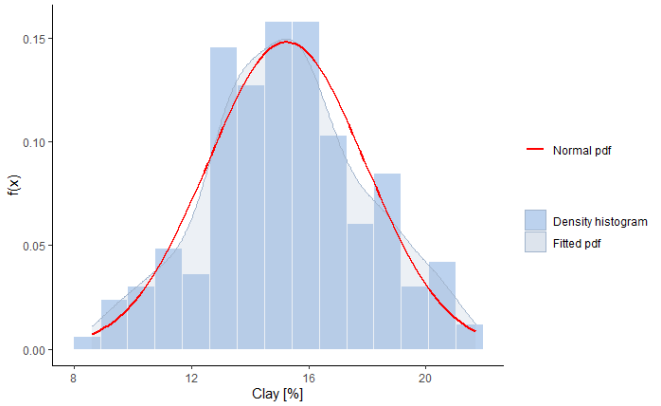


Figure 4: Clay density histogram

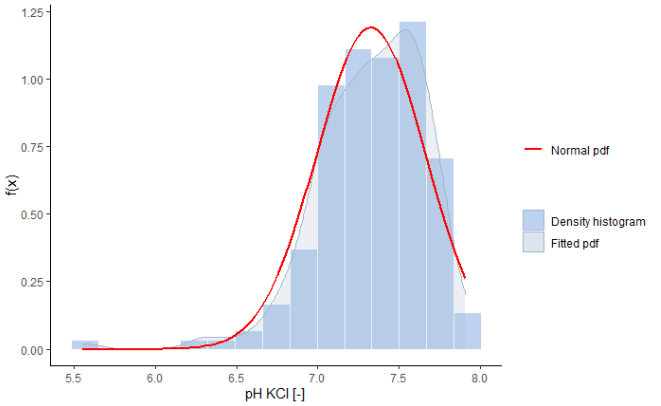


Figure 5: pH KCl density histogram

As it was graphically seen in the section 2.2, the CEC histogram contains some high values affecting the normality of its distribution. These high values are allegedly linked to the difference of land use.

Concerning the beat index's distribution, it is far from being normal. This could either be explained by the land use of the parcel impacting the beat index's measure or by the under-sampling of the zone resulting in a high number of NaN values contained in the dataset.

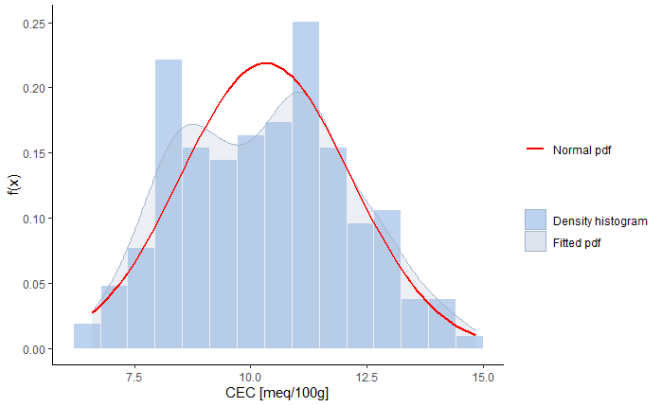


Figure 6: CEC density histogram

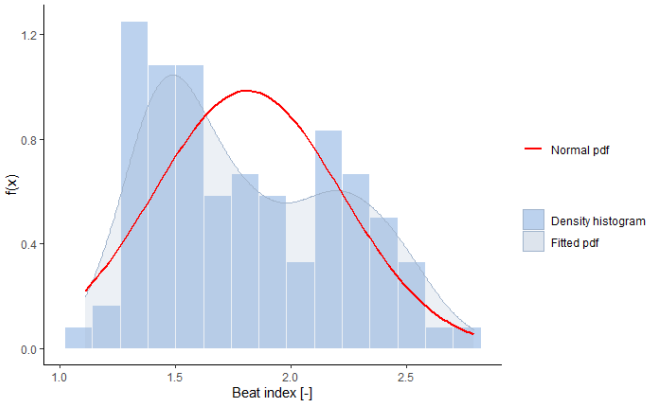


Figure 7: Beat index density histogram

2.4 Correlation

As it has been observed before, there is a strong correlation ($\rho = 0.75$) between *clay* and *CEC* as these parameters are closely related. This was to be expected since the presence of clay has a positive impact on the CEC of the soil as written in the scientific literature in the section 1.2.

The other major correlation value is noticed between the beat index (*IB.samples*) and the *CEC* ($\rho = -0.96$). This negative correlation can be explained by looking at the beat index's formula. As mentionned in the section 1.2, the CEC's intrinsic properties are mostly due to the soil's clay content for the permanent charges and to organic matter (OM) for the variable charges. As the permanent charges parameters involved in the beat index formula are immutable, the only variable parameter affecting the index is the OM. As seen on the IB's formula 1 the OM parameter is on the denominator, it is therefore presumed that a low CEC value which is mostly due to a low OM content, will create a high IB index as shown by the negative correlation value. A positive correlation was expected between *IB.samples* and *clay* but this is not the case here ($\rho = -0.61$).

It is observed that X and Y are correlated with each other but this does not seem relevant since this correlation seems to depend on the shape of the plot. Moreover, X does not seem to be correlated with any other variable, which is not the case for Y, which is correlated with the variables *CEC* and *IB.samples*.
For the first one, this correlation seems to be due to the slope of the parcel, favoring the accumulation of organic matter in the bottom of it.
The second correlation of Y, according to *IB.samples* this time, seems to be the result of the arrangement of the two zones of the plot. Indeed, the bulk of the first zone is to the north and *vice versa*. These two areas having a different activity directly impacting the beat index, this correlation is the result.

Finally, it is noticed that the correlation between *CEC* and *pH KCl* is negative ($\rho = -0.37$). This goes against what have been said in the section 1.2 of the report: "hight values of pH tend to increase the CEC and *vice-versa*". Still, it is important to note that the CEC of a soil can be affected by a number of factors, including the type and amount of clay minerals present, the type and amount of organic matter present, and the pH of the soil. Therefore, it is possible that the negative correlation between the pH of the soil in a KCl solution and the CEC could be due to a combination of these factors.

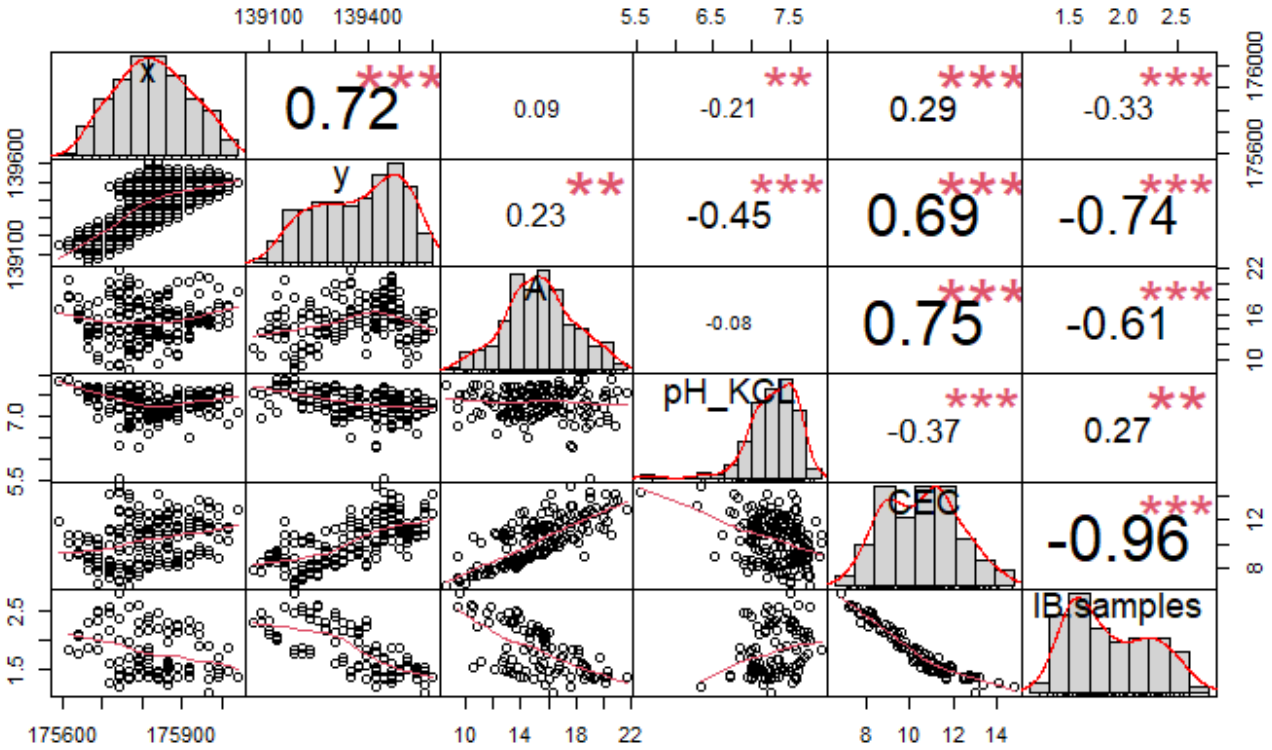


Figure 8: Correlation chart between the coordinates and the physico-chemical variables

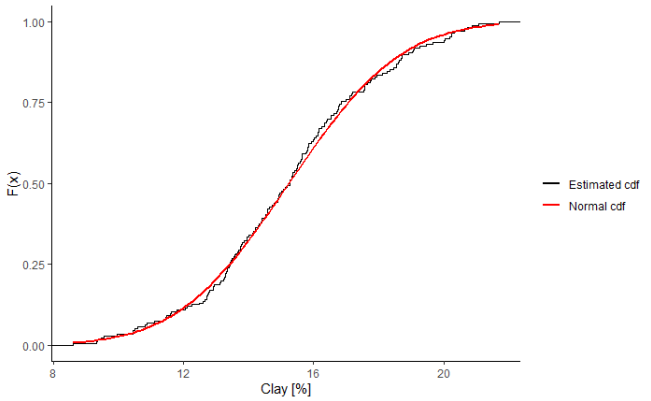
2.5 Comparison of CDF with and without outliers

CDF stands for "Cumulative Distribution Function." It is a function that gives the probability that a random variable X is less than or equal to a particular value x.
The cumulative distribution function (CDF) is defined as

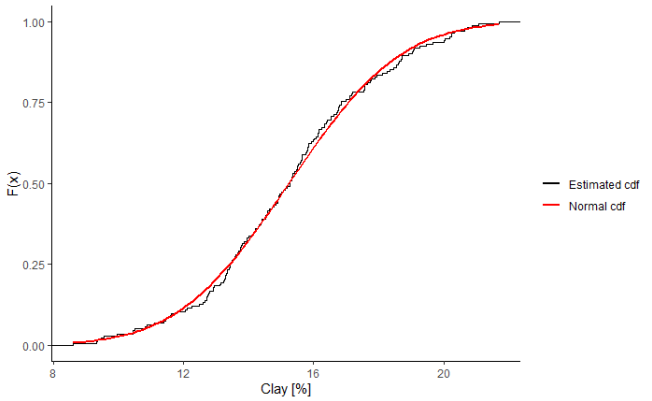
$$F(x) = P(X \leq x)$$

where F(x) is the CDF, X is the random variable, and x is a particular value.

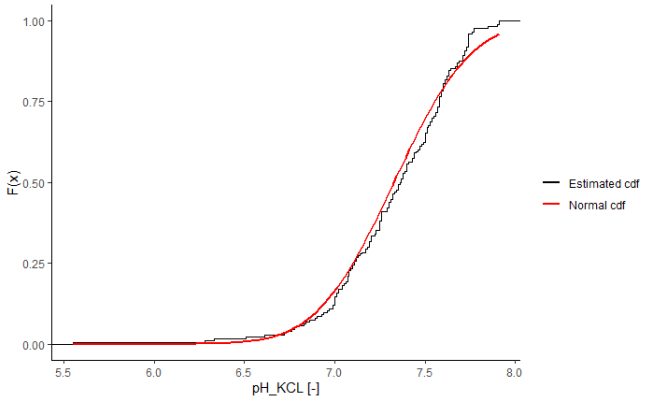
The figures 9a and 9b show that the withdraw of the outliers will not change the shape of the distribution. It will also be the case for the CDF of the *CEC* and *IB.samples* as seen in the figures in the section 10.3. On the other hand, the suppression of outliers has changed the shape of the CDF of the *pH KCl* as can be seen in the graph 9d compared to 9c. However, it can not be said that the estimated CDF does fit the normal CDF better after that manipulation.



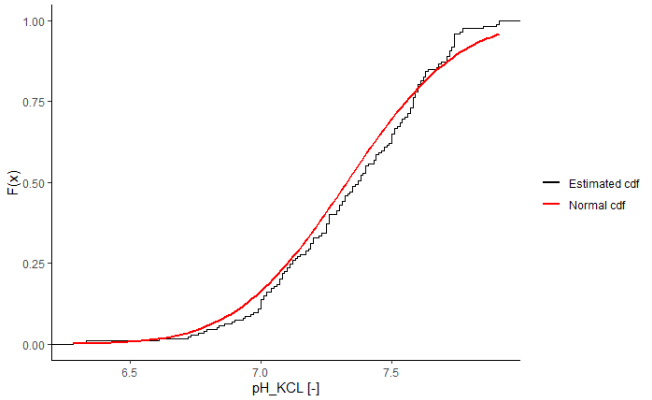
(a) Clay CDF with outliers



(b) Clay CDF without outliers



(c) pH KCl CDF with outliers



(d) pH KCl CDF without outliers

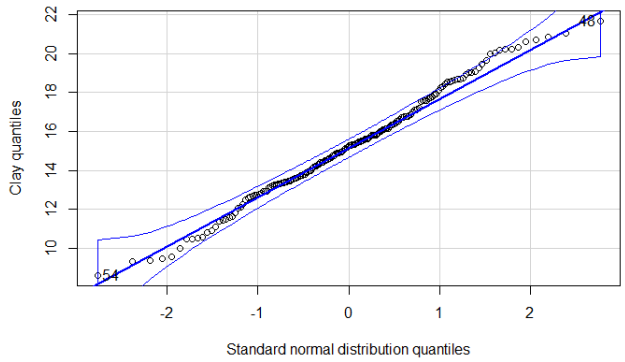
Figure 9: CDF

2.6 QQ plot analysis

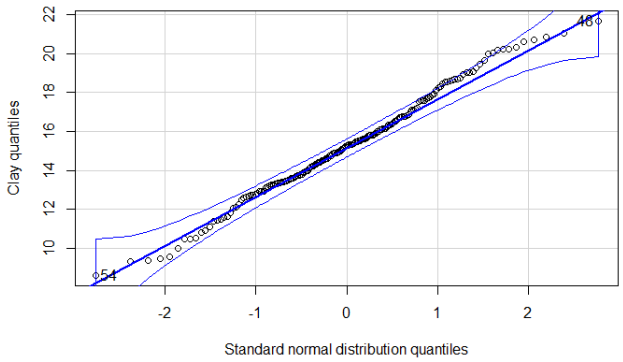
A quantile-quantile plot (QQ-plot) is a graphical tool used to compare a given distribution to a theoretical distribution, in this case a normal distribution. This diagram is used to evaluate the normality assumption. This hypothesis is considered to be correct if all the points are present approximately in the form of a straight line.

The clay's QQ-plot in the figure 10a is relatively similar to those of *CEC* and *IB.samples* and will therefore be used to discuss these three distributions. It is noticed that these plots do not have high outlier values, so even if the curve does not perfectly match the normal distribution, it is fair to say that they are relatively close to it. Consequently, it is hard to notice any difference concerning the QQ-plots with and without outliers for these three variables as seen on the figures 28b and 28d in the section 10.4.

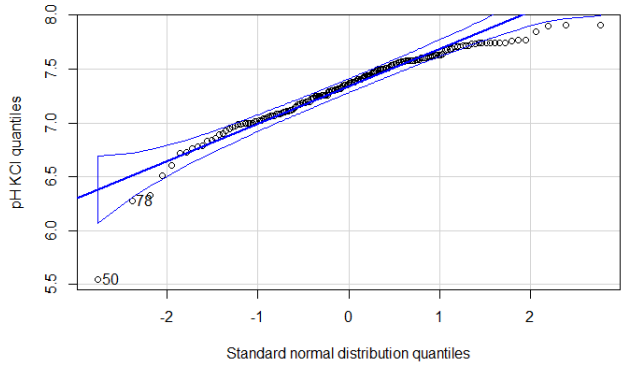
Concerning the *pH KCl*'s QQ-plot, it is noticed that a vertical orientation shift is observed after removing the outlier (pH = 5.55) as seen on the figure 10d.



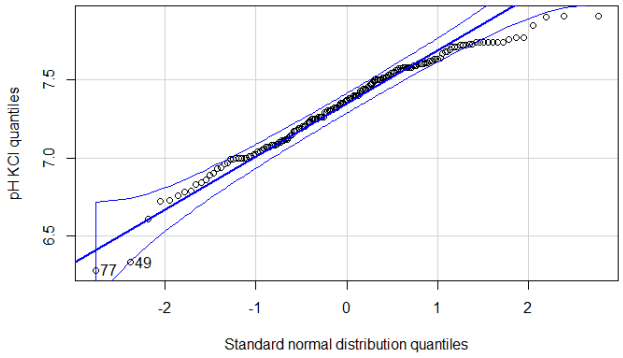
(a) Clay QQ plot with outliers



(b) Clay QQ plot without outliers



(c) pH KCl QQ plot with outliers



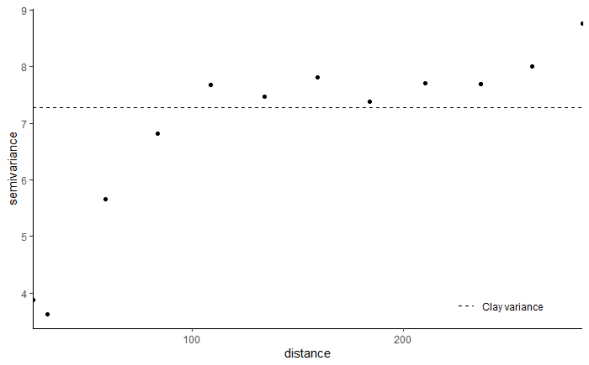
(d) pH KCl QQ without outliers

Figure 10: QQ plots

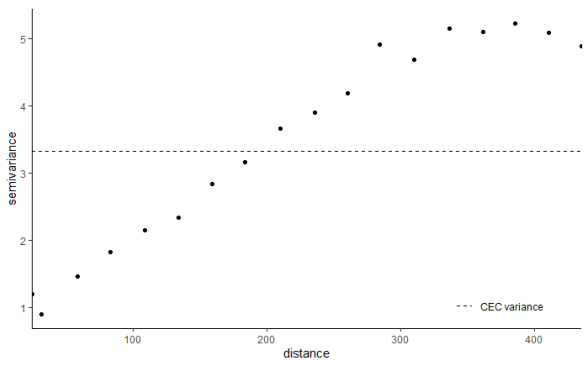
3 Spatial dependency : analysis and modelling

This part of the work is essential for the prediction part. Indeed, the analysis of spatial dependence will be useful to make the predictions in the rest of the report. The goal with this analysis is to build the variogram for each variables.

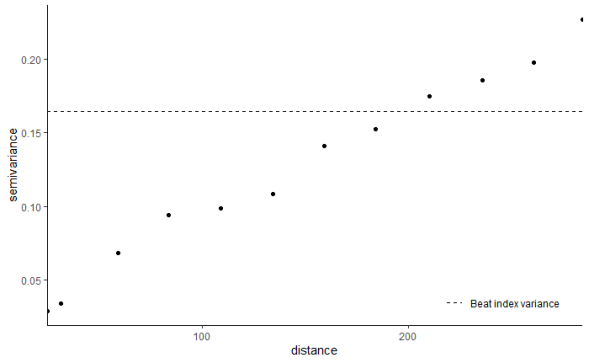
A simple plot of the variograms without treatments shows that it does not tend towards the variance of each variable as seen on the figure 11. This proves that the variables are not stationarity of order 1 which is necessary. To reach a stationarity of order 1, it is require to remove the trend of the variable. This is what it is done in several steps in this section.



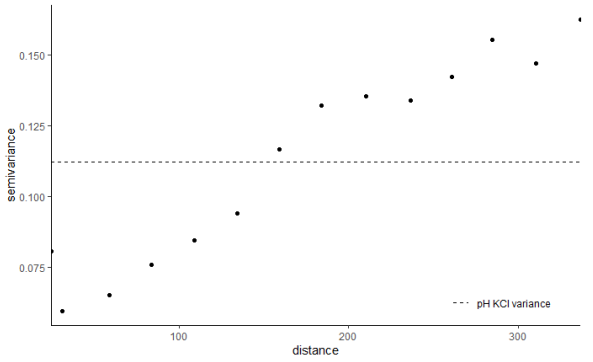
(a) Clay



(b) CEC



(c) Beat index



(d) pH KCl

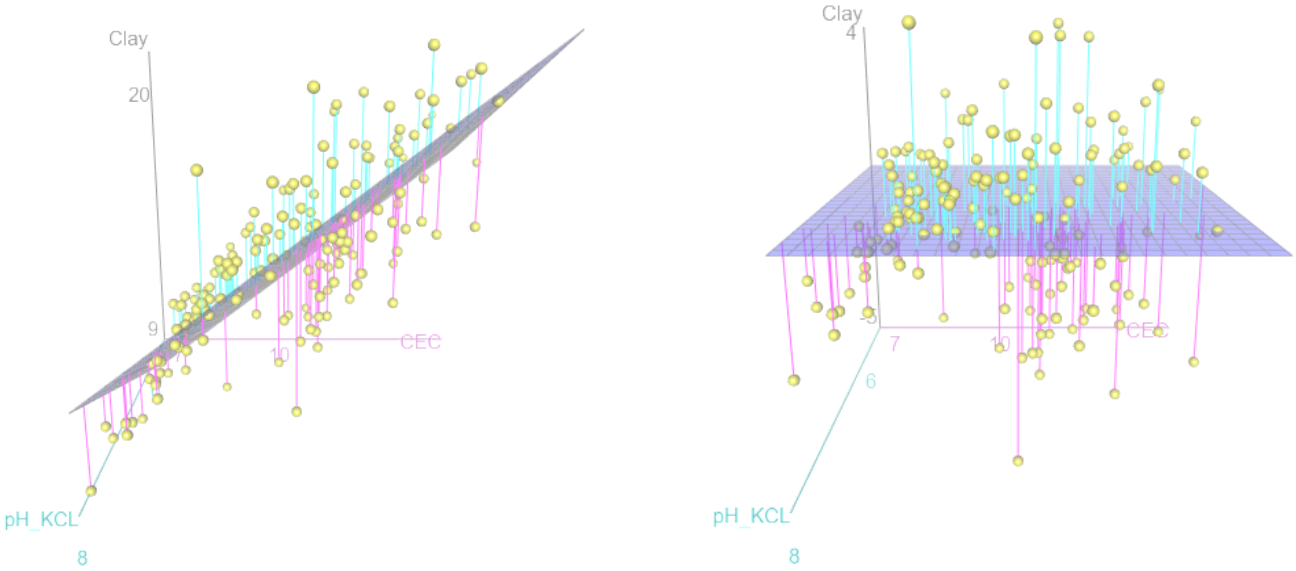
Figure 11: Variogram before trend removal of the four variables

First, 3D graphs of the points were made to highlight the trend of the variable in relation to two others. The trend for each variable was tested as a function of X and Y but no conclusive results of any trend were observed, as it can be seen on the figures 12 and 13.

Figure 12: Clay as function of X and Y

Figure 13: Beat index as function of CEC and Y

The analysis was therefore turned to the other variables given. With these the results were more conclusive. Indeed, as it can be seen on the figure 14a there is a trend for clay as a function of *CEC* and *pH KCl*. These three variables show a trend between them and for this reason the trend in *CEC* was evaluated as a function of *clay* and *pH KCl* and the same for *pH KCl* as it can be seen on the figures 29a and 29e. Finally, a 3D plot of the residuals shows that the trend is well removed as it is shown on the figures 14b, 29b and 29f.
The three first variables are now good but not yet for*IB.samples*. This variable depends on the other three, so a choice must be made as to which two variables to choose to remove the trend. In this case, *clay* and *CEC* were chosen because they showed the strongest one. The results of the 3D plot of the trend as well as the 3D plot of the residuals without the trend are shown on figures 29c and 29d.
In this analysis, *IB.samples* was not chosen to visualize and remove the trend due to its under-sampling compared to the other variables and therefore its choice would have resulted in a loss of data.

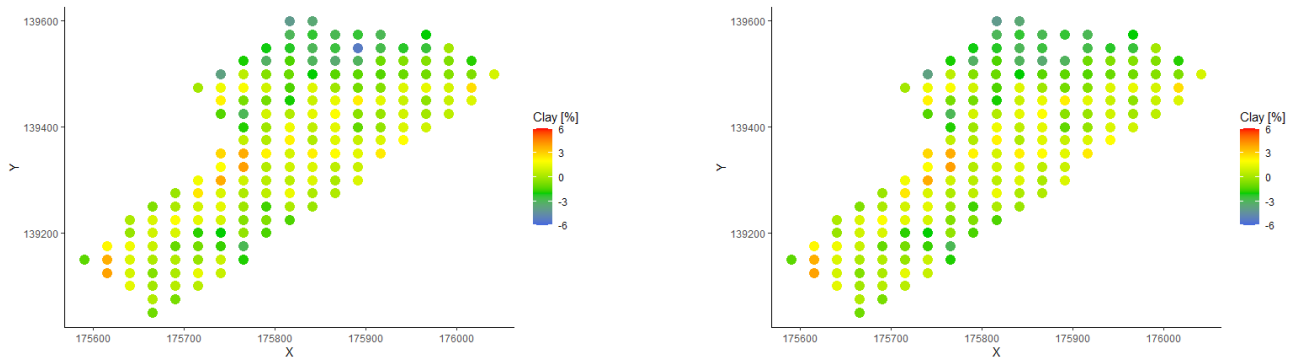


(a) Clay data

(b) Clay residues

Figure 14: 3D point cloud of the clay samples

It is possible to observe on the 3D graphs of the residuals that some values deviate greatly from the plane as can be seen on the figures 29d and 29f for instance. These values are considered as outliers and it will be correct to remove them. The method used is to ignore the residuals with a value greater than the mean plus or minus three times the standard deviation. After its implementation, it is noticed that one value was considered as an outlier for the clay variable. For *CEC*, no values are considered as such while two are for *pH KCl* and only one for *IB.samples*. These results are shown on the figures 15b, 26b, 26d and 26f.

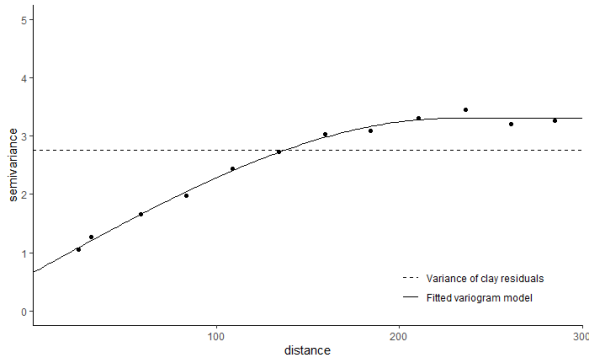


(a) With outliers

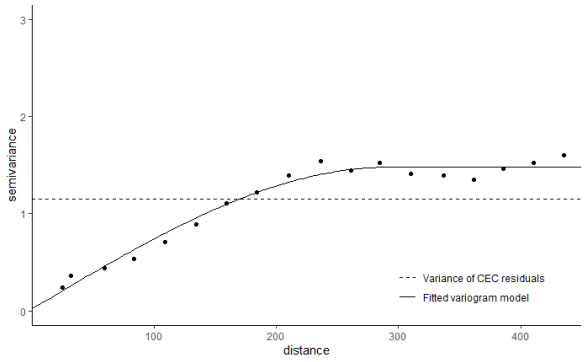
(b) Without outliers

Figure 15: Clay residual maps

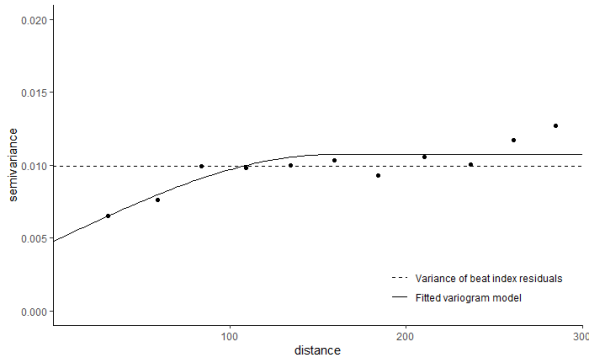
Now that the data are processed, it is possible to make variograms for each of these variables without the outliers. The result of these variograms is shown on figure 16.
As said before, these variograms will be useful for the analysis of spatial dependence since they measure dissimilarity as a function of distance. The expected trend of this one is a monotonically increasing function with ||h|| with no negative values taken into account. These two conditions are well observed on the four variograms shown in the figure. Moreover, it is also expected that the variograms tend towards the variance of the studied variable. This is partly observed on all the variograms. Indeed, they tend towards a value around the variance but not exactly towards the variance itself. This could be partly explained by the two modes of exploitation of the study area. A part by part study could possibly make the variograms of each variable tend towards the respective variance of each part of the area.



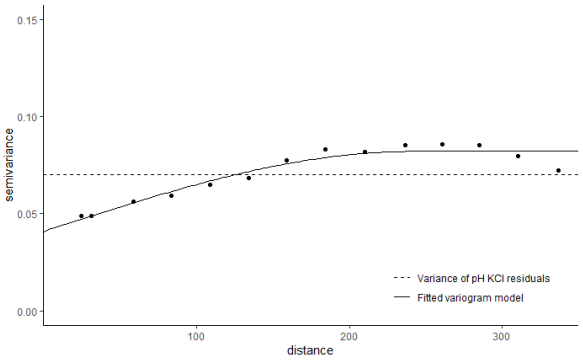
(a) Clay



(b) CEC



(c) BI



(d) pH KCl

Figure 16: Variogram of the four variables

The equation used to model these variograms is the following:

$$\gamma(h) = \begin{cases} c_0 + c\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3\right) & if\ h < a, \\ c_0 + c & if\ h \geq a \end{cases} \tag{2}$$

where c_0 is the value of the semivariance for a distance of $h = 0$, c is the difference between the variance and the value of c_0 and a is the value of the distance needed for the model to reach its plateau. These parameters were evaluated to best fit the model of each variable. The equations with these parameters are presented below.

$$\gamma(h)_{clay} = \begin{cases} 0.78 + 2.07\left(\frac{3}{2}\frac{h}{224.42} - \frac{1}{2}(\frac{h}{224.42})^3\right) & if\ h < 224.42, \\ 0.78 + 2.07 & if\ h \geq 224.42 \end{cases} \tag{3}$$

$$\gamma(h)_{CEC} = \begin{cases} 0.20 + 0.28\left(\frac{3}{2}\frac{h}{239.33} - \frac{1}{2}(\frac{h}{239.33})^3\right) & if\ h < 239.33, \\ 0.20 + 0.28 & if\ h \geq 239.33 \end{cases} \tag{4}$$

$$\gamma(h)_{BI} = \begin{cases} 0.0048 + 0.006\left(\frac{3}{2}\frac{h}{157.02} - \frac{1}{2}(\frac{h}{157.02})^3\right) & if\ h < 157.02, \\ 0.0048 + 0.006 & if\ h \geq 157.02 \end{cases} \tag{5}$$

$$\gamma(h)_{pHKCl} = \begin{cases} 0.035 + 0.043\left(\frac{3}{2}\frac{h}{214.16} - \frac{1}{2}(\frac{h}{214.16})^3\right) & if\ h < 214.16, \\ 0.035 + 0.043 & if\ h \geq 214.16 \end{cases} \tag{6}$$

4 Variables prediction

Various variables have been sampled in the field at given locations. The goal of this section is to predict the value of those variables at locations where there was no sampling. This is achieved by using two methods, namely the Inverse Distance Weighting method (IDW) and the kriging method.

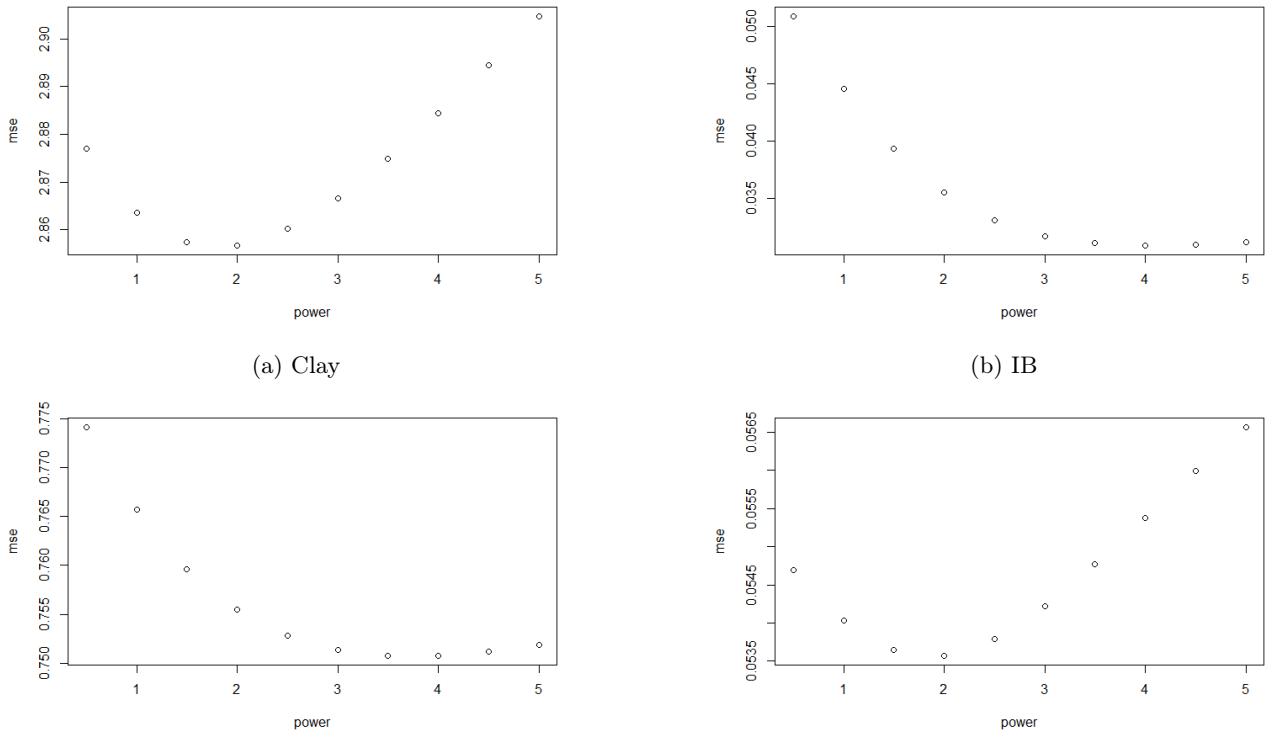
As explained in section 3, there is no spatial dependency between any of the variables and the X and Y coordinates. Therefore the predictions are directly calculated using the raw data and not the residuals.

4.1 Inverse Distance Weighting (IDW)

This first method gives a weight λ_i to the predicted value based on the inverse of the distance between a given point x_i and the sample point x_0. The rate at which the weights decrease depends on the value of θ. For all distances $||\mathbf{h}||_{0i}$, the weights are defined as

$$\lambda_i = \frac{\frac{1}{||\mathbf{h}_{0i}||^\theta}}{\sum_{j=1}^{n} \frac{1}{||\mathbf{h}_{0j}||^\theta}} \qquad where\, \theta > 0 \tag{7}$$

In order to apply the IDW method, it is first needed to determine the optimal value of the parameters *nmax* and θ, where *nmax* is the number of nearest observations used. This is achieved by cross-validation, i.e. the value at a sample location is computed without considering the value at this location using the IDW method. This is done for all the sample locations and for various arbitrary values of θ and *nmax*. Then, θ and *nmax* are chosen in order to minimize the mean squared error (MSE) between the predicted value and the actual value, as shown in figure 17.



(a) Clay

(b) IB

(c) CEC

(d) pH KCl

Figure 17: Mean Squared Error (MSE) for different values of θ computed for four different variables

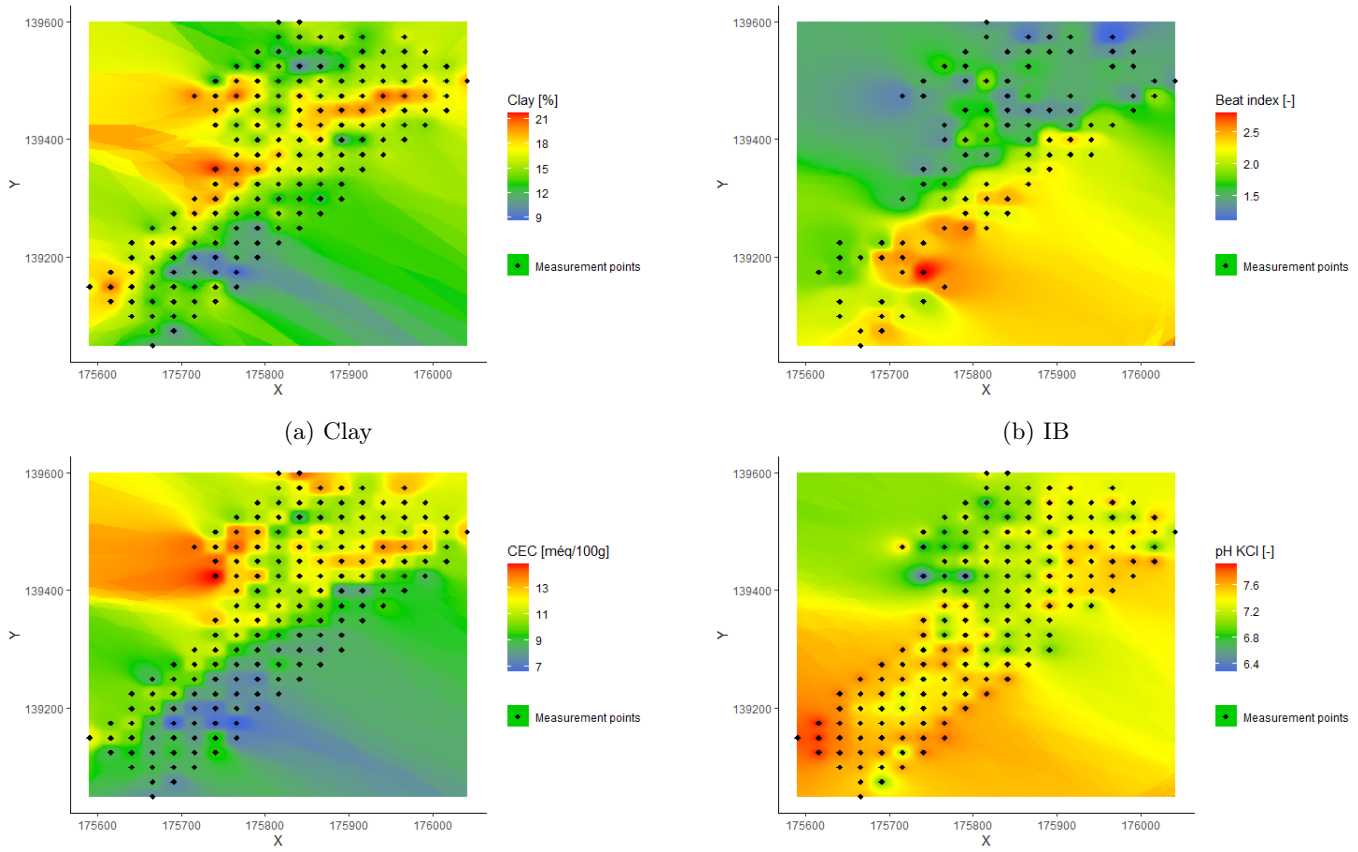
The values of θ and *nmax* are chosen based on figure 17. The optimal values are displayed on table 1. It is important to note that the *maxdist* parameter was set to 290 in order to have a prediction on the whole grid.

Table 1: Optimal values of θ and *nmax* minimising the MSE

Variable	Clay	CEC	pH KCl	Beat index
θ	2	4	2	4
nmax	5	6	21	17
mse	2.856	0.75	0.0535	0.0309

It should be noted that the lower the value of θ, the smoother the results are. Indeed, for a higher value of θ, the values of the weights decrease much faster with the distances. Furthermore, for a same distance the weight will be lower than with a lower value of θ.

When comparing figure 18c and 18d, the impact of smoothness is particularly visible. Prediction of CEC where made using a value $\theta = 4$. Predictions around a sampling point tend to converge very quickly towards the actual value, creating very important changes for a small distance. This is represented by squared "patches" of colors around sampling points. This behaviour, called an artifact, is less present for the predictions of pH KCl where a value of $\theta = 2$ was used. It should be noted that in this case, the grid size has a value of 1. Using a greater grid size may reduce this behaviour. A more in depth discussion is presented in section 4.3.



(a) Clay

(b) IB

(c) CEC

(d) pH KCl

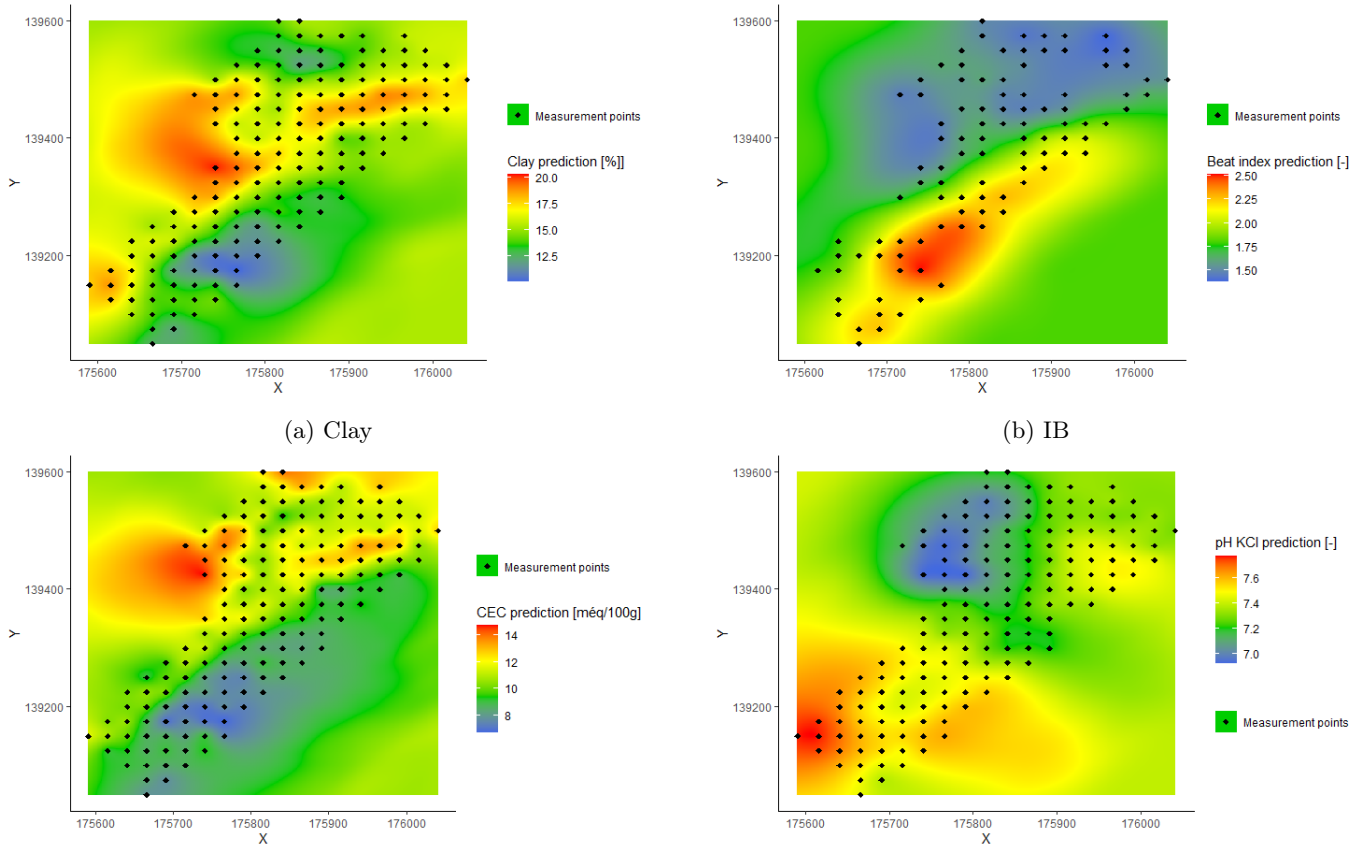
Figure 18: Predicted values for four different variables using the Inverse Distance Weighting (IDW) method

Another way to visualize the smoothness is to represent the contour plots or the 3D surface plots for the given variables. These plots are displayed in the appendix at page 24 for reader's information.

4.2 Kriging method

The second method is the kriging method also know as the Best Linear Unbiased Predictor (BLUP). This implies that the weights used to calculate the predicted values must fulfill the following conditions (i) the mean of the error is null and (ii) the variance of the error is minimal.

First the experimental variogram is computed. Then, a modeled variogram is defined after the experimental variogram and is fitted using a non-weighted least squares method, specifically the ordinary least squares (OLS) method, as represented in figure 16. This allows to have a continuous function that can be applied for any distance h, where the experimental variogram only gives discrete values. Finally, the weights are computed solving a system of equations containing values from the modeled variogram such as the solutions respect the conditions stated before. This is done by introducing a Lagrange multiplier.
The predicted values for all four variables are plotted and shown in figure 19. When comparing those results with the results from the IDW method, an absence of square "patches" is observed, confirming that this solution is indeed smoother.



(a) Clay

(b) IB

(c) CEC

(d) pH KCl

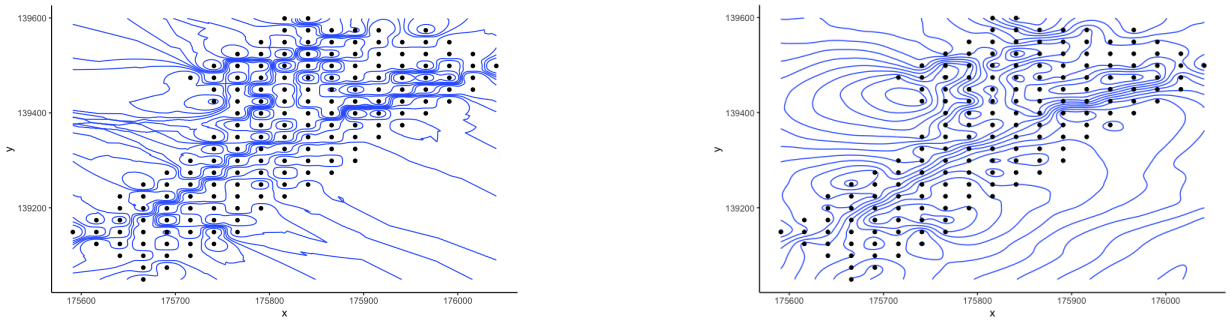
Figure 19: Predicted values for four different variables using the Krigin method

Figures 19a and 19c both present a clear separation roughly along the upward diagonal. This corresponds to the two zones described in section 1.

4.3 Comparison between IDW and kriging

The IDW and kriging method are both exact interpolators, meaning that the predicted value is the same as the sample value for a given sampling point. However, kriging is the preferred method since there are no artifacts unlike the IDW method, as mentioned previously. Those artifacts stand out when observing figure 20a. For the space between two neighbour points, the predicted value varies greatly as illustrated by the contour lines. However this is not the case when using the kriging method where values are smoother as it can be seen on figure 20b.

Another way to visualize the smoothness is to plot the variance of prediction. This is straight forward for kriging as the gstat object returned by the *krige* function computes the variance of prediction, but not for the IDW method. A solution is to use a modelled variogram and the IDW's weights but due to lack of time this solution was not further investigated.



(a) IDW ($\theta = 4$)

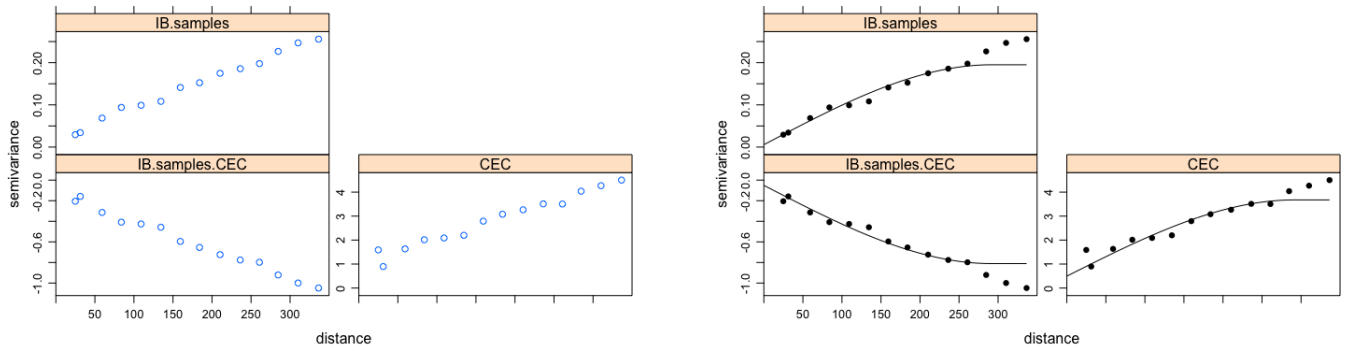
(b) Kriging

Figure 20: Contour plots for CEC predicted values using two different methods

5 Cokriging prediction of the IB variable using CEC data

Cokriging is the multivariate equivalent of kriging where additional variables are used to improve the interpolation predictions. This extra variable if often referred as the covariable [3]. In this case, *CEC* is the covariable and is used since it has the highest correlation with *IB.samples*. The cokriging method is particularly useful when some values are missing, or when measuring the target value is more complicated [4]. In this dataset the variable IB is undersampled compared to the other variables.
First, the experimental cross-variogram is build. Then, a linear model of co-regionalisation (LMC) is fitted. It should be noted that all three variograms and cross-variograms are first fitted then each partial sill is approach using the least squares method, therefore the range parameter is not optimized [5].

Both cross variograms in figure 21 show a negative slope. This is explained by the fact that *CEC* and *IB.samples* are negatively spatially correlated. Indeed the correlation has a value of $\rho = -0,96$. Furthermore, for a distance $h = 0$, the semivariance is not null. This represents the variability that is not explained by the spatial dependence, however this value is relatively small and is lower than $-0,1$.



(a) Variogram and cross variogram

(b) Variogram and cross variogram with the LMC

Figure 21: Variogram and cross variogram for cokriging

Now that the LMC has been modelled, the extra samples of *CEC* are used to, hopefully, improve the predictions of the beat index. Figure 22 shows the result. When comparing this result with figure 19b, the covariable does not seem to bring much information above the upward diagonal part of the figure, unlike the lower part of the figure where the predicted value of *IB.samples* is lower. Both mentioned figures are presented side by side in the appendix at page 24 for ease of comparison.

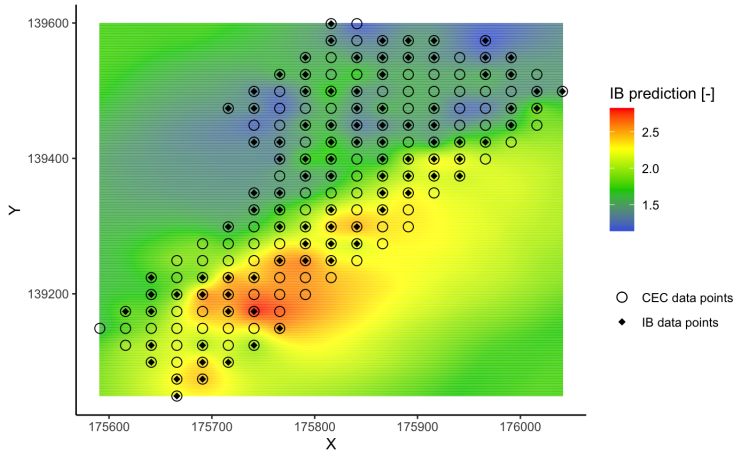
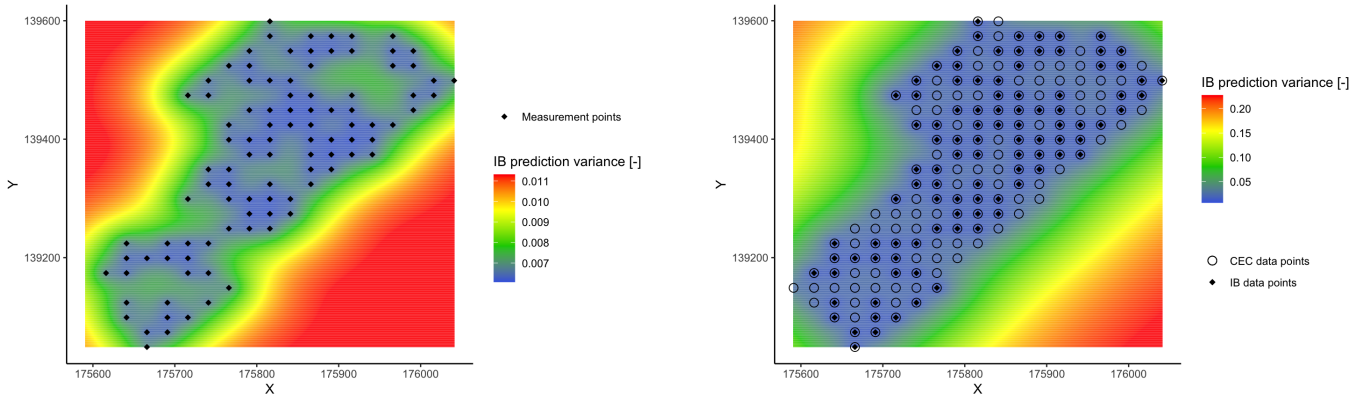


Figure 22: IB prediction using cokriging with CEC data

5.1 Comparing kriging and cokriging variance of prediction

Another way to visualize the information brought by the use of a covariable is to compare the variance of prediction when using the kriging or cokriging method. This is shown on figure 23.



(a) Kriging

(b) Cokriging

Figure 23: Prediction variance

As it can be observed for the kriging method, the variance at a sampling point is really low and is slightly greater where no sampling was done. However is it not easy to visualize the difference when using the cokriging method since the color scale is not the same. In order to resolves this issue, the variance of prediction difference between the two methods is computed and plotted as shown on figure 24. Now it is easy to observe that the covariable brings little to no information in this particular case. Indeed, the difference in variance of prediction is close to 0 around all the points. This could be explained by the fact that the points distribution is very dense. Therefore, an extra variable brings little information even if this variable is highly correlated.

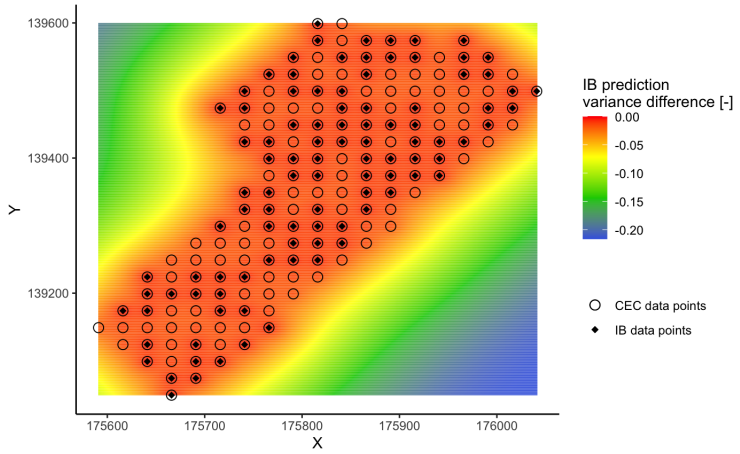


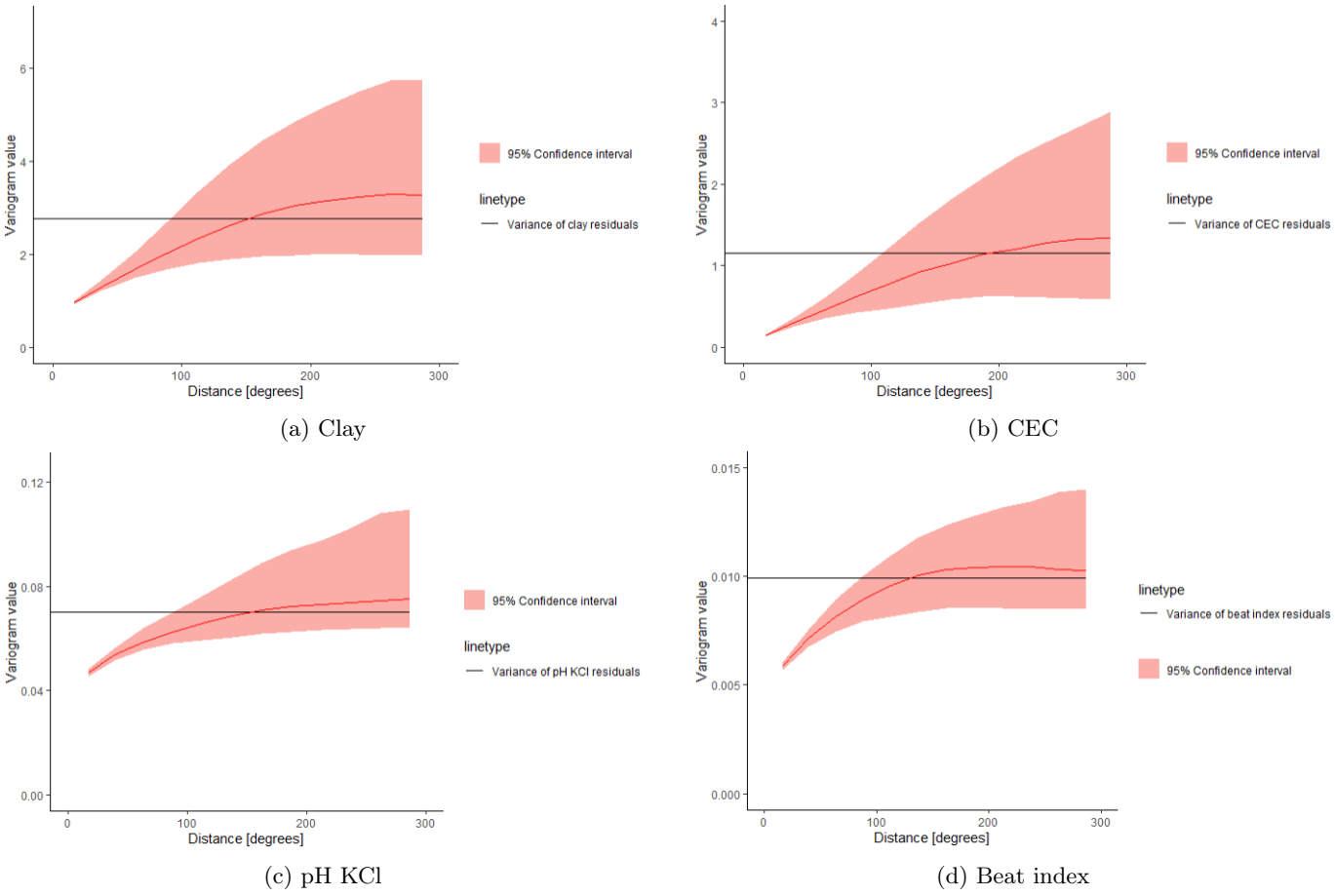
Figure 24: Prediction variance difference between kriging and cokriging ($\widehat{\sigma}^2_{krig} - \widehat{\sigma}^2_{cok}$)

Another observation that can be made is that the variance of prediction of kriging is lower than cokriging. This is surprising, since the opposite was expected. The mean of variance prediction for kriging is 0.00866 and for cokriging this values is 0.0699. There is a factor of 10 between the two means however thoses values are still small. Furthermore, the variance of prediction is homogeneous around the observations for both methods. An explanation for this high variance of prediction for cokriging is that the linear model of co-regionalisation is not completely appropriate. An alternative could be to use a non-linear model of co-regionalisation or the use of robust estimators [6] but this method should be used with caution and can not be applied in every situation.

6 Validation of the variogram with simulation

As an additional topic, it was chosen to validate the models via simulations presented through the mean variogram with a confidence interval (95%). These variograms were obtained via a hundred simulations on a grid size fixed at 5. Note that the size of the grid was increased from 1 to 5 for this simulation, 1 being the value defined for the predictions, allowing a non-negligible saving of time during the execution of the code.

The results of this validation are shown in the figure 25 below.



(a) Clay

(b) CEC

(c) pH KCl

(d) Beat index

Figure 25: Confidence interval (95%) for the variogram of the different variables. The red line represents the mean

The simulations confirm that the use of a variogram model using both a nugget and spherical model is appropriate. There is indeed a spatial dependency, especially for small distances. Using a variogram build only with the nugget model will not highlight the spatial dependency for small distances.
These different confidence intervals allow us to see that for each variable, the variance of the residuals is present in the confidence interval, which allows us to validate the variograms constructed in the section 3. This validation will be useful for the discussion part to validate the hypothesis made.

7 Discussion

While analyzing the results and writing the report, it was noticed that some of the results seemed incorrect. The purpose of this section is to discuss these results and to highlight the inconsistencies and how these could possibly be corrected.

The first major source of error is the fact that the plot was taken as one whole area when it has two areas with different land uses that have a direct impact on the sampled variables. Considering the plot in two zones would have allowed us to work with more accurate residual values for each zone with the result of eliminating more or less outliers resulting in improved predictions

In the first instance, this assumption has an impact on the different steps of the spatial dependency analysis.
First, when the variograms are displayed in relation to the raw values, they give the impression that the variables are not stationary of order 1. However, displaying the variogram for each zone via the raw values would perhaps show that the values tend towards variance with distance and thus that for each zone stationarity of order 1 for the four variables has already been reached. However, as seen in the section 6, the variance is within the confidence interval which allows us to conclude that our variograms are not that bad.
Secondly, despite the improvement of the variograms after removing the trends, these trends are not the best. Indeed, the trend with respect to *IB.samples* is removed via A and CEC. However, *IB.samples* is calculated via CEC which is itself calculated as a function of A. The same thing is observable for the trend of A which is removed with respect to CEC and *vice versa*. Even if this results in better variograms, it still seems imprecise or not correct.
To finish on this part of the spatial dependence, a better analysis could have been to consider that a spatial dependency is observed on the 3D graphs for the X and Y variables, making predictions on the residuals and adding the mean would have allowed a greater precision and a more correct analysis with respect to the theory.

The global quality of the predictions are acceptable and seem plausible. Predictions made with the IDW method are not as good as the predictions made using the kriging or cokriging method, especially for high values of θ, as illustrated with the respective contour plots.
The two disadvantages concerning the IDW method is the presence of artifacts as discussed in section 4.3 and the fact that the quality of predictions decrease rapidly for larger distances. The prediction converge towards the mean of the observations, providing no information concerning the target values at unsampled locations. In other words IDW works well for interpolation but yields poor results for extrapolation.
Despite the good performances of cokriging, this method does not provide more information when comparing with the results with the univariate kriging. Again, this could be explained by the fact that only one zone was considered instead of two, thus introducing more disparencies in the data used.
Finally, when observing all the predictions made it seems that the variables do not respect the stationary of order 1. This does not corresponds to the hypothesis stated before based on the analysis made using the 3D scatter plots done in section 3. This could be easily fixed by using the residuals instead of the direct values. Predictions made using the residuals respective to each zone could even improve the results. However, this seems to be unnecessary. Indeed, when taking into account the variance of the residuals considering only one zone, they are still inside the confidence interval thus validating our results.

8 Conclusion

First, a spatial dependency analysis was made to investigate the relationship within the variables. Should they exist, the trend and the outliers were then taken out and the respective variograms were build. This is followed by the prediction of the variables using various methods namely the Inverse Distance Weighting (IDW), kriging and cokriging. It is shown that the prediction made using the IDW method yields good results for relatively small distances although those results are not as smooth as the results from the kriging or cokriging method. Predictions made with the kriging method surprisingly yield the best results. No extra information is gained using a covariable despite the high correlation with the target variable. This could be explained by the fact that the linear model of co-regionalisation is not appropriate in this particular case.

Finally, the variograms were validated by running 100 simulations. In order to reduce computation time, a grid size of 5 instead of 1 was used. The spatial dependency is confirmed for all four variables by rejecting the use of a variogram build using only a nugget model.

9 References

[1] Ray R. Weil , Nyle C. Brady. The nature and properties of soils. file:///C:/Users/PC/Downloads/The%20nature%20and%20properties%20of%20soils%20(Pearson,%202017).pdf. Accessed: 2022-12-18.

[2] Selim Kapur, Sabit Erşahin. Soil security for ecosystem management. https://link.springer.com/content/pdf/10.1007/978-3-319-00699-4.pdf. Accessed: 2022-12-18.

[3] Zia Ahmed. Geospatial data science in r : Co-kriging. https://zia207.github.io/geospatial-r-github.io/cokriging.html. Accessed: 2022-12-18.

[4] DG Rossiter. Co-kriging with the gstat package of the r environment for statistical computing. http://www.itc.nl/rossiter/teach/R/Rck.pdf, 2007.

[5] Edzer Pebesma. R documentation : Fit a linear model of coregionalization to a multivariable sample variogram. https://search.r-project.org/CRAN/refmans/gstat/html/fit.lmc.html. Accessed: 2022-12-18.

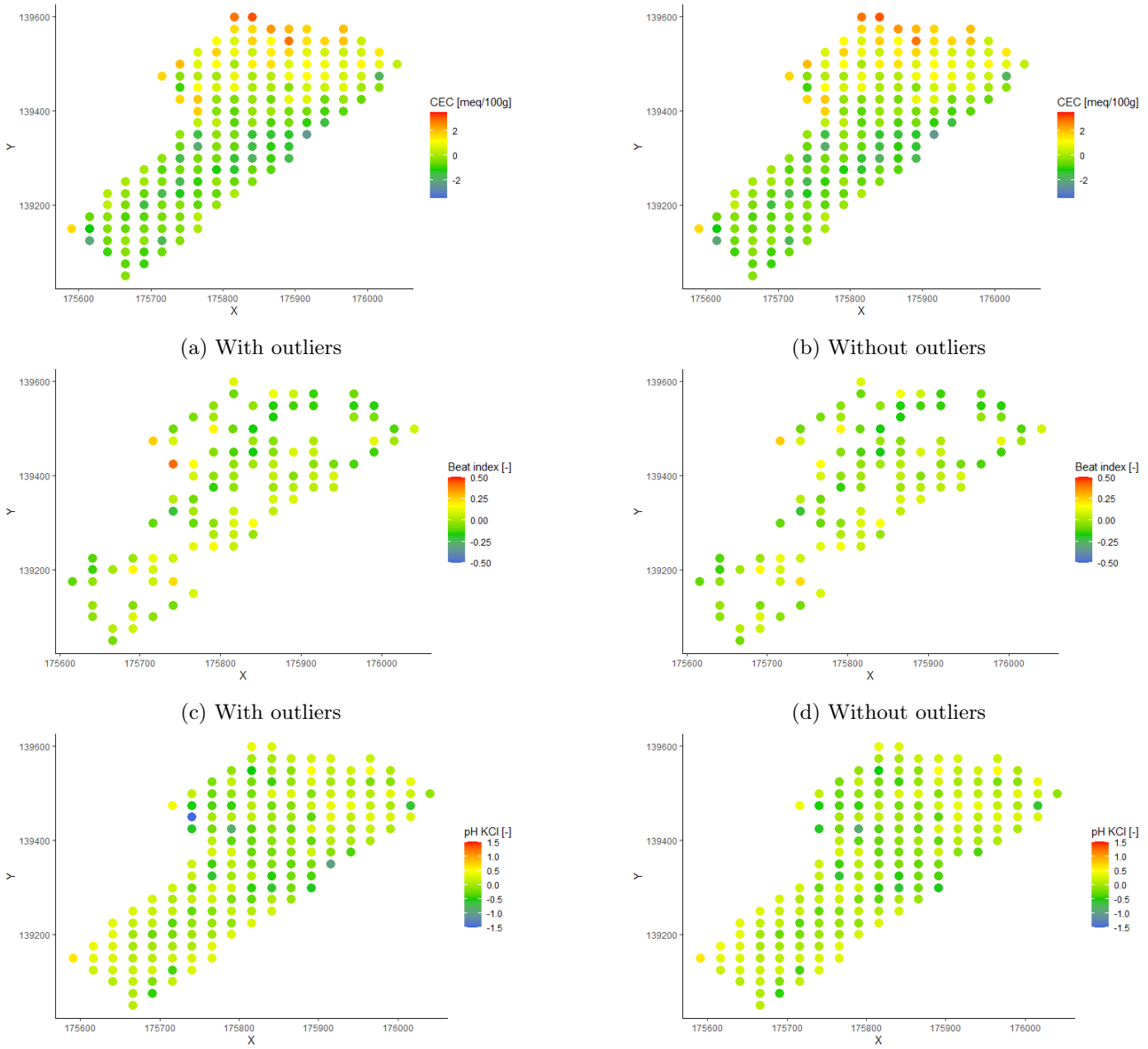
[6] RM Lark. Two robust estimators of the cross-variogram for multivariate geostatistical analysis of soil properties. *European Journal of Soil Science*, 54(1):187–202, 2003.

10 Appendix

10.1 Software used for the analysis

This analysis was done using the R environment via the Rstudio software. For ease of visualization, it is a RMarkDown file that was chosen to make this analysis.

10.2 Residual maps with and without outliers



(a) With outliers

(b) Without outliers

(c) With outliers

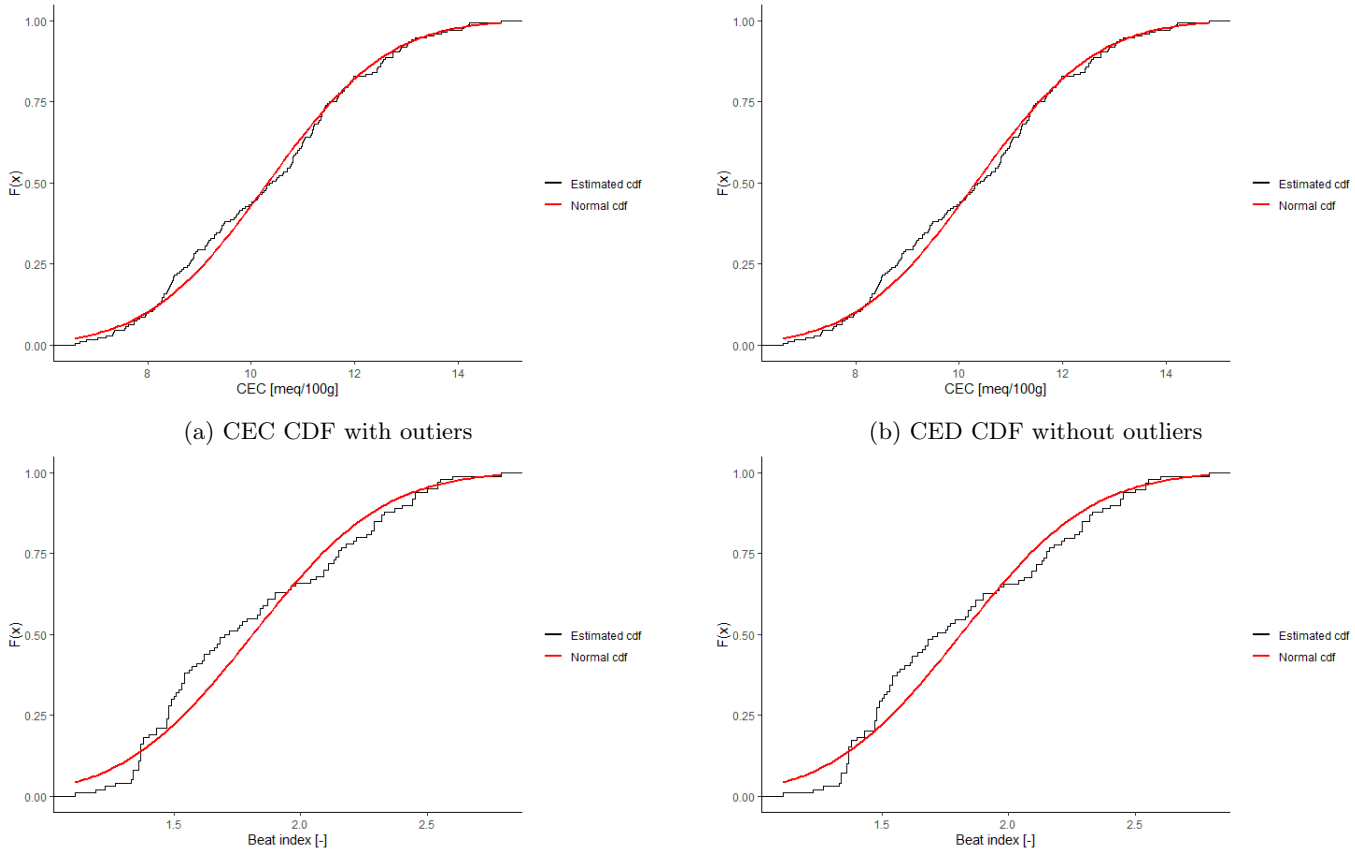
(d) Without outliers

(e) With outliers

(f) Without outliers

Figure 26: Variables residual maps

10.3 CDF plots with and without outliers



(a) CEC CDF with outiers

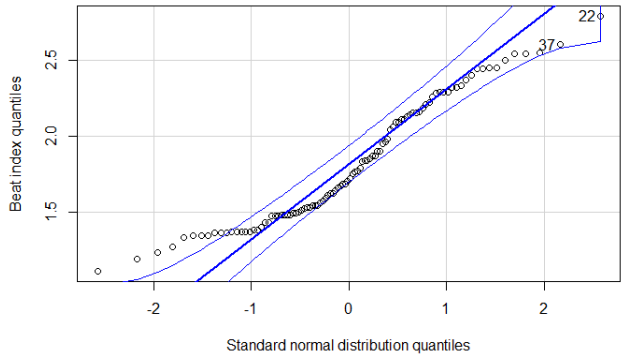
(b) CED CDF without outliers

(c) IB CDF with outliers

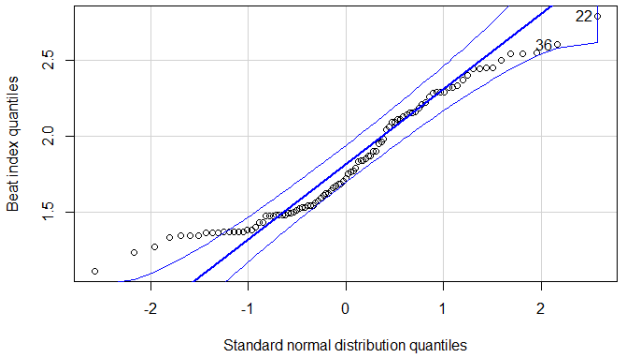
(d) Clay CDF without outliers

Figure 27: CDF plots

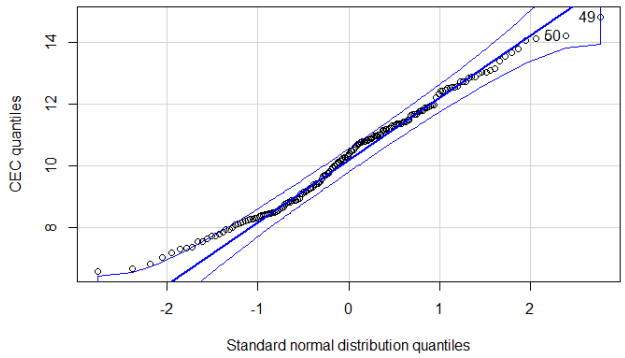
10.4 QQ plots with and without ouliers



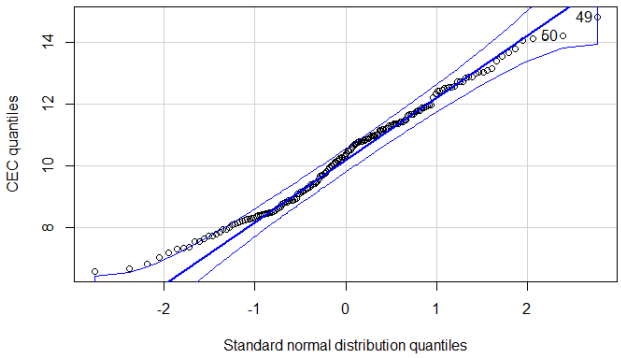
(a) IB QQ plot with outliers



(b) IB QQ plot without outliers



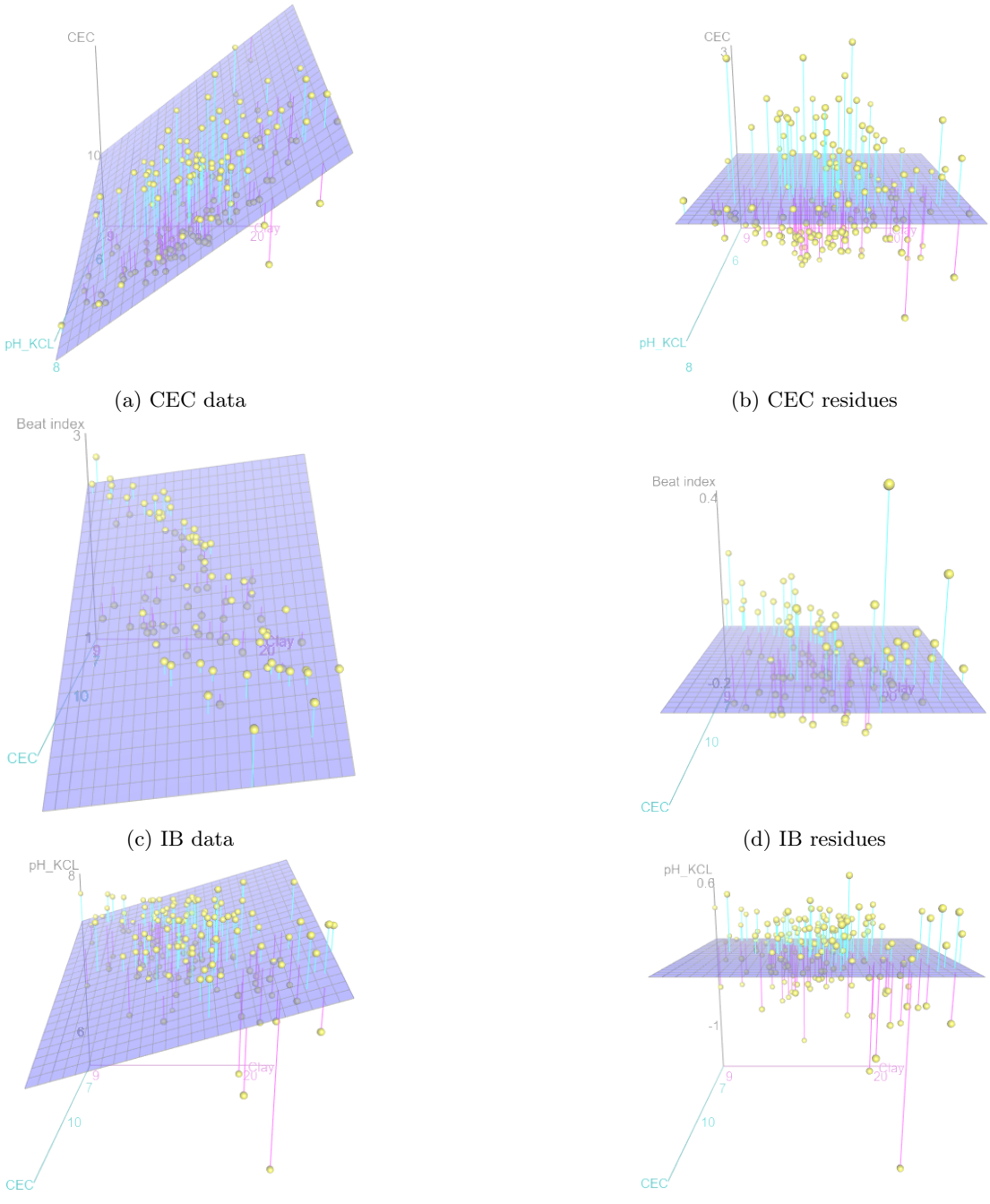
(c) CEC QQ plot with outliers



(d) CEC QQ plot without outliers

Figure 28

10.5 3D point cloud of the three other variables



(a) CEC data

(b) CEC residues

(c) IB data

(d) IB residues

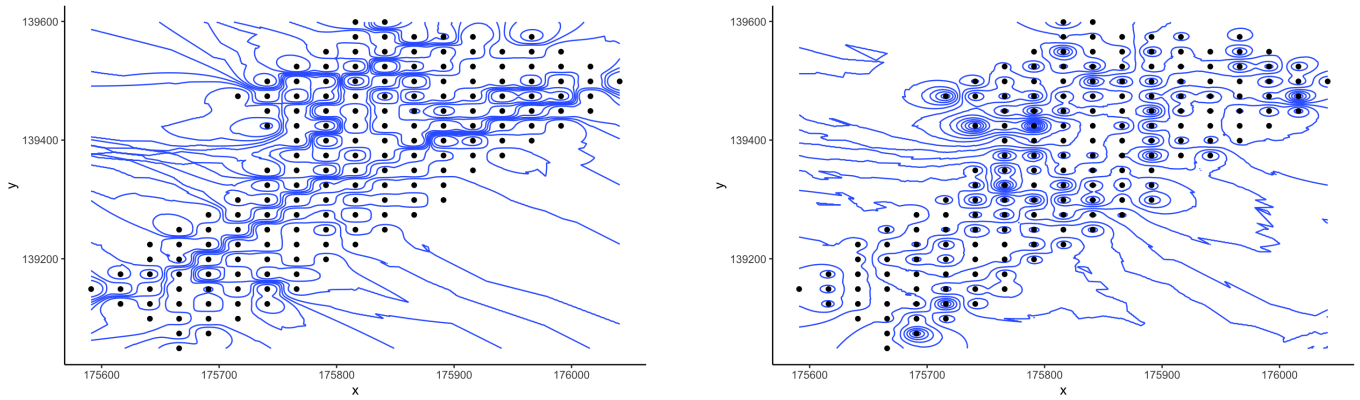
(e) pH KCl data

(f) pH KCl residues

Figure 29: 3D point cloud of each variable

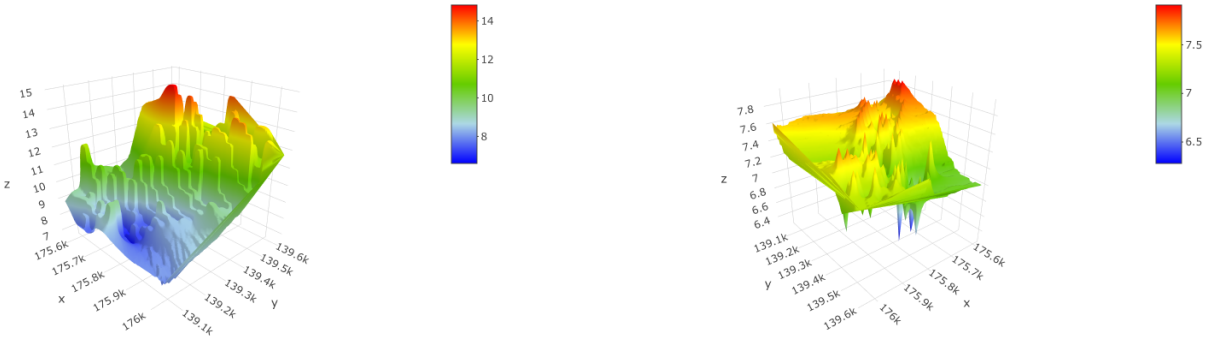
10.6 Inverse Distance Weighting : contour and 3D plots

Figures 30 and 31 show that predictions with a lower value of θ are smoother that a with a higher value.



(a) CEC - $\theta = 4$ (b) pH KCL - $\theta = 2$

Figure 30: Contour plots using IDW

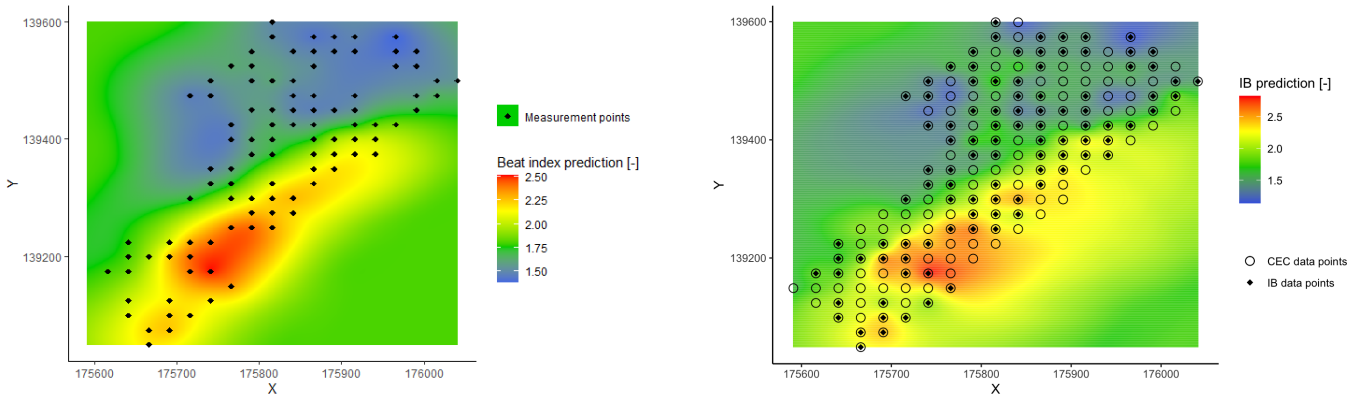


(a) CEC - $\theta = 6$ (b) pH KCL - $\theta = 2$

Figure 31: 3D plots using IDW

NB : It should be noted that for *CEC*, the power is 6 for the simulation because for a power of 4, Rstudio crashes completely. However, the observations should remain the same.

10.7 IB prediction : kriging vs cokriging



(a) Kriging (b) Cokriging

Figure 32: Comparing prediction of IB using two different methods