Alterra WageningenUR & University of Twente

Predicting the effect of reducing the nutrient surplus on the nutrient drainage in the province of Fryslân

Anne Mossink
13-06-2016

ALTERRA WAGENINGEN UR

UNIVERSITEIT TWENTE.

Foreword

This thesis is the outcome of a study about the effect of the nutrient surplus on the nutrient drainage. Ten weeks working at AlterraWageningenUR are dedicated to deliver a sound and proper report.

The project, is in response to a project from Wetterskip Fryslân. They would like to know what the effects are of different measures on the nutrient drainage to waterbodies in the province of Fryslân. They asked Alterra to think of measures and to calculate their impact. I was asked to work out a method that can predict the effect of decreasing the nutrient surplus.

My research had its ups and downs. In the beginning it was hard to get started, because the goal of the study was unclear to me. Furthermore, time was needed to better understand the statistics that was needed to do this research. Later on, researching went well, so I started to do extra tasks and started to optimize my results. However, in the end, I was running out of time. I discovered a mistake in my calculations that I made earlier on and the mapping with GIS took more time than I expected.

Luckily, with the help of my supervisors and some weekend- and late night work in the end, I established to finish the report. I could not have made this report without my supervisors: Peter Schipper: who made sure, that I did not get lost in the complexity of the nutrient cycles and who provided free coffee and Maarten Krol, who could help me with the statistics. Both supervisors had a role in establishing the direction of the report. This was useful, because at one point I did not know towards which goal I was working. Furthermore, I would like to thank Erwin van Boekel and Piet Groenendijk, who helped me with understanding the model STONE. I am also thankful to Harry Massop, for taking the time to explain some features in ArcGIS. In the end I would like to thank Job Oude Vrielink, he read my report and commented on it.

Summery

The water quality in the water bodies of the province of Fryslân is insufficient according to the requirements of the water framework directive (WFD). To improve the water quality in Fryslân, wetterskip Fryslân and AlterraWageningenUR are investigating measures, that can be applied on a local level, like a farm, to decrease the nutrient drainage[1].

This research in specific, investigates the influence of decreasing the nutrient surplus[2] on the nutrient drainage to the waterbodies from the grasslands in the province of Fryslân. Different examinations are executed for the effects of nitrogen and phosphorus, furthermore datasets are divided into data collected at sandy soils and data collected at clay soils. With the results from the examination a tool will be developed. A tool that is able to quickly predict the effect of decreasing the nutrient surplus in a specific area.

First, literature study is done, to understand the nutrient cycle. Using literature, different factors that determine the nutrient drainage are found. The statistical significance of the influence of these factors are analysed, with a t-test and a correlation study. At hand of the t-test, the correlation study and the literature study the factors are combined and filtered from the dataset. With the remaining datasets, multiple regression analyses are done. For each dataset an analysis is done that takes into account nominal and numerical variables and one is done that takes into account only numerical values. Furthermore, for sandy soils the dataset is divided in sub-datasets with a certain characteristic of the dataset, for instance: irrigation is present. This sub-dataset is aimed to find more precise relations between the nutrient drainage and the nutrient surplus for areas that share the same characteristics. In the last step predictions are made for decreasing the nutrient surplus is sandy soils with 5%. The effect of such a reduction on the nutrient drainage is calculated.

An important aspect of all the equations obtained from the multiple regression analysis is the validation and verification. To make predictions with the obtained equations, it is useful to address the uncertainties. Therefore, the equation is validated on: its linearity, its correlation coefficient, the error, the variables and parameters involved, the best fit with the original dataset and the possible indications of overfitting. The most accurate and probable equations that are derived from the multiple regression analyses are used for prediction making.

The regression analyses for nitrogen do show a valid relation between the nitrogen surplus and the nitrogen drainage. Decreasing the nutrient surplus with 5% causes a decrease of the nutrient drainage in range of 0.67-5.26%. The obtained equations can be used to make predictions, however improvements can be made. This research suggests that, taking into consideration bigger datasets, considering more specified data and involving other variables like: average highest groundwater level or the storage of nutrients in the soil, can improve the accuracy of the regression equation and the precision of prediction making.

[1] Nutrient drainage consists of: the leaching of nutrients and the run-off of nutrients to surface water.
[2] Nutrient surplus consists of: nutrient deposit by fertilizers, nutrient absorption by vegetation and atmospheric deposition.

Contents

List of figures and tables

Figures:

Tables:

List of abbreviations

EU: European Union
WFD: Water framework directive (directive 2000/60/EC)
DAW: Deltaplan Argrarisch Waterbeheer / Delta plan agricultural water management.
GIS: Geographic informatics systems
RMSE: Root mean square error
RRSE: Root relative square error

1. Introduction

In the Netherlands water quality has been insufficient in many water bodies for the last decades. The main source of this pollution is assumed to be agricultural. The fertilizer use of the intensive Dutch agricultural sector causes a major nutrient emission, endangering the water quality of the Dutch water systems (Oenema, van Liere, & Schoumans, 2004).

In 2000 the European Union (EU) introduced the water framework directive (WFD), which led to a new integrated approach of water management in Europe. One of the topics that the WFD addressed, was the improvement of the water quality. This was necessary, because the water quality endangered the biodiversity in flora and fauna. An objective of the WFD was to achieve a good water quality in the ground- and surface waterbodies in Europe in 2015, aiming to restore the biodiversity and ecology in the waterbodies (Rijkswaterstaat, Ministerie van Infrastructuur en milieu, sd). The main causes of the poor water quality in the Netherlands in 2000, were: the increased water temperature, the effluent of industries and the intensified agriculture (Planbureau voor de leefomgeving, Centraal bureau voor statistiek and WageningenUR, 2015).

In the last years, water quality has improved, including the lower concentration of nitrogen and phosphorus (Figure 1). However, more than half of the surface waterbodies in the Netherlands do not fulfil the requirements for nutrients specified in the WFD, at the moment (Planbureau voor de leefomgeving, Centraal bureau voor statistiek and WageningenUR, 2015). In addition, according to van Gaalen, et al. (2016) the requirements will not be achieved in 50% of the water bodies, in 2027. As a reaction to this prospect the government released the Deltaplan Argrarisch Waterbeheer (DAW)[3], to fulfil the requirements of WFD in 2027. DAW aims to improve the water quality at a company scale. It provides local tailor-made measures, that fit to the specific characteristics of a farm (LTO Nederland, 2013).

One of those measures is the reduction of the nutrient surplus. The nutrient surplus consists of nutrient deposit by fertilizers, nutrient absorption by vegetation and the atmospheric deposition. In practice, a nutrient surplus decreasing measure can be the reduction of fertilizer use or the change of vegetation to more nutrient demanding plants. A fragment of the nutrient surplus is drained to the water bodies. The Nutrient drainage consists of nutrient leaching and nutrient run-off. The majority of the drainage originates from nutrient leaching (Schipper P. , Applying nutrient measures in practice, 2016).

[3] Deltaplan Agrarisch Waterbeheer is a development policy regarding agricultural water management.

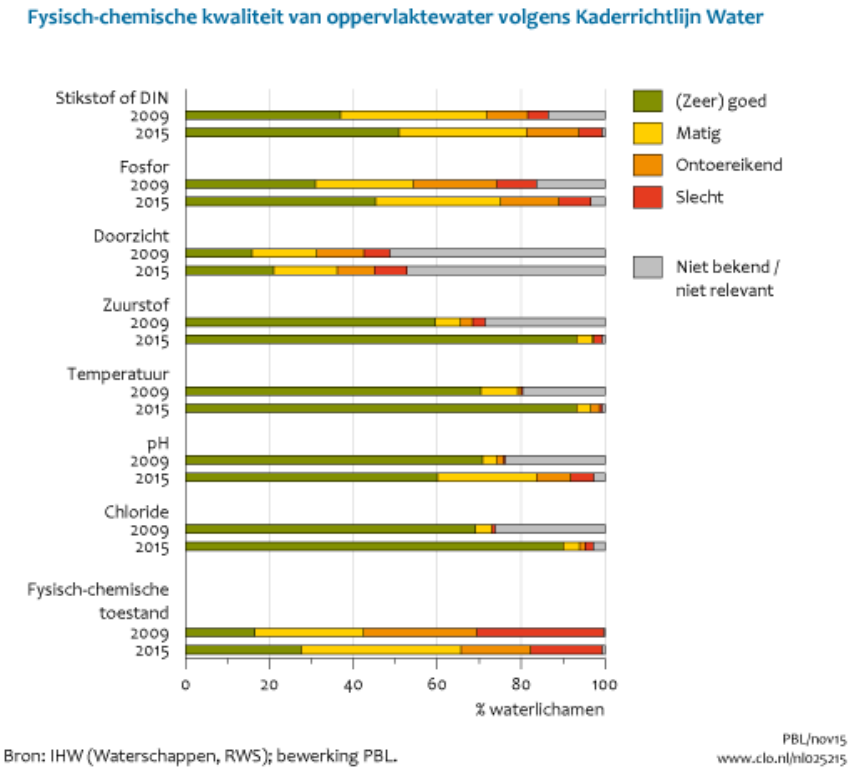


Figure 1: Development of surface water quality factors from 2009 to 2015 (Planbureau voor de leefomgeving, Centraal bureau voor statistiek and WageningenUR, 2015). The last row, 'Fysisch-chemische toestand', represents the development of the water quality in total.

2. Research aim

To improve the insufficient water quality measures should be taken. Current measures, that are taken in the 5th Dutch nutrient action program on fertilizers, do not have the intended effects on the amount of nutrients in water (Groenendijk, et al., 2015). Therefore, alternative measures to achieve the water quality goals are necessary. One of those measures is the reduction of the nutrient surplus. In this research, the effect of decreasing the nutrient surplus on the nutrient drainage in the waterbodies of the province of Fryslân is examined. This measure complies with the strategy developed in the DAW-program (Schipper & van Boekel, 2016) (LTO Nederland, 2013).

Previous research established the relations between factors that determine the nutrient drainage and the nutrient balance in the soil. However, the exact relations among factors remain insecure (Oenema, van Liere, & Schoumans, 2004) and assumptions to describe these relations are often made (Groenendijk P. , 2016). This research is executed in addition to a current research done with the model STONE (Annex 3), executed by Alterra, in the province of Fryslân. STONE is developed to simulate the consequences of fertilizer use on the emission of nutrients to groundwater and surface waters. Nowadays, there is a growing need to use a modelling tool, that predicts the effects of nutrient drainage decreasing measures at a local scale (DAW-program). This tool should have a small running time and should be simple to handle. My research will illustrate how such a tool can be developed, using relationships between nutrient drainage to surface waters and nutrient decreasing measures. The tool will be used to indicate, if measures have a significant effect on the nutrient drainage and the tool will do a rough estimation of this effect on the nutrient drainage. Furthermore, the tool can be included in other models that have a need of a nutrient prediction tool. In those other models simplicity and a fast running time can be required.

2.1. Organisation

The organisation where this research is done is Alterra. It is a research institute connected to Wageningen University. Their research is focussed on the green living environment. They investigate in: policy, management and design effects on the spatial environment. The department that oversees this research is focussed on water management. It investigates the effect of substances and water composition on water quality (Alterra, 2016).

That department is doing a research to possible DAW-measures to improve the nutrient concentration in the waters of Fryslân with 6 polders in specific. This is done in cooperation with Wetterskip Fryslân[4]. The measures are established by the two organisations and are calculated with the STONE model provided by Alterra (Alterra, sd).

2.2. Location

In this research the water quality in the Dutch province Fryslân is examined. This is done by extrapolating from six polders to the control area of Wetterskip Fryslân (Figure 2). The polders are divided into areas with three soil types. The soil of Dongeradiel and Schalsum consist of clay, De Lits and De Linde consists of sand and Fjouwer and Echten consist of peat. In the area, water quality is currently insufficient. Nitrogen is troubling in most of the lakes (Figure 3), while phosphorus is a problem in most rivers (Figure 4). The cause is assumed to be agricultural, blaming cattle in specific. The biggest use of soil in the area is agriculture, with an occupation range of 73-92% in the polders (Schipper & van Boekel, 2016).

Figure 2: Control area of wetterskip Fryslân, containing the 6 examined polders (Schipper & van Boekel, 2016).

[4] The 'waterschappen' or 'Wetterskip' is the governance institution for water systems in the Netherlands. In this case the institution of the province Fryslân.

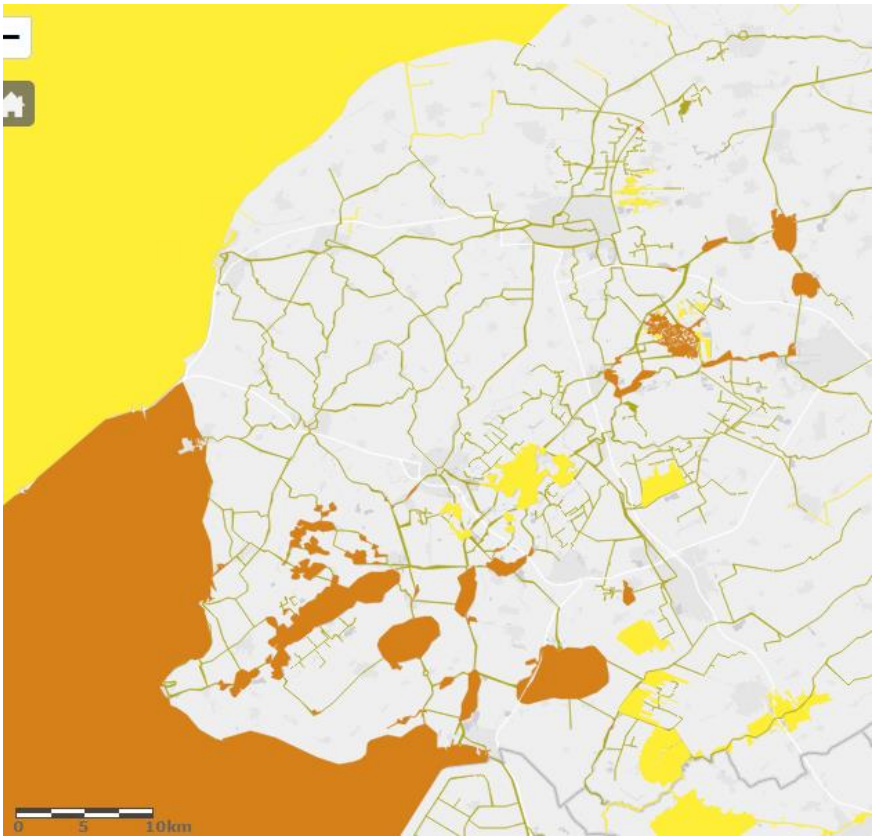


Figure 3: Nitrogen condition in surface waters in Fryslân. Green: sufficient, Yellow: average and orange: insufficient (Planbureau voor leefomgeving, 2015).

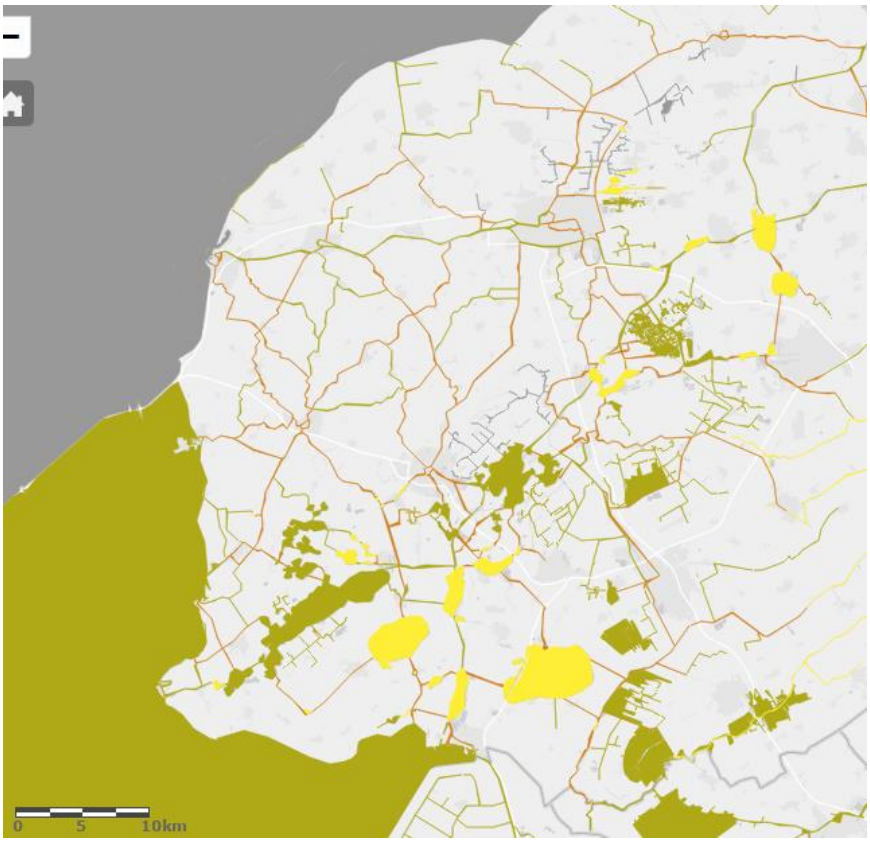


Figure 4: Phosphorus condition in surface waters in Fryslân. Green: sufficient, Yellow: average and orange: insufficient (Planbureau voor leefomgeving, 2015).

3. Research questions

To develop the tool a few stages are followed within the research. First of all, literature study to understand the concepts and processes behind the tool is done. Second, the factors that are involved in the processes that influence nutrient drainage are examined. Third, an analysis of the data and an examination of the relations within the data is executed. Fourth, the results are implemented into practice and visualised. The tool represents the estimated effect of decreasing the nutrient surplus in the waterbodies of the province Fryslân. The main objective therefor is:

What is the estimated effect of decreasing the nutrient surplus on the leaching and surface run-off of nutrients into the waterbodies of the province of Fryslân?

This main objective is separated into sub-questions to answer the different aspects related to the main objective. Aspects are: to understand the specific factors that are involved in the nutrient cycle, the way these factors are related in the nutrient processes, the significance of these factors and the way they contribute to the leaching and run-off in specific. These aspects conclude in a manner to ascertain general relations between the amount of nutrients in water and factors within the nutrient cycle. To investigate this the questions below are constructed:

1. *Which factors influence the nutrient leaching and surface run-off significantly?*
2. *To which extent do the influencing factors relate to each other?*
3. *Which mathematical relations describe the influence of decreasing the nutrient surplus on the nutrient loads leaching and running off to waterbodies?*
4. *How well do these mathematical relations fit to the original set of data?*

The rules obtained from the mathematical part are implemented on areas in Fryslân, aiming to do a forecast on the effect of decreasing the nutrient surplus. The research question for this step is:

5. *What are the estimated effects of decreasing the nutrient surplus, at a specific area with its specific characteristics?*

The research is schematically presented in Annex 1.

4. Limits and boundaries

In total, 8 mathematical relations will be established. Calculations are executed for the soil types sand and clay. Each soil type has a calculation for Nitrogen (N) and phosphorus (P). Furthermore, calculations are made with two different sets of independent variables. In the non-nominal set the nominal values are left out. In the nominal set all values are taken into account. In addition, the nominal factors are examined more closely. For instance, datasets that have the characteristic ‘wet soil’ are examined on their own.

The focus is on finding mathematical relations that represent the relation between the nutrient surplus and the nutrient drainage well. Therefore, attention is paid to the soundness of the formula and on the procedure that is followed.

4.1. Report outline

To understand more about the principles of nutrient leaching and run-off and their factors, literature is studied first (Chapter 5). The literature is used to examine and process the data correctly. The processing of the data itself is described by the methodology (Chapter 6). The methodology explains

the working of the models STONE and Weka, the choices made executing this research and the procedures followed to achieve a mathematical relation. The main procedure is a multiple regression analysis. In chapter 7, the 8 results of the research are presented. In the last chapters: discussion, conclusion and recommendations are stated. In the recommendations an advice is included on how the results can be used to predict the effect of decreasing the nutrient drainage.

5. Hydrological cycle, Nutrient cycle and Eutrophication

To determine which factors influence the nutrient drainage, it is necessary to understand the processes behind it and the effects of nutrients in water. The factors are separated in the nutrient cycle and the hydrological cycle. The nutrient cycle regulates the amount of nitrogen and phosphorus. The hydrological cycle regulates the amount of water. Besides influencing the nutrient drainage, the factors influence each other. This results in a complex system and difficulties in making predictions derived from the cycles. (Oenema, van Liere, & Schoumans, 2004). These cycles are presented in Figure 5 and Figure 6.

5.1. Hydrological cycle

The hydraulic factors influence the water quantity that is discharged in a certain period of time. Precipitation, irrigation and positive seepage influence the supply of water to the system, while evaporation, negative seepage, run-off and leaching drain the system. In between water supply and drainage, the water is stored in the soil and vegetation (Perry, Robbins, & Barnes, 1988) (Figure 5).

The hydraulic factors are dependent on the availability of drain pipes, the weather conditions, the soil type and the soil composition. The soil type and composition affect: the water retention in soil, the saturation of the soil (GT-class) and the time it takes to dispose the rain water into the surface water (Perry, Robbins, & Barnes, 1988). For instance, in a coarse soil structure like sand, water leaches quicker (Perry, Robbins, & Barnes, 1988) and when there are cracks in clay, more nutrient run-off should be considered. (van Boekel, et al., 2012).

The water retention is an important aspect of the hydrological cycle. When water is leached quickly after a shower, more nutrients are disposed into the surface water, because: the soil, the plants and the bacteria have less time to absorb and process the nutrients. Every soil and vegetation type have their own water retention characteristic.

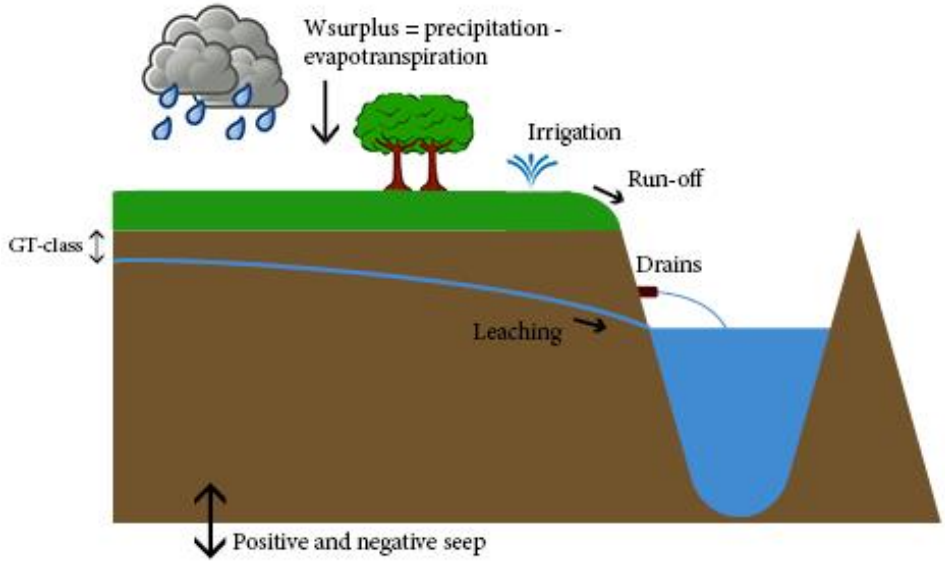


Figure 5: Hydrological cycle (drainage=leaching +run-off).

5.2. Nutrient cycle

In the nutrient cycle different substances play a part. In this research a distinction between nitrogen and phosphorus is made.

The nutrient cycle (Figure 6) consists of soil recovery, denitrification, nutrient seepage and the nutrient surplus. The nutrient surplus consists of: the use of fertilizers, the atmospheric deposition and the absorption of nutrients by crops. Nutrients that are not conserved in the system, are disposed into the waterbodies. The major intake of nutrients into a water body is from run-off during rains, from leaching and from water that originates from other water bodies (The University of Waikato, sd).

The quantity of nutrients in the cycle is dependent on the areas' characteristics, for instance soil type (Schipper & van Boekel, 2016). These characteristics determine the size of the conservation within the soil.

In contrast to nitrogen, phosphorus is not subjected to denitrification or atmospheric deposition (The University of Waikato, sd). Phosphorus enters the system by seepage or fertilizer use. The phosphorus that is not conserved within the soil or absorbed by vegetation is disposed into the waterbodies (The University of Waikato, sd).

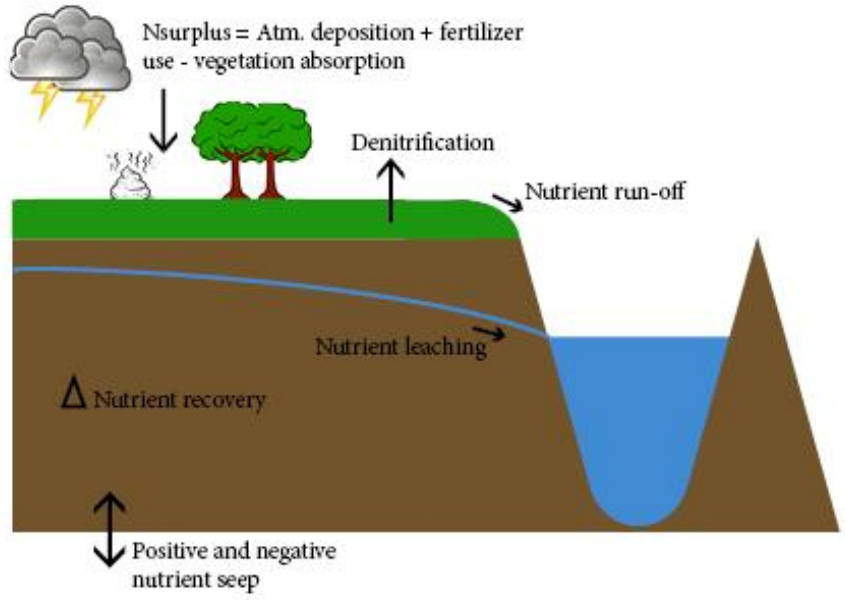


Figure 6: Nutrient cycle (Nutrient drainage = nutrient leaching + nutrient run-off).

5.3. Eutrophication

If the amount of nutrients in water is too high, eutrophication occurs. Eutrophication has a considerable ecological footprint on the water quality, due to the creation of mortal conditions. Eutrophication occurs when the concentration of nutrients in water exceeds a limit, in that case plankton growth is stimulated. When this occurs algae blooms appear. The plankton in its turn, creates low oxygen conditions and prevents sun light from reaching submerged plants. These conditions harm the flora and stimulate decomposers. The decomposition of flora further depletes oxygen, with mortal conditions for animals as a result (BBC, 2014) (Hoekstra, 2013).

6. Methodology

First, the main part of the methodology, the regression analysis is discussed. That chapter is followed by a step-by-step explanation of the methodology. An overview of the methodology is presented in annex 1. The choices made within the methodology result in 6 considerations that are executed in the research. These considerations are evaluated separately.

Considerations:

1. Distinction of data into clay soil and sand soil.
2. Distinction of data into phosphorus drainage and nitrogen drainage.
3. Consideration of nominal and numerical attributes (nominal analysis) or only numerical attributes (non-nominal analysis).
4. Investigation of a dataset which contains areas with no drains, only.
5. Investigation of a dataset which contains areas with one soil physical unit.
6. Investigation of a dataset which contains areas with single soil type, no presence of drains and no presence of drainage.

The considerations 4,5 and 6 are investigations to a dataset with a specific spatial characteristic. These investigations are executed for sand, because clay does not have an appropriate amount of data. Other investigations are done as well, for example to the GT-class, however these investigations did not deliver results accurate enough for prediction making.

In this research the following assumptions are made: Nitrogen and phosphorus do not influence each other and are therefore considered to be two separate investigations. This conclusion is derived from literature and from the low correlation. In this research abbreviations or terms are used to indicate the different processes. These are shown in annex 2.

6.1. Regression analysis

To obtain the relation between the drainage of nutrients and the nutrient surplus a multiple regression analysis is done. The regression analysis examines the interrelation among the complex factors and processes within the nutrient cycle. The input data that will be analysed is provided by the model output from the model STONE (Annex 3). This model is developed by Alterra WageningenUR (Alterra, sd). With the program Weka (Annex 3), developed by (University of Waikato, 2016), the regression analysis is executed. The outcomes are: a regression equation, the equation's correlation coefficient and the equation's confidence (measured by RMSE and RRSE). The regression method described below is founded on the procedure followed for clay. The same method is applied for sand.

The focus of this research is on taking into account nutrient surplus and nutrient drainage, because the relation between those two factors is examined. If Weka does not take these factors into account, the influence of these factors is assumed to be not significant enough or the input data for the surplus would be insufficient.

Regression analysis is a statistical data mining process. Establishing the regression rule, takes into consideration back-ground theory, to determine the factors that will be included in the analysis. It searches for a mathematical relation within the statistical distributions of the input factors (SONDZ, sd) (Sykes, 1993). It takes into account the factors and their direct and indirect influence on the nutrient leaching and run-off.

Multiple regression is a technique, which allows additional factors to enter single regression separately. (Sykes, 1993) Single regression estimates a line that represents the relation between a

dependent (y) and an independent (x) factor. For each point of the regression line an error can be determined. The errors indicate the accuracy of the point to the estimated relation. These estimated relation points are predicted within a certain standard deviation (Figure 7) (McClave, Benson, Sincich, & Knypstra, 2011).

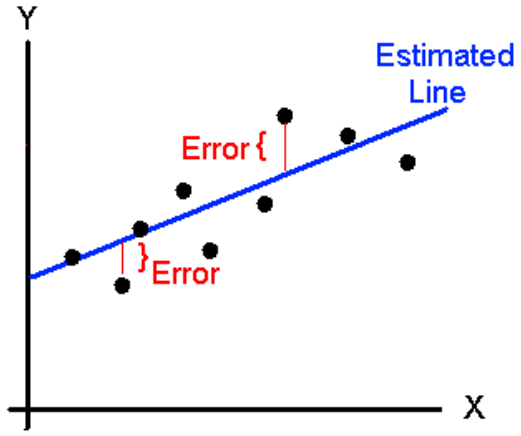


Figure 7: Single regression line.

Combining multiple single linear regressions results into multiple linear regression. The regression form of multiple regression is:

$$y = \beta_0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \varepsilon \qquad (1)$$

with:
y = dependent variable (nutrients drained)
β_0 = a constant (value of y, when all X's are zero)
β = weight factor for input variables.
X = independent variable or input factor
E = the noise or error.

In a multiple regression, each of the variables has its own distribution, which is taken into account in the calculation of y (Figure 8). Factors with a major variance, are allocated less weight in Weka. Their distribution allows an inaccurate determination of the relation. (Sykes, 1993) For each value of y on the estimated line, the distribution is uniform. This means that for all points y, the deviation of a point is the same.

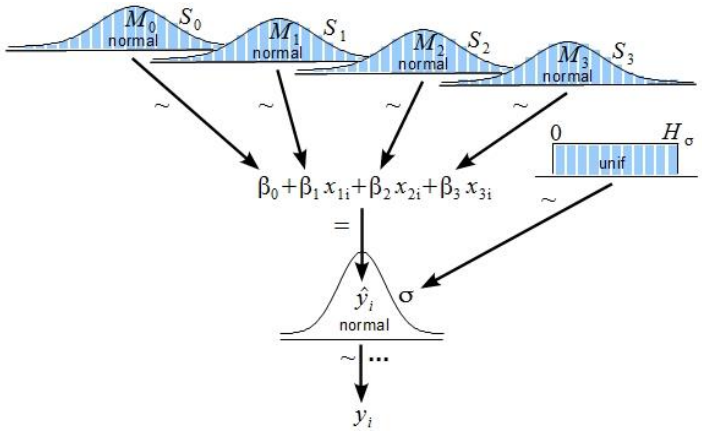


Figure 8: Combination of multiple regressions.

6.2. Data preparation

The first step in this research is the preparation of data. The data derived from STONE (Annex 3) is filtered and prepared for the regression analysis in Weka. The data that is in the end considered in the regression analysis is presented in Table 1.

First of all, areas that fulfil the requirements: grass and Fryslân are collected. Second, the data sets are divided into clay soil and sand soil. From these sets data is again filtered into data that influence the nitrogen drainage and into data that influence the phosphorus drainage.

Other data is combined:

- Nrecovery consists of: mineral and organic recovery.
- Nseep consists of: NH_3 and NO_4 seepage.
- Nsurplus consist of: The absorption of nitrogen, the atmospheric deposition and the fertilizer use.
- Wsurplus consist of: precipitation minus evaporation.
- GT-class is divided in the classes: wet, middle and dry (Annex 2).
- The soil physical unit is divided in clay types: A, B, C, D and E (Annex 2).

The same is done for Pseep, Psurplus and Precovery.

The data preparation above results into the initially considered factors, that are used to calculate the nutrient drainage: nutrient surplus, soil physical unit, GT-class, presence of drains, presence of irrigation, water surplus, water drainage, seepage, nutrient seepage, denitrification and nutrient recovery or intake by soil. The size of the data set is n=77 for clay and n=175 for sand.

Table 1: The variables that are considered in the regression analysis.

Name	Variables or Abbreviation	Definition
Nutrient drainage	Ndrain or Pdrain	The total amount of nutrients disposed into the surface water by nutrient leaching and run-off.
Nutrient surplus	Nsurplus or Psurplus	The surplus of nutrients in an area.
Soil physical unit	Soil physical unit	The composition of the soil.
Gt-class	Gt-class	The fluctuations of the groundwater level.
Presence of drains	Drains	The presence of tile drains.
Presence of irrigation	Irrigation	The application of irrigation.
Water surplus	Wsurplus	Amount of rainwater left after evapotranspiration.
Water drainage	Wdrainage	The total amount of water disposed into the surface water.
Seepage	Wseep	The seep of water through soil in or out the area.
Nutrient seep	Nseep or Pseep	The amount of nutrients transported by seep.
Denitrification	Ndenitrification	Denitrification of nitrogen by bacteria.
Nutrient recovery	Nrecovery or Precovery	Storage change of nutrients in soil.

6.3. Data examination

With the use of theoretical knowledge and statistical tests, choices are made about which data is used.

6.3.1. T-test

The students' t-test is used to decide which input factors are the most reliable to use (McClave, Benson, Sincich, & Knypstra, 2011). The deviations calculated with the t-test are compared to the means. Data with a lower deviation compared to the mean are more easy to predict, therefor preference is given to data with a lower t_c-value. The t_c-value is calculated with equation 2:

$$t_c = \frac{t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right)}{\bar{x}} \quad (2)$$

With:
$\bar{x}$ = mean
n = number of data
$t_{\alpha/2}$ = coefficient dependent on n-1 degrees of freedom and the confidence interval.
s = standard deviation: $s=\sqrt{var}$ (3) with var=variance
variance = $var = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ (4)
x_i = single value

The t_c-values are calculated for a significance of 95% and presented in annex 4. Remarkable is the amount of data needed to achieve reliable distribution. Ndrain should have a data size above n=500. Concerning the independent variables, the amount of data is insufficient as well.

6.3.2. Visualisation

To achieve an initial indication of the data, the data is visualised in plots. From the plots some conclusions are drawn:

-The graphs Ndrain-Nsurplus and Pdrain-Psurplus do not show a clear relation for sand, however for clay it is possible to conclude a linear relationship (annex 5, Figure 16).
-Factors concerning phosphorus do not relate to nitrogen factors.
-The drain plots show that the presence of drains in the area influences the amount of nutrients leached. Drains lower the ground water level, creating aerobic grounds and limiting denitrification (The University of Waikato, sd) and the vegetation's absorption (Perry, Robbins, & Barnes, 1988) (annex 5, Figure 17).
-Wsurplus is not related to the drainage in clay (annex 5, Figure 18). However, according to the theory there is a relation between rain and leaching (Groenendijk, et al., 2015) A possible explanation is that precipitation and evaporation inside the area, do not differ much from each other.
-Similarities are found between Wseep and Wdrainage (annex 5, Figure 19).

6.3.3. Correlation

In addition, the correlation is examined. If variables influence each other in a significant extent, the relation will become untrustworthy, because indirectly factors would be taken into account twice. If the correlation coefficient is close to zero, there will be no relation. If the correlation coefficient is near r=1 or r=-1, the correlation will be perfect and if r > 0.7, the correlation is assumed to be strong (Doorn & Rhebergen, 2006). Therefore, when two factors have a high correlation outside of -0,5 <= r <= 0,5, one of them should be discarded from the regression analysis. (SONDZ, sd). The correlation coefficient or Pearson's r is calculated with equation 5 (McClave, Benson, Sincich, & Knypstra, 2011):

$r = \frac{SSxy}{\sqrt{SSxx \times SSyy}}$ (5) consisting of:

$SSxy = \sum XiYi - \frac{(\sum Xi)(\sum Yi)}{n}$ (6), $SSxx = \sum Xi^2 - \frac{(\sum Xi)^2}{n}$ (7) and $SSyy = \sum Yi^2 - \frac{(\sum Yi)^2}{n}$ (8)

With: Xi = single independent value, Yi = single dependent value and n = number of data.

The validity of the correlation is calculated with equation 9 and tested to the null hypothesis (equation 10) and the alternative hypothesis (equation 11) (Wilks, 2006). The correlation results are presented in annex 6.

$t = r\sqrt{\frac{n-2}{1-r^2}}$ (9), H_0: r=0 (10), H_a: r≠0 (11)

The correlations are compared with the theory. In clay, the results show that denitrification has a positive relation with the nutrient drainage. This means that, if denitrification increases, nutrient discharge increases. However, in theory, when denitrification increases the amount of nutrients in the cycle becomes less and therefor the nutrient drainages should reduce. In addition, equation 9 shows that the correlation is not sufficient. Therefor denitrification is not taken into account in the regression analyses.

Furthermore, differences are found in the relation between Wsurplus and Ndrain. In clay there is a positive relation in sand a negative one. A positive relation could be caused by the bigger transport of nutrients in run-off water. A negative relation could be caused by a bigger outward seepage or a better saturation of the soil.

High correlations are found between Wdrainage, Nseep or Pseep and Wseep. These correlations are found for clay and sand. Another high correlation is found between Psurplus and Precovery. Insufficient correlations are also found. Nitrogen drainage has an insufficient correlation with denitrification. In the case of phosphorus, the correlation between Psurplus and any other factor is disputable. Furthermore, for sand soils is the relation between Psurplus and Pdrain invalid according to the correlation test.

6.3.4. Selection of factors

With the knowledge from the data examination above, choices are made. First of all, due to the high correlations among Wseep, Wdrainage and Nseep or Pseep, Wdrainage and Nseep or Pseep are neglected. Besides, trial and error in Weka showed that the regression for Wseep is the most reliable and showed that Wseep did not exclude nutrient drainage and nutrient surplus at the same time. Furthermore, denitrification is not taken into account, because its correlation is invalid and it makes overfitting more likely.

Despite the disputable correlation factors in phosphor, the regression in phosphorus is calculated. Also Precovery and Psurplus are both taken into account despite the high correlation, because if one of the two is left out, regression analysis results would be inaccurate. Besides, both values are dependent on other aspects. Precovery depends on the minerals in the soil and Psurplus depends on vegetation intake.

In conclusion the following 9 attributes are considered for nitrogen: Ndrain, Nrecovery, Nsurplus, Wsurplus, Wseep, soil physical unit, GT-class, drains and irrigation (Annex 2). The following 9 attributes are considered for phosphorus: Pdrain, Precovery, Psurplus, Wsurplus, Wseep, soil physical unit, GT-class, drains and irrigation (Annex 2).

6.4. Executed tests.

With the achieved knowledge from the data examination tests and by executing trial and error tests, different tests are chosen to be executed. Earlier in the research process, the tests were separated in sand and clay and in phosphorus and nitrogen. Furthermore, a test without nominal values is done, because it is more difficult to allocate weight to nominal values and nominal values incline overfitting in regression analysis. In addition, tests are done that examine datasets with certain characteristics. A test is executed that examines a dataset that is characterised by no presence of drains. A test is executed that is characterised by the same influence of the soil. And last, a test is executed that does consider a dataset with: the same soil, no drains and no irrigation. Tests for the GT-class were executed as well, but did not deliver accurate results.

Weka is used to establish the regression equation. The equation is plotted in a Ndrain-Nsurplus graph with the trend line of the equation and the prediction interval of the equation. It is possible to use this graph as an indicator for the height of the drainage at a certain surplus. It also shows what values the predicted values can attain. The trend line indicates the average drainage for a certain surplus. Furthermore, the prediction interval is calculated, with a confidence of 95% (equation 12 and 13) (McClave, Benson, Sincich, & Knypstra, 2011)

Prediction interval (PI):

$$PI = \bar{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_{xx}}} \quad (12) \qquad \text{and} \qquad S = \sqrt{\frac{\sum (y - \bar{y})^2}{n-(k+1)}} \quad (13)$$

With:
$\bar{x}$ = mean independent value (Nsurplus)
$\bar{y}$ = predicted value (Ndrain)
Y = STONE value (Ndrain)
$t_{\alpha/2}$ = coefficient dependent on n-(k+1) degrees of freedom and the confidence interval.
k = number of independent factors.
n = amount of data
x_i = single value (Nsurplus)
SS_{xx} = equation 7.

Predictions about the effect of decreasing the nutrient surplus are made with the assumption that all factors have constant values. The surplus of a single dataset is reduced, while the other variables in the equation stay the same. Applying this method to the nominal equation results in a prediction considering specific characteristics of the dataset like soil physical unit=H and no drains presence. The non-nominal equation does not take into account these variables. Therefore, the non-nominal equation is applicable in areas that have unknown specific characteristics.

6.4.1. Presence of drains

Annex 7 shows that the mean of nutrient drainage increases for sand, if there are drains presence in the area. Besides, annex 5, Figure 17 show visually that a difference is expected between no drains present and drains present. On the other hand, annex 7 also implies that nutrient leaching decreases when drains are presence in clay. This contradiction is possible, due to the many aspects that are influenced by the drains. If drains are applied in an area, the water balance in the soil would change. More nutrients will be leached to groundwater instead of the water bodies. And due to the change of water pressure within the soil the amount of seep changes. On the contrary, it is believed that drains lower the amount of nutrients leached, due to increasing mineralization in an area with drains (Scholefield, et al., 1993).

6.4.2. Soil physical unit

The soil physical unit determines the nutrient leaching to water bodies. The soil structure determines the water drainage through the soil and the nutrient intake by water (Perry, Robbins, & Barnes, 1988). Annex 7, Table 27 shows a difference between the soil physical units. However, the regression equations do not value the soil physical units as different influences in most cases (Chapter 7). Due to the indication that soil physical unit can influence the drainage, a research to a dataset that contain soil physical units: F, H and I is done. In addition, an extra analysis is executed to the dataset of soil physical units: F, H and I. The presence of drains and the presence of irrigation are left out as well.

6.5. Overfitting

A common phenomenon in regression is overfitting. It occurs when regression is fitted to well to the data. The regression line predicts existing data precisely, but new data does not fit with the regression equations. Often there are too many variables relatively to the number of observations (Frost, 2015). Overfitting can be solved by testing the regression with a new similar data set (Hawkins, 2003) or with a regularization rule (Ng, sd).

In Weka 10 folds cross-validation uses a separate data set to validate to original dataset (Annex 3) (Witten, 2013). However, in the statistics community, cross-validation is considered as a poor validation (Hawkins, 2003). However, Hawkins (2003) states that it is better to use all of the compounds of the dataset for cross-validation than to split the model and use a separate part from the analysis as a validation set.

Two rule of thumbs are mentioned to determine the right amount of data and to avert overfitting. The data set should be between 10 × k and 15 × k (14) (Frost, 2015) or n=100 + k (15), with k as the number of independent variables (Anglim, 2011). These rules of thumb are applied on the executed tests in annex 9.

6.6. Verification tests

The verification of the analysis is examined by testing some of the properties of the distributions. The size of the data set indicates the accuracy (annex 9). The students' t-test is conducted to evaluate the accuracy of the individual input data (annex 4). Furthermore, three regression indicators indicate the reliability of the equation. Namely: the correlation coefficient for the whole regression equation (equation 5), the root mean squared error (RMSE) and the root relative squared error (RRSE). R represents the strength of the relation. RMSE and RRSE are used to examine the deviation in regard to the estimated line (Witten, 2013). The equations are given below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pi - xi)^2} \quad (16)$$

With Pi is the predicted value and xi is the observed value for a certain instance i.

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(Pi-xi)^2}{\sum_{i=1}^{n}(xi-\bar{x}i)^2}} \quad (17)$$

With $\bar{x}i$ as the mean of the observed values.

The regression equation itself is validated, by comparing the equation with the actual data from STONE and among other regressions results. The regression equations are also tested on their linearity, with residual plots (Shalabh, sd) (Statwing, sd). In a residual plot the variable Nsurplus is compared to the error (difference between the predicted value and the real value). If there is no indication of a relation between Nsurplus and the error, the regression equation is assumed to be linear.

7. Results

The regression that is executed takes into account the following variables: Ndrain or Pdrain, Nrecovery or Precovery, Nsurplus or Psurplus, Wsurplus, Wseep, soil physical unit, GT-class, drains and irrigation. The hydraulic variables are presented in mm/yr. and the substances are measured in kg/ha/yr. (Annex 2). In all cases, the most accurate and most probable regression equation is chosen to use for predictions. The regression should be in accordance with the theory and leaves out as many independent variables as possible. The results for all tests are discussed below.

The results are presented in annex 10. They consist of the equation with its indicators: R, RMSE and RRSE. Furthermore, the obtained regression equations are plotted. The plotted equations are accompanied by the trend line that represents the average prediction and by lines that represent a prediction deviation of 95%. As a validation the predicted drainage is plotted against the real drainage derived from STONE. At last, a prediction line is established that represents nutrient reduction at a single farm or dataset. This single prediction is done with data derived from STONE and the equation from the regression analysis. An initial point in the dataset is chosen. From this point a line is drawn, that shows the effect of changing the nutrient surplus, while other factors are assumed to be constant. To express the results more clearly the plots only visualise the core of the graph. Outside the core interval a few values are present (Annex 12). In addition, to prove the linearity of each equation residual plots are presented in annex 11.

7.1. Nitrogen

Nitrogen drainage is influenced significantly by the nitrogen surplus. This is shown in annex 5, Figure 16 and in the slopes of the regression equations' averages (Annex 10, top right graphs). Considering data from STONE, the fluctuations in the data is significant (Annex 5, Figure 16). Therefore, the regression equation will show the same fluctuations (Annex 10, top right graphs).

7.1.1. Clay

The regression analysis executed for clay resulted in the following regression equation:

$$Ndrain = 0.115Nsurplus + 0.051Wseep + 6.932 \text{ if 'soil physical unit=D'} - 5.754 \text{ if 'Gt-class=Middle or Wet'} + 3.956. \quad (18)$$

N = 77, R = 0.82, RMSE = 5.03, RRSE = 57.34%.

The nominal regression equation set for nitrogen has a clear relation between nitrogen surplus and nitrogen drainage (Annex 10.1, top right). The regression does not involve a large number of independent values in the regression equation in regard to the small number of data used. Furthermore, many nominal values are valuated the same. For instance, soil type A, B, C and E are considered as one and the presence of drains does not matter. The correlation of the equation indicates that the relation between nitrogen surplus and nitrogen drainage is significant enough. However, the error and the relative error have high values, this indicates a large prediction interval and a high deviation of predicted values in regard to the average.

The Ndrain-Ndrain predicted graph (Annex 10.1, bottom left), shows that the regression equation fits to the data of STONE. This is confirmed by the slope of the trend line of the graph and the intersection with the x-axis. In an ideal situation the parameter of the slope would be one and the intersection with the x-axis would be zero.

The bottom right graph of annex 10.1 shows the prediction line for two points with the same characteristics. Both points represent an area in the province of Fryslân. The values of these initial points are presented in Table 2. These points are plotted against the predicted values of the

regression equation. The prediction lines are parallel to each other, due to the corresponding parameters. The single prediction lines lay within the course of the graph, but do deviate a bit from the average trend line. In addition, the prediction interval is calculated for data point 1 (Figure 9). The figure shows that the deviation of the prediction at a 95% interval is widespread.

Table 2: Used point characteristics for the nominal equation of clay.

	Point 1	Point 2
Nsurplus	141.79 kg/ha/yr.	139.90 kg/ha/yr.
Wseep	1.13 mm/yr.	-249.14 mm/yr.
Soil physical unit	D	D
Drains	No	No
Irrigation	No	No
Gt-class	Wet	Wet

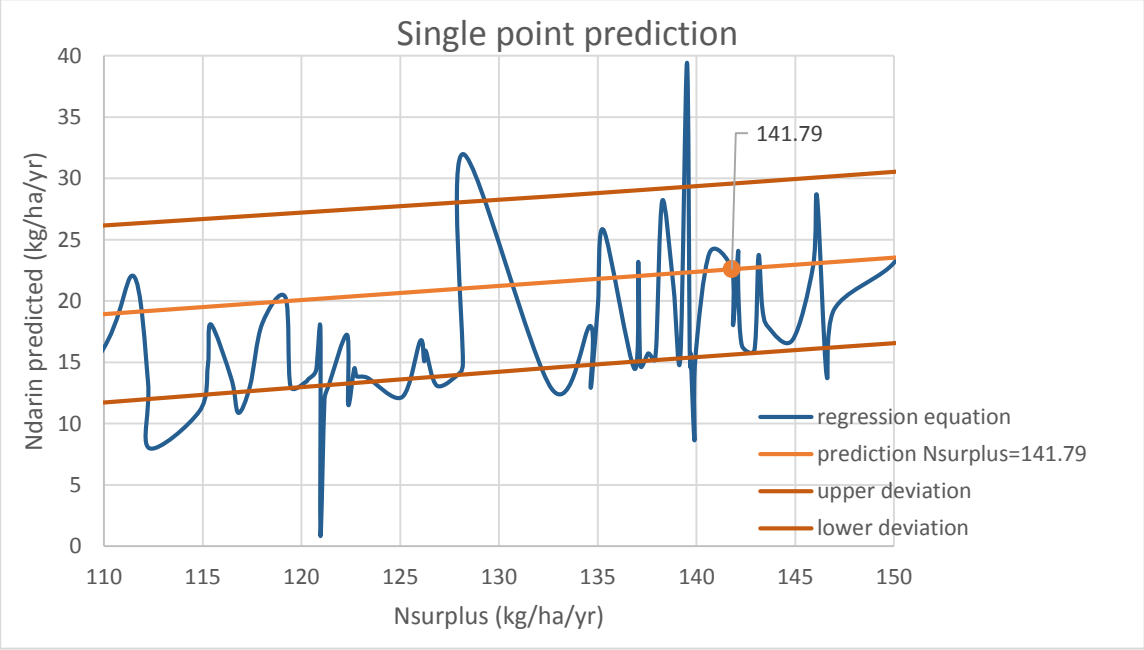


Figure 9: Single point prediction and its prediction interval (95%) in clay with starting point Nsurplus=141.79.

Considering numerical values only resulted in the non-nominal regression equation:

$$Ndrain = 0.115Nsurplus + 0.057Wseep + 0.093Wsurplus - 0.160Nrecovery - 31.187. \quad (19)$$

N = 77, R = 0.84, RMSE = 3.09, RRSE = 53.98%

Compared to the nominal equation, the non-nominal equation involves water surplus and nitrogen recovery as well. Both the nitrogen surplus and the seepage have similar parameters with the nominal equation. The non-nominal equation has slightly better R, RMSE and RRSE values, than the nominal equation. This means that the equation fits better with the data derived from STONE. This is also visible, when comparing the Ndrain-Ndrain predicted graphs (Annex 10.1 and annex 10.2, bottom left) The points' characteristics of the single prediction lines (Annex 10.2, bottom right) are presented in Table 3. Due to the equal parameter Nsurplus, the non-nominal single point predictions and the nominal single point predictions are parallel to each other.

Table 3: Used point characteristics for the non-nominal equation of clay.

	Point 1	Point 2
Nsurplus	141.79 kg/ha/yr.	134.81 kg/ha/yr.
Wseep	1.13 mm/yr.	-97.13 mm/yr.
Wsurplus	368.52 mm/yr.	369.90 mm/yr.
Nrecovery	5.15 kg/ha/yr.	-9.85 kg/ha/yr.

7.1.2. Sand

The regression analysis executed for sand resulted in the following regression equation:

Ndrain = 0.049Nsurplus + 0.026Wseep + 3.603 if 'soil physical unit=F,H or I' + 0.361 if 'soil physical unit=G' - 1.405 if 'GT-class= Wet' + 2.490 if 'Drains=Yes' + 2.571 if 'irrigation=yes' + 10.047. (20)

N = 175, R = 0.75, RMSE = 4.28, RRSE = 66.65%

The nominal regression equation for sand involves many variables, especially nominal variables. All the nominal values do contribute to the value of nutrient drainage, except the presence of drains. The R value of 0.75 and the RMSE of 4.28 show that the regression equation does fit well, with the data from STONE. The Ndrain-Ndrain predicted graph (Annex 10.3, bottom left), confirms this. On the contrary, the RRSE indicates that predicted data can have a significant deviation with the average trend line of the STONE data.

The high fluctuations of the prediction equation in regard to the average trend line is also shown in (Annex 10.3, top right graph). These fluctuations are partly caused by the differences in the areas' characteristics. For example: areas with irrigation, often have a higher drainage than other areas and therefore show an overestimation at the top of the graph. Furthermore, the graph and the average trend line show, that the nitrogen drainage does not depend on the nitrogen surplus in a high extent.

The bottom right graph of annex 10.3 shows the prediction line for two points with the same characteristics. The initial points represent areas, with values from Table 4. The slope of the single point prediction lines shows a minor change when decreasing the nitrogen surplus. Additionally, the graph and the corresponding values show that the prediction lines are parallel to each other and about parallel to the regression equation line, within the given interval. However, this parallel trend between the single prediction line and the regression equation is only seen within the interval of 90-110. Outside this interval, extreme values do not show a clear course of the graph. In addition, for point 1 the prediction interval is calculated (Figure 10). The figure shows that the deviation of the prediction at a 95% interval is widespread.

Table 4: Used point characteristics for the nominal equation of sand.

	Point 1	Point 2
Nsurplus	101.29 kg/ha/yr.	104.23 kg/ha/yr.
Wseep	25.36 mm/yr.	-193.15 mm/yr.
Drains	No	No
Soil physical unit	H	H
Irrigation	No	No
Gt-class	Middle	Middle

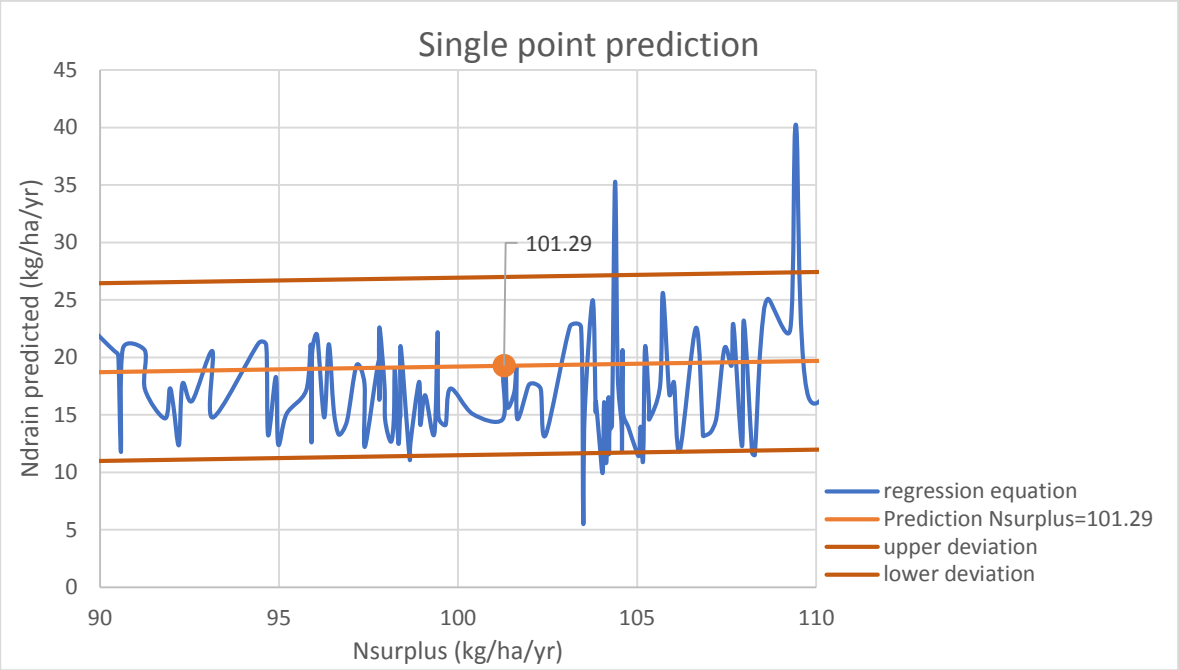


Figure 10: Single point prediction and its prediction interval (95%) in sand with starting point Nsurplus=101.29.

Taking into account numerical values only results in non-nominal equation 21:

Ndrain = 0.130Nsurplus + 0.023Wseep - 0.068Wsurplus - 0.050Nrecovery + 28.146. (21)

N = 175, R = 0.75, RMSE = 4.25, RRSE = 65.97%.

The non-nominal equation is dependent on less factors than the nominal equation and its parameters do not comply with the nominal equation, except for seepage. The non-nominal equation has a more favourable R by 0.01. Also the RMSE and RRSE are slightly better. The R, RMSE, RRSE and the Ndrain-Ndrain prediction graph (Annex 10.4, bottom left) indicate a good fit of the equation to the STONE data.

The single predictions with the non-nominal equation are estimated with the points from Table 5 in annex 10.4, bottom right. Due to different parameters of the nominal and non-nominal equations, the slopes of the single predictions differ. The slope for the non-nominal equation is steeper.

Table 5: Used point characteristics for the non-nominal equation of sand.

	Point 1	Point 2
Nsurplus	107.44 kg/ha/yr.	104.23 kg/ha/yr.
Wseep	124.64 mm/yr.	-193.15 mm/yr.
Wsurplus	327.54 mm/yr.	330.09 mm/yr.
Nrecovery	29.75 kg/ha/yr.	28.34 kg/ha/yr.

7.1.3. Extra tests

To examine the influence of the single characteristics within an area some tests are executed. Most of the regression analyses on filtered datasets did not deliver accurate results. However, for sand results were gained for three tests:

-A dataset with data that is characterised by: no drains.

-A dataset with data that is characterised by: soil physical units: F, H and I.

-A dataset with data that is characterised by: soil physical units: F, H and I, no drains present and no irrigation present.

The parameters of the regression equations derived from the extra tests are presented in annex 8. As is shown in the annex, the regression indicators do not differ much from each other. Also parameters show similarities. Seepage is similar for all equations. Nutrient surplus and the constant β are similar between the soil physical unit F, H and I test and the non-nominal test.

7.2. Phosphorus

Phosphorus drainage is influenced differently in clay and sand. Clay shows a positive trend when decreasing the phosphorus surplus, while sand gives a negative trend. Considering the STONE data, significant fluctuations between data points are shown (Annex 5, Figure 16).

7.2.1. Clay

The following regression equation is obtained for phosphorus in clay:

$$Pdrain = 0.463Psurplus + 0.007Wseep – 0.004Wsurplus - 0.514Precovery + 0.230 \text{ if 'soil physical unit=A or B'} + 0.745 \text{ if 'soil physical unit=C'} + 0.284 \text{ if 'soil physical unit=D'} - 0.297 \text{ if 'GT-class=wet'} + 0.376 \text{ if 'Drains=no'} + 1.701. \quad (22)$$

N = 77, R = 0.87, RMSE = 0.40, RRSE = 48.74%

The regression equation of phosphorus in clay does contain many variables. The prediction of the data is very accurate, due to the high regression coefficient R and the low error RMSE. This Fit is supported by the Ndrain-Ndrain predicted graph (Annex 10.5, bottom left), because the value of a point on the x-axis do not differ much from the value of a point on the y-axis. Also the trend line has a slope of almost 1 and the intersection with the x-axis is almost zero. This indicates that no overestimation or underestimation is done. However, the fit will be better, if the RRSE is decreased.

Notable is the gradient of the surplus-drainage graph (Annex 10.5, top right). The negative trend when the phosphorus is increased does not comply with the theory. According to the theory, the capacity of the soil to bind the phosphor decreases, when the surplus increases. In that case, drainage is increased, due to the less intake of the soil. In addition, the single point prediction (Annex 10.5, bottom right graph) does not comply with the trend of the total data set. The single point predictions (Table 6) are very steep and do not predict values that correspond with any other data within the given data set.

Table 6: Used point characteristics for the nominal equation of clay.

	Point 1
Psurplus	-2.00 kg/ha/yr.
Wseep	-83.09 mm/yr.
Wsurplus	388.16 mm/yr.
Precovery	-5.17 kg/ha/yr.
Soil physical unit	A
GT-class	Dry
Drains	Yes

In regard to the nominal regression equation, the non-nominal equation does fit better to the data of STONE, because R, RMSE and RRSE are more favourable. Also the amount of variables within the equation is smaller (Equation 23).

$$Pdrain = 0.437Psurplus + 0.006Wseep - 0.493Precovery + 0.665. \quad (23)$$

N = 77, R = 0.90, RMSE = 0.35, RRSE = 43.27%.

The non-nominal regression equation shows the same problems as the relation between phosphorus surplus and phosphorus drainage (Annex 10.6). Also, the single point prediction does not comply with the trend of the dataset (Table 7).

Table 7: Used point characteristics for the non-nominal equation of clay.

	Point 1
Psurplus	-2.00 kg/ha/yr.
Wseep	-83.09 mm/yr.
Precovery	-5.17 kg/ha/yr.

7.2.2. Sand

Regarding sand, no regression equation is found taking into account all considered values of the nominal equation. Weka does not take into account Psurplus as a determining factor. The factors: soil physical unit, Wsurplus and Precovery are considered to be more important. Neglecting these three factors, does provide a result. This result is an equation that does not contain many variables and has one nominal value, namely GT-class (Equation 24).

$$Pdrain = 0.018Psurplus + 0.003Wseep - 0.086 \text{ if 'GT-class=middle'} + 0.240 \text{ if 'GT-class=wet'} + 0.887. \quad (24)$$

N = 175, R = 0.87, RMSE = 0.33, RRSE = 49.36%.

The accuracy of the equation is better for data points with a drainage near zero. High surplus drainage values in STONE do differ more from the predicted value (Annex 10.7, bottom right). The overall accuracy is significant: the regression coefficient is high and the error is low. However, the relative error RRSE has a high value.

In the top right graph of annex 10.7, the relation between the phosphorus surplus and the drainage is presented with its prediction interval of 95%. There is a small positive relation and the graph fluctuates a lot. The prediction interval has a wide ranch, therefore predictions can differ significantly from the average value. The top values of the regression equation do not fit with the trend of the dataset.

A point prediction is made with the initial point from Table 8. The single point prediction line is steeper than the average trend line. However, the prediction line is drafted, inside the boundaries of the values presented in the graph (Annex 10.7, bottom right graph).

Table 8: Used point characteristics for the nominal equation of sand.

	Point 1
Psurplus	0.02 kg/ha/yr.
Wseep	-99.83 mm/yr.
GT-class	Dry

The non-nominal regression equation does not differ much from the nominal equation:

$$Pdrain = 0.013Psurplus + 0.003Wseep + 0.008\ Wsurplus - 1.6541. \quad (25)$$

N = 175, R = 0.87, RMSE = 0.33, RRSE = 48.29%.

The non-nominal equation has a better regression coefficient and a smaller error than the nominal equation. However, the fit of the predicted data on STONE seem worse (annex 10.8, bottom left graph). The trend of the predicted drainage increases, when the phosphorus surplus increases (annex 10.8, top right graph). The non-nominal single point prediction is executed with the data from Table 9 and is less steep than the single point prediction of the nominal equation (annex 10.8, bottom right graph).

Table 9: Used point characteristics for the non-nominal equation of sand.

	Point 1
Psurplus:	0.02 kg/ha/yr.
Wseep	-99.83 mm/yr.
Wsurplus	329.10 mm/yr.

7.3. Testing linear relation

All executed tests above are tested on their linearity with a residual plot. These residual plots are presented in annex 11. The clay plots (Figure 20 and Figure 21) both show a weak linear relation. This weak relation is determined by an extreme value. Neglecting this extreme value results in a negative linear relation. Therefore, there is no proof that the equation for sand is not linear.

The sand plots (Figure 22 and Figure 25), show a small indication that the relations are not linear. They show a double bow plot (Shalabh, sd). A double bow plot indicates that there is no similarity in the variance of the estimated values and that the equation can be a binomial distribution. The linearity problem can be solved by a transformation of parameters or by applying the weighted least square method (Shalabh, sd) (Statwing, sd). However, the transformation did not work, because the extreme values did not appear to have similar values. The weighted least square method is not executed.

To do a proper linearity analysis, nominal values have to be divided. It is possible that the presence of drains does show a relation and that this relation is corrected by a wet GT-class (Statwing, sd). The linearity test on two of the nominal values are shown in Figure 23 and Figure 24. Both test do not contrast with the residual plot of the entire data set. Other tests are not executed, because the amount of available data would be too small. In conclusion, the regression equation for sand is assumed to be linear, because the indications for a non-linear relation are not significant enough.

In the case of phosphorus (Figure 26, Figure 27, Figure 28 and Figure 29), for all cases, no indication is found that the regression analysis is not linear.

8. Discussion

This chapter discusses the research aspects that can be improved and the issues that were left out. Main discussion points are: the accuracy of the equations, the resemblance of the equations with the theory and the precision of prediction making.

During the preparation of the data generalizations were made and values were left out. Nutrient seepage, seepage and water drainage were generalized and denitrification is left out. Other regression results would be obtained when the data preparation was done differently. Furthermore, in the data preparation more attention could be paid to the values that are measurable. Because measurable values are applied in practice easier.

An important aspect is the validity of the equations. It is possible that the equation is too accurate (overfitting) or not accurate enough. The high t_c-values for the individual input disputes the accuracy of the data that is used. Furthermore, the regression of phosphorus is questioned, due the low regression coefficients with Pdrain and the high coefficient between Psurplus and Precovery. The input values' deviations, also question the accuracy. These deviations express themselves in the high RRSE for the regression equations. In addition, the prediction interval of the regression equation is high, due to the deviations among input data. A larger amount of input data could solve this. Despite these uncertainties, the regressions equations themselves have a R and RMSE that do not indicate a bad accuracy and do not imply overfitting. Furthermore, the data predicted with the equation fits well with the data from STONE. Still, more accurate regression equations can be established.

In general, the equations of phosphorus are more accurate established than the nitrogen equations. Also the non-nominal equations have better accuracy values. The non-nominal values also use less variables and do not contain nominal values. Therefore, the change at overfitting is smaller. If a model is more complex than another model that fits equally well, chance of overfitting is higher (Hawkins, 2003).

Considering accuracy, two results are not in accordance with the theory:
-The reduction of nutrient drainage in clay when drains are presence. In general, an increase is expected (Schipper P. , 2016). However, nutrient drainage can be reduced, due to the increasing mineralization in an area with drains (Scholefield, et al., 1993).
-When the phosphorus surplus in clay increases the phosphorus drainage in clay decreases. However, a positive relation is expected. This error is probably due to the phosphorus that has been stored in the soil for years. If a surplus of phosphorus was applied on the land ten years ago, the effects are still noticeable in the drainage of this year (Schoumans, Willems, & van Duinhoven, 2008). Also an inconvenient set of data could have changed the relation in a wrong way. This negative relation between the surplus and drainage is also found in the data derived from STONE.

A better understanding of the relations between drainage and the influence factors of drainage, will increase the accuracy of the regression equations. More research to the exact relation between drainage and the influence factors should be done. The nutrient cycle is dependent on many factors and therefore, it is complex (Oenema, van Liere, & Schoumans, 2004). The processes in the soil are considered as a black box, factors are generalized and small influences are neglected.

One of those processes that can be considered separately, is the nutrient storage in soil. In the current analysis recovery is examined. However, nutrient drainage is not only dependent on the amount of nutrients that are stored in a year. Nutrient drainage is also dependent on the amount of nutrients that is already stored in the soil. The extra storage of nutrients in the soil in a year, does not have an immediate effect on the nutrient drainage. The time between nutrient recovery and the drainage of the same nutrients could be years, especially in the case of phosphorus.

Further research will also address the uncertain values that are present in the nutrient cycle. Weather conditions, water balances and the disposal of fertilizers, do effect the nutrient cycle and are unpredictable (Perry, Robbins, & Barnes, 1988). Also STONE has its uncertainties, mainly the seepage and the physical or chemical processes in the soil (Groenendijk P. , 2016). To cover these uncertainties the equation should be validated with real field data instead of the data derived from STONE. Another option is the consideration of different input factors. Instead of nutrient surplus, vegetation type or the amount of cows could be considered. And instead of GT-class the average highest groundwater level or the average lowest groundwater level could be considered. Other uncertainties are the diversity of farm characteristics (Oenema, van Liere, & Schoumans, 2004). Different outcomes can appear between farms that seem to have the same characteristics.

Another aspect that influences the outcome of the regression analysis are the extreme values. Extreme values, derived from the STONE dataset, influence the equation in a high extent and cause insecurities. Regressions are fitted in a way to get a smaller error with the extreme values, with overfitting or an increased error towards average values as a possible consequence. High peaks and high errors between the predicted value and the real value, describe these extreme values. These values are characterized by high seep parameters and high or low nutrient surpluses. However, in most cases extreme values are determined by a combination of factors. All datasets for clay and sand, for phosphorus and nitrogen show extreme values. Executing the regression analysis without these extreme values, do not improve the equation. Due to the unpredictability of extreme values, no predictions can be done with nutrient surpluses, outside a certain interval (Annex 12). Furthermore, predictions made with values outside the interval do not confirm the linearity of the dataset. Due to the extreme value in the dataset of clay linearity for the equation of clay is not confirmed. When this value is left out, the trend line in Figure 20 and Figure 21 shifts significantly.

Accurate single point predictions are made for all regression equations, except for phosphorus in clay. Predictions made with the single prediction line share the direction of the accompanying surplus-drainage graph. However, none of the lines are parallel to the trend lines. This indicates that, the single prediction line does not comply with the dataset from the equation at points remote to the initial point (Annex 12). Therefore, predictions made with a reduction of the nutrient surplus bigger than 10% are assumed to be invalid. Also, predictions made outside the core interval are assumed to be invalid, because these points do not always follow the trend of the regression equation. The most accurate predictions are made with the equations of sand. The single prediction lines of sand do not differ much from the average trend line of the dataset.

The slope of the single point predictions do differ between the non-nominal and the nominal regression equations, due to the different parameter of Nsurplus. A possible explanation is that the dependency of factors on each other, is represented in the parameter of Nsurplus. This is the case, when determining factors are not taken into account in the non-nominal equation. Furthermore, the influence of seepage in the datasets is constant. Therefore, seepage is not dependent on other factors of the regression equation.

Relations are predicted more precise when the relation is established with a more specific analysis for a dataset with specific characteristics. However, the regression analyses of characteristic datasets show that it does not matter, whether data is filtered even further or not (Annex 8). On the other hand: the graphs (Figure 17 and Figure 23), Annex 7 and the fact that nominal values get allocated as determining factors, indicate that these factors are clustered as a group in the whole dataset. It is possible that, even sized datasets will show more clearance on that aspect. Although, a random test, with a smaller dataset did not show any indication of that being the case.

9. Conclusions

In this research the relation between nutrient surplus and nutrient drainage is determined with a multiple regression analysis. From this research conclusions can be drawn.

The influence factors vary per soil type and per linear regression approach, non-nominal or nominal. The only factor they had in common was seepage. The equations did not always involve nutrient surplus, because nutrient drainage is not dependent on nutrient surpluses in a high extent. But an equation that did involve nutrient surplus is chosen.

The nominal values taken into consideration do not show a major influence on the nutrient drainage. Soil physical units like: A, B, C or F, H, I are valued the same for nitrogen in sand. Besides, the non-nominal equations perform as well as the nominal equations. Furthermore, sand is differently subjected to nominal values like drains and irrigation, than clay. This is due to the different coarse structure of the soil (Perry, Robbins, & Barnes, 1988).

The established regression equations do show a significant correlation. The equations also fit well on the original dataset from STONE. The relationship between the nutrient surplus and the nutrient drainage is found to be statically significant. It is possible to draw two conclusions from the correlation coefficients and the errors (RMSE and RRSE):
-A test for an area with specific characteristics does not improve the regression equation.
-The non-nominal equation has slightly better verification indicators than the nominal equation.
In addition, the non-nominal equation is less likely to be subjected to overfitting, because it does not consider nominal values and there are less factors involved, what a more probable equation suggests (Hawkins, 2003). In general, the nominal and non-nominal equations share the same parameters for the variables Wseep. The same applies to Nsurplus in clay soils. This indicates a good estimation, because the values are found twice with different input.

Predictions that are done with the regression equation are estimated within a significant prediction interval. The prediction lines that forecast the effect of decreasing the nutrient surplus show similarities with the trend line, within a certain interval. Despite not taking into account indirect relations and the insecurities of the regression equations, it is possible to make an estimation about the effect of decreasing the nutrient surplus. This estimation is made, considering a maximum reduction of the nutrient surplus of 10%.

This research suggests that predictions of the effects of nutrient surplus decreasing measures on the nutrient drainage should be done with a non-nominal equation. This does not mean that nominal values do not affect the nutrient drainage. Datasets and single point prediction lines with the same characteristics appear close to each other and datasets with the same characteristics can be concentrated. Furthermore, relations that show the effect of nominal values on the nutrient drainage are found.

In conclusion, the non-nominal regression equations obtained for sand and the non-nominal equations obtained for nitrogen in clay do predict the nutrient surplus. However, to be more accurate issues have to be dealt with. The number of data should be bigger. Also the role of extreme values within the equations should be examined more precise.

10. Recommendations

Further research can be performed in a couple of fields: the examination of other and bigger datasets, the consideration of other factors that determine the nutrient drainage, a more specific aimed regression analysis and the consideration of other measures that decrease the nutrient drainage.

To improve the validity of the research, it is advised to use a bigger dataset. Due to the small dataset, clay and in a less extent sand, can be subjected to overfitting and inaccurate prediction making. The validity can also be improved by performing a control test with another dataset. A regression with other datasets derived from other models can deliver different results. In addition, validity is improved, when results are tested with other data than the data derived from STONE. Real data, measured in the field should be used as comparison material. For instance, nutrient concentrations: measured in drains, in surface waters or at pumping stations, serve as control sample. Last, examination of the influence of extreme values can administer uncertainties. Leaving out extreme values or examining the role of extreme values permit other data to be estimated more precise.

Extra literature study will provide better insight on the processes, on which the regression equations are based. The nutrient cycle and the factors in it are not well established. Literature study contributes by considering the influence of the factors on each other. Consideration of these indirect influences will improve the prediction of the nutrient surplus. On the other hand, it will make the model more complex. Literature study will also give more insight on the established theories and relations and their similarities with the regression equations.

Further research is also possible to taking into account other factors during the regression analysis. Instead of using GT-class, it is possible to use the highest average groundwater level or the lowest average groundwater level. Furthermore, to do a proper examination of the effects of phosphorus surplus on the drainage, multiple years of phosphorus recovery or the initial storage in the soil should be taken into account. However, this will complicate the regression model complex and the required data is difficult to collect. A similar examination can be done to improve the regression for nitrogen.

Recommended is a further development of the prediction tool. At the moment, the tool only represents the effect of the nutrient surplus on the nutrient drainage at grassland on clay or sand soils. It is possible to add other features to the tool like:
1. Other types of land use and soils. For instance, the effect of nutrient surpluses in corn fields or the effect on peat soils.
2. Extrapolation to areas outside of Fryslân. It is possible to apply the method used on all areas in the Netherlands.
3. The examination of other measures that are proposed by Alterra to decrease the nutrient surplus. The method followed in this research can be used to examine the effect of the Gt-class or soil improvement. This research establishes some predictions about the effect of decreasing the GT-class, but a research aimed at that aspect alone, would establish clearer and more trustworthy relations. Furthermore, GT-class is not considered in all the obtained relations.
4. The involvement of the tool in another prediction tool, for example one that involves a costs prediction. Due to the low running time of this tool, implementation in other models is possible.
5. The subdivision of the data into groups with specific spatial characteristics. This research is executed for nitrogen in sand, but not for clay and phosphorus. It is possible to conclude from this research that a regression analysis to areas with specific characteristic do not improve the predictions made with the obtained equations. However, this research also implies that nominal values have a different impact on the nutrient drainage, thus tailor made regression equations for

specific areas will have different outcomes. At the moment, this kind of research is restricted by the insufficient amount of data.

11. Applicability on farm level

To illustrate the relevancy of this research, the implementation of the regression equations into practice is explained. Two case studies are executed using the regression equations established for nitrogen in sand. The first case study uses data derived from STONE calculations and predicts an overall effect of decreasing the nitrogen surplus. The second case study uses mapped data and predicts the effect for two areas. It is possible to visualise the results of the case studies in the province of Fryslân. This will give a clearer understanding of the effect of decreasing the nitrogen surplus. However, this is not executed in this research. Furthermore, this case study will consider the selection of data and the selection of data related to mapping the effects.

To do a proper spatial analysis, two aspects should be considered: the variables used in the regression equations and the spatial data used. It is important to choose regression equations that only involve variables, whose values can be determined easily in practice. Variables that are known by the farmer or are based on spatial datasets are eligible. Taking into consideration nutrient recovery or denitrification, obliges to making assumptions, while a variable like the presence of drains is easier to determine. Furthermore, a seepage map is not that accurate and different values for seepage are found in different maps of the same area (de Lange, et al., 2010). During the examination of the maps in ArcGIS, inconsistencies between spatial data was found. For instance, the soil physical unit was allocated differently at some places. A possible reason to explain this difference, is that some maps are not up to date. In time, loamy sand soils had become sandier during the years. Furthermore, every year, the groundwater level differs and irrigation is applied on a different location.

11.1. Case study 1: STONE data.

Using only STONE data, a prediction is done that indicates the effect of decreasing the nitrogen surplus in the area that is characterised by grassland and sandy soil. Two assumptions are made:
-A reduction of the nitrogen surplus is assumed to be possible for areas with a surplus higher than 100kg/ha/yr. (Schipper P. , 2016).
-To stay within the validity range, the nitrogen surplus is decreased with 5%. The accuracy of a 5% reduction will be higher than the accuracy of a 10% reduction.

About 50% of the STONE plots has a nitrogen surplus higher than 100kg/ha/yr. (Figure 11). This concerns 82 plots with different sizes. Each plot consists of multiple grids with a size of 250x250 m^2 spread across the province of Fryslân (Figure 12) (Wolf, et al., 2003). Reducing the nitrogen surplus, two different results are obtained. According to the nominal regression a nitrogen surplus reduction of 5% will decrease the surplus with 0.67-5.26%. The non-nominal regression predicts a bigger effect, between 1.87-12.35%. The bigger reduction of the non-nominal equation is caused by the higher value of the regression coefficient for the dependency on Nsurplus. Also, the predictions made with the non-nominal equation contain more high values than predictions made with the nominal equation.

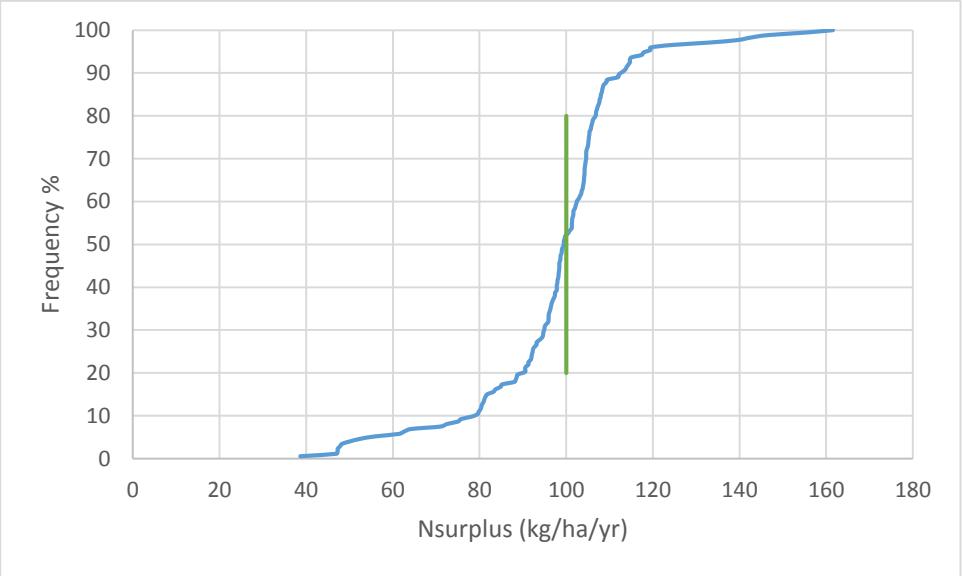


Figure 11: Percentage of values with a surplus higher than 100kg/ha/yr. in sand.

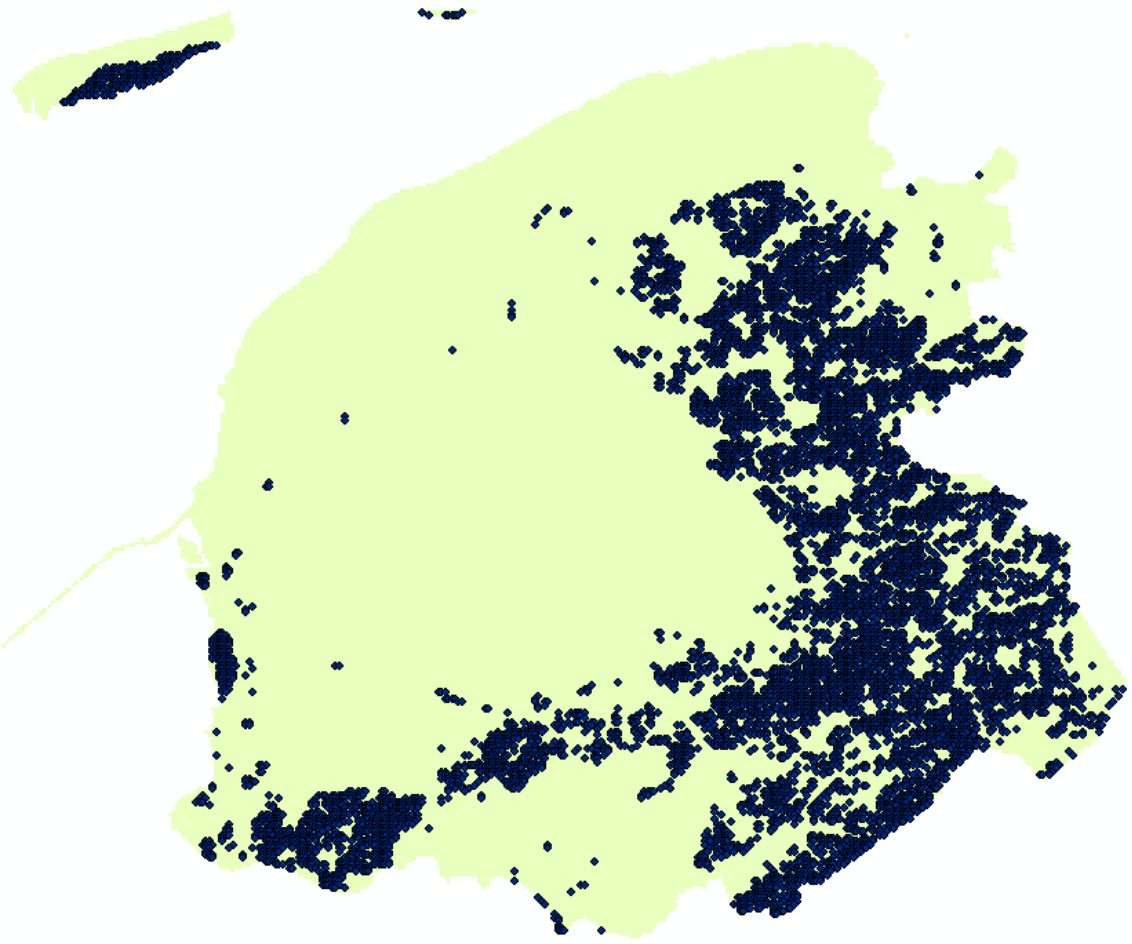


Figure 12: Areas with grassland and sandy soils in the province of Fryslân.

The next chapter will be left out in the final thesis report. Due to problems with data, the speed of my computer running ArcGIS and time, I will not be able to execute this, within 5 days.

11.2. Case study 2: Mapped data.

The second case study uses spatial data to predict the factors that influence the nitrogen drainage. The prediction is done with the nominal equation, because the non-nominal equation involves Nrecovery, which is not determined easily. All variables within the nominal equation can be derived from spatial data, except for nitrogen surplus, which is derived from STONE. Spatial data is less based on generalisations than the data from STONE. Therefore, more precise predictions can be done with the spatial data. Furthermore, the spatial data is more recent. The maps collected to execute the study are:

- The land use (Hazeu, et al., 2014).
- The soil type, divided in 24 soil physical units (PAWN) (Wösten, et al., 2013).
- The location of STONE plots and the accompanied nitrogen surplus (Wolf, et al., 2003).
- The seepage (Nederlands hydrologisch instrumentarium, 2016).
- The GT-class (Wösten, et al., 2013).
- The presence of Drains. (Massop & Schuiling, 2016)
- The presence of irrigation (Massop, Schuiling, & Veldhuizen, 2013).

The spatial data obtained are presented in Table 10 and Table 11. These tables also present the nitrogen drainage calculated with the regression equation.

Table 10: Current nitrogen drainage, calculated with the regression equation and the spatial data derived from the maps.

	Area 1	Area 2
Nitrogen surplus	101.97 kg/ha/yr	107.64 kg/ha/yr
Seepage	0.19 mm/dag = 69.35 mm/yr	0.43 mm/dag = 156.95 mm/yr
Soil physical unit	I	H
GT-class	Wet	Wet
Presence of drains	No	No
Presence of irrigation	No	No
Predicted nitrogen drainage	19.06 kg/ha/yr	21.63 kg/ha/yr

Table 11: Predicted nitrogen drainage, when the nitrogen surplus is reduced with 5%. Calculated with the regression equation and the spatial data derived from the maps.

	Area 1	Area 2
0.95 × Nitrogen surplus	96.87	102.26 kg/ha/yr
Seepage	0.19 mm/dag = 69.35 mm/yr	0.43 mm/dag = 156.95 mm/yr
Soil physical unit	I	H
GT-class	Wet	Wet
Presence of drains	No	No
Presence of irrigation	No	No
Predicted nitrogen drainage	18.81 kg/ha/yr	21.37 kg/ha/yr

According to the nominal regression a nitrogen surplus reduction of 5% will decrease the nitrogen surplus in area 1 with 1.31% and in area 2 with 1.20%. This prediction is made with new input data. To validate the outcome of the regression equation, the outcome of Table 10 is compared to the simulated drainage value of STONE. STONE simulates a nitrogen drainage of 19.15 kg/ha/yr for area 1 and a drainage of 22.34 kg/ha/yr for area 2. Both outcomes are within range of the prediction interval.

11.3. Conclusion

To put the regression equation into practice, up-to-date data is necessary. This can be spatial data or data received from an interview with a farmer. The effect of reducing the nutrient surplus, can be predicted in the whole area that is characterised by grassland and sandy soil. This is possible by extrapolating the method of case study 2. An overlay of maps will provide all the needed variables to determine the nutrient drainage at a specific location.

12. References

Abernethy, M. (2010, April). *Data mining with Weka, Part 1: introduction and regression*. Retrieved from IBM developersWorks: https://www.ibm.com/developerworks/library/os-weka1/

Alterra. (2016). *wateronderzoek bij Alterra*. Retrieved from WageningenUR: http://www.wageningenur.nl/nl/Expertises-Dienstverlening/Onderzoeksinstituten/Alterra/Expertisegebieden/Water.htm

Alterra. (n.d.). *STONE*. Retrieved April 6, 2016, from Alterra, WageningenUR: http://www.wageningenur.nl/nl/Expertises-Dienstverlening/Onderzoeksinstituten/Alterra/Faciliteiten-Producten/Software-en-modellen/STONE.htm

Anglim, J. (2011, April 28). *Rules of thumb for minimum sample size for multiple regression*. Retrieved from Cross Validated: http://stats.stackexchange.com/questions/10079/rules-of-thumb-for-minimum-sample-size-for-multiple-regression

BBC. (2014). *Eutrophication*. Retrieved from BBC, Science, pollution: http://www.bbc.co.uk/schools/gcsebitesize/science/edexcel/problems_in_environment/pollutionrev4.shtml

de Lange, W., Delsman, J., Prinsen, G., van Bakel, J., Massop, H., & Hoogewoud, J. (2010). *Vergelijking NHI 2.0 PAWN-STONE.* Nationaal hydrologisch instrumentarium. Retrieved from http://www.nhi.nu/ref_modelv2/Vergelijking%20NHI2.0%20met%20PAWN%20en%20STONE2.3%20v0.8.pdf

Doorn, P. K., & Rhebergen, M. P. (2006, December 15). *statistiek voor historici*. Retrieved from Leiden University: http://www.let.leidenuniv.nl/history/RES/stat/html/les10.html

Frost, J. (2015, September). *The danger of overfitting regression models*. Retrieved from The minitab blog: http://blog.minitab.com/blog/adventures-in-statistics/the-danger-of-overfitting-regression-models

Groenendijk, P. (2016, April). Working with STONE. (A. Mossink, Interviewer)

Groenendijk, P., Renaud, L., van der Salm, C., Leusink, H., Blokland, P. W., & de Koeijer, T. (2015). *Nitraat en N- en P-uitspoeling bij de gebruiksnormen van het 5de NAP.* Wageningen: Alterra Wageningen UR. Retrieved from http://library.wur.nl/WebQuery/wurpubs/fulltext/343644

Hawkins, D. M. (2003, October). The problem of overfitting. *Journal of chemical information and computer science*, 1-12.

Hazeu, G. W., Schuiling, C., Dorland, G. J., Roerink, G. J., Naeff, H. S., & Smidt, R. A. (2014). *Landelijk grondgebruiksbestand Nederland versie 7 (LGN7).* Wageningen: Alterra WageningenUR. Retrieved from http://content.alterra.wur.nl/Webdocs/PDFFiles/Alterrarapporten/AlterraRapport2548.pdf

Hoekstra, A. Y. (2013). *Water.* Universiteit Twente.

LTO Nederland. (2013, Januari). Deltaplan Agrarisch Waterbeheer. Retrieved from http://agrarischwaterbeheer.nl/system/files/documenten/pagina/brochure_daw_januari_2013.pdf

Massop, H. T., Schuiling, C., & Veldhuizen, A. (2013). *Potentiele beregeningkaart 2012.* Wageningen: Alterra WageningenUR. Retrieved from Potentiele beregeningskaart 2012 : update landelijke potentiele beregeningskaart voor het NHI op basis van landbouwmeitellingen 2010

Massop, H., & Schuiling, C. (2016). *Buisdrainagekaart 2015.* Wageningen: Alterra WageningenUR. Retrieved from http://edepot.wur.nl/370378

McClave, J. T., Benson, P. G., Sincich, T., & Knypstra, S. (2011). *Statistiek.* Pearson.

Nederlands hydrologisch instrumentarium. (2016, Juni). Retrieved from http://www.nhi.nu/nl/index.php/data/nhi-lhm/uitvoer/v302/kwel/

Ng, A. (Director). (n.d.). *The problem of overfitting* [Motion Picture]. Retrieved May 2016, from https://class.coursera.org/ml-005/lecture/39

Oenema, O., van Liere, L., & Schoumans, O. (2004, July 1). Effects of lowering nitrogen and phosphorus surpluses. *Hydrology*, 289-301.

Perry, C. A., Robbins, V., & Barnes, P. L. (1988). *Factors affecting leaching in agricultural areas and an assessment of agricultural chemicals in the ground water of kansas.* U.S. Geological survey. Retrieved from http://pubs.usgs.gov/wri/1988/4104/report.pdf

Planbureau voor de leefomgeving, Centraal bureau voor statistiek and WageningenUR. (2015). *Algemene fysisch-chemische waterkwaliteit KRW*. Retrieved from Compendium voor de Leefomgeving: http://www.compendiumvoordeleefomgeving.nl/indicatoren/nl0252-Fysisch-chemische-waterkwaliteit-KRW.html?i=25-107

Planbureau voor leefomgeving. (2015). *Waterkwaliteit Kaderrichtlijn Water 2015 (KRW)*. Retrieved from Geoservice: http://geoservice.pbl.nl/website/Arcgisonline/basicviewer_KRW/index.html?appid=bf816666fbe14e02a331f1dfb8038587

Rijkswaterstaat, Ministerie van Infrastructuur en milieu. (n.d.). *Handboek water*. Retrieved maart 2016, from Kenniscentrum InfoMil: http://www.infomil.nl/onderwerpen/klimaat-lucht/handboek-water/wetgeving/waterwet/doelstellingen/waterkwaliteit/

Schipper, P. (2016, May). Applying nutrient measures in practice. (A. Mossink, Interviewer)

Schipper, P. (2016, June). The integration of the models SWAP and ANIMO in STONE. (A. Mossink, Interviewer)

Schipper, P., & van Boekel, E. (2016, April 13). Herkomst van stikstof en fosfor in het oppervlaktewater voor zes polders in het beheergebied van Wetterskip Fryslân. (A. Mossink, Interviewer)

Scholefield, D., Tyson, K. C., Garwood, E. A., Armstrong, A. C., Hawkins, J., & Stone, A. C. (1993). Nitrate leaching from grazed grassland lysimeters: effects of fertilizer input, field drainage, age of sward and patterns of weather. *European Journal of soil science*, 601-613.

Schoumans, O., Groenendijk, P., Renaud, L., van Dijk, W., Schröder, J., van den Ham, A., & Hooijboer, A. (2012). *Verhoogde nitraatconcentraties in het Zuidelijke zandgebied.* Alterra, WageningenUR. Wageningen: Alterra.

Schoumans, O., Willems, J., & van Duinhoven, G. (2008). *30 vragen en antwoorden over fosfaat in relatie to landbouw en milieu.* Alterra, WageningenUR, Wageningen. Retrieved from Alterra:

http://content.alterra.wur.nl/webdocs/internet/corporate/prodpubl/boekjesbrochures/30vragen_fosfaat.pdf

Shalabh. (n.d.). *Linear regression analysis, model adequacy checking*. (Indian institute of technology Kanpur) Retrieved May 2016, from http://nptel.ac.in/courses/111104074/Module4/Lecture17.pdf

SONDZ. (n.d.). *regressie analyse.* Retrieved April 2016, from www.sondz.nl: http://www.sondz.nl/downloads/regressie.pdf

Statwing. (n.d.). *interpreting residual plots to improve your regression*. Retrieved May 2016, from Statwing: http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#outlier-header

Sykes, A. (1993). *An introduction to regression analysis.* University of Chicago Law School. Retrieved from http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1050&context=law_and_economics

The University of Waikato. (n.d.). *Nutrient Cycling*. Retrieved April 2016, from Science on the farm: http://sci.waikato.ac.nz/farm/content/nutrientcycling.html

University of Waikato. (2016, April 13). *Weka 3: Data Mining Software in Java*, Weka 3.8. Retrieved from Machine learning group at the University of Waikato: http://www.cs.waikato.ac.nz/ml/weka/index.html

van Boekel, E., Roelsma, J., Massop, H., Hendriks, R., Goedhart, P., & Jansen, P. (2012). *Nitraatconcentraties in het drainwater in zeekleigebieden.* Alterra, WageningenUR. Wageningen: Alterra. Retrieved from http://library.wur.nl/WebQuery/wurpubs/fulltext/256852

van Gaalen, F., Tiktak, A., Franken, R., van Boekel, E., van Puijenbroek, P., Muilwijk, H. (2016, January) Waterkwaliteit nu en in de toekomst.
http://www.pbl.nl/sites/default/files/cms/publicaties/PBL_2016_Waterkwaliteit%20nu%20en%20in%20de%20toekomst_1727.pdf

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.).

Witten, I. H. (Director). (2013). *Predictive Analytics Training with Weka (Linear regression)* [Motion Picture]. New Zealand.

Wolf, J., Beusen, A., Groenendijk, P., Kroon, T., Rötter, R., & van Zeijts, H. (2003, January). The integrated modeling system STONE for calculating nutrient emissions from agriculture in the Netherlands. *Environmental Modelling & Software*, 597-617.

Wösten, H., de Vries, F., Hoogland, T., Massop, H., Veldhuizen, A., Vroon, H., . . . Bolman, A. (2013). *BOFEK 2012, de nieuwe, bodemfysische schematisatie van Nederland.* Wageningen: AlterraWageningenUR.

Annex 1: Research methodology

The steps taken in this research are presented Figure 13.

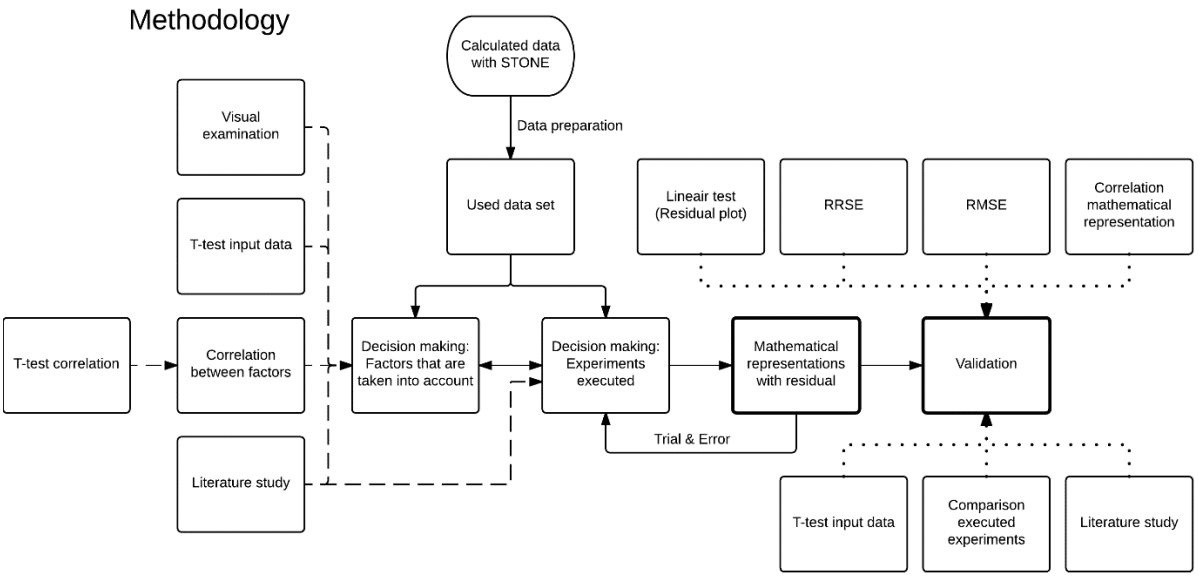


Figure 13: The research steps taken (solid lines), the factor analysations (dashed lines) and the validation tests (dotted lines).

Annex 2: Definitions and factors nutrient cycle

In the research abbreviations are use. A lists of these abbreviations with their explanation and units is given below.

Table 12: Lists of factors in the nutrient cycle and their abbreviations.

In research (abbreviation)	Explanation	Consists of	Unit	Data type
Ndrain	The total amount of nitrogen disposed into surface water. (nutrient discharge/nutrient drainage)	Leaching and run-off	Kg/ha/yr. or concentration	Numerical
Pdrain	The total amount of phosphorus disposed into surface water.	Leaching and run-off	Kg/ha/yr. or concentration	Numerical
Nsurplus	Surplus of nitrogen amount.	Vegetation, fertilizer use and atmospheric deposition.	Kg/ha/yr.	Numerical
Psurplus	Surplus of phosphorus amount.	Vegetation and fertilizer use.	Kg/ha/yr.	Numerical
Soil physical unit	The composition of the soil.		A, B, C, D, E, F, G, H, I, J, K	Nominal
Wseep	The seepage of water through soil in or out the area.		mm/yr.	Numerical
Nseep	Amount of nitrogen transported by seepage.		Kg/ha/yr.	Numerical
Pseep	Amount of phosphorus transported by seepage.		Kg/ha/yr.	Numerical
Nrecovery	Storage change of nitrogen in the soil.	Organic and mineral recovery.	Kg/ha/yr.	Numerical
Precovery	Storage change of phosphorus in the soil.	Organic and mineral recovery.	Kg/ha/yr.	Numerical
Ndenitrification	Denitrification of nitrogen by bacteria.		Kg/ha/yr.	Numerical
Wsurplus	Amount of rainwater left after evapotranspiration.	Evapotranspiration crops, evapotranspiration soil and precipitation.	mm/yr.	Numerical
GT-class	Fluctuations in the groundwater level or saturation.	Summer and winter groundwater level	Wet, Middle, Dry	Nominal
Irrigation	The application of irrigation.		Yes, no	Nominal
Drains	The presence of tile drains.		Yes, no	Nominal
Wdrainage	Amount of water disposed into surface water.	Leaching and run-off	mm/yr.	Numerical

Substitution of Soil physical units

The system of soil physical units is classified into 21 different soil types. Considering clay, 5 types of soils are present, in the selected data. Sand has 6 types of soil (Wösten, et al., 2013). In the calculations the soil types are referred to as: A, B, …, or K. This substitution is done to make sure soil physical units are taken into account as categorical data in the regression analysis (Table 13).

Table 13: The soil types taken into account in this research.

In this research	Explanation	In STONE data (code)	Subscription in STONE data
Clay			
A	Coarse sand	15	zavel_M8 M8 Mn25A
B	Light clay	16	lichtklei M10, M11, R3, R10 gMn85C
C	Dense clay	17	zwaarklei M22, R7 Mn45A
D	Clay on peat	18	kleiveeneu M18, R5 Rn44Cv
E	Clay on sand	19	kleizand M7 MOb72
Sand			
F	Sand on peat on sand	5	meerveen V13, V14 zWp
G	Drift sand	7	stuifzand Z4, Z27 Zn21
H	Light loamy coarse sand	9	podzolZ8 Z8 Hn21
I	Loamy coarse sand on loam	11	podzlZ8x Z8x Hn23x
J	Coarse sand with humus	12	enkeerdz Z16 zEZ21
K	Dense loamy sand	13	beekeerd Z20 pZg23

GT-Classes

'GT-class' or 'grondwatertrap', represent the fluctuations in the groundwater level. It can be divided into different classifications (Table 14). The classification wet-middle-dry is taken into account in this research.

Table 14: Categorization of GT-class. With GHG=mean highest groundwater level measured in winter.

In this research	In literature and STONE data	Metric value
Wet	Gt-class I, II, III, V and V*	GHG < 40cm below surface
Middle	Gt-class IV and VI	40cm below surface < GHG < 80cm below surface
Dry	Gt-class VII and VIII	GHG > 80cm below surface

GHG = mean highest groundwater level measured in winter.

Annex 3: Use of models

The STONE model

The process-based STONE-model generates nutrient emission datasets. A specific combination of input variables is used to calculate 6405 different scenarios that represent areas in the Netherlands. STONE consist of multiple sub models, that represent for instance the hydrology or the nutrient storage (Wolf, et al., 2003)(Figure 14). One of those models is SWAP. SWAP calculates the moisture of the soil. Another model is ANIMO. ANIMO (Figure 15) describes the N and P controlling processes in each layer of soil in a deterministic and integrated way (Schipper P. , The integration of the models SWAP and ANIMO in STONE., 2016) The aim of STONE is to simulate the consequences of fertilizer use for the emission of nutrients to groundwater and surface waters. It combines factors, like aerobic condition and groundwater level, into a value that represents nutrient leaching or run-off (Alterra, sd).

The output of STONE consists of plots. Each plot contains 46 attributes that represent factors of the nutrient cycle at an area of 250x250 m^2 somewhere in the Netherlands (Groenendijk P. , 2016).

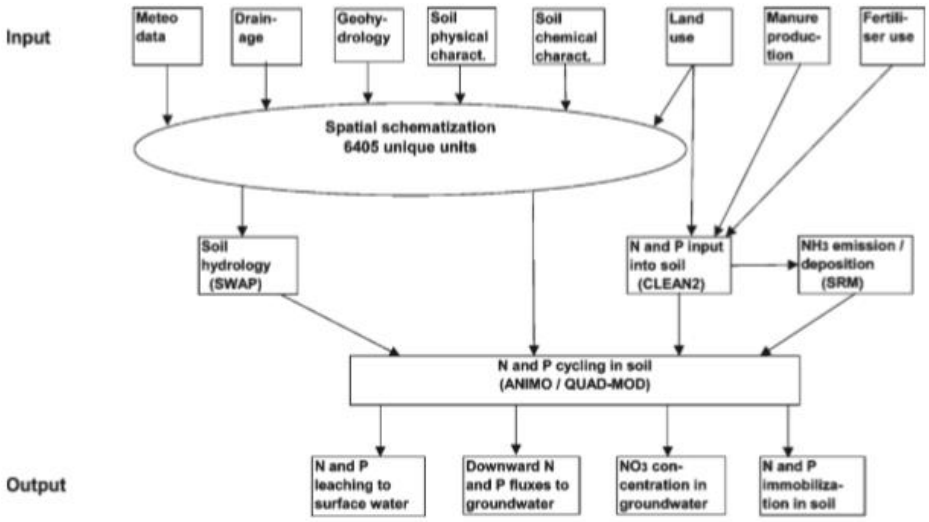


Figure 14: Overview of input data, modelled processes in different components and output of the STONE modelling system (Wolf, et al., 2003).

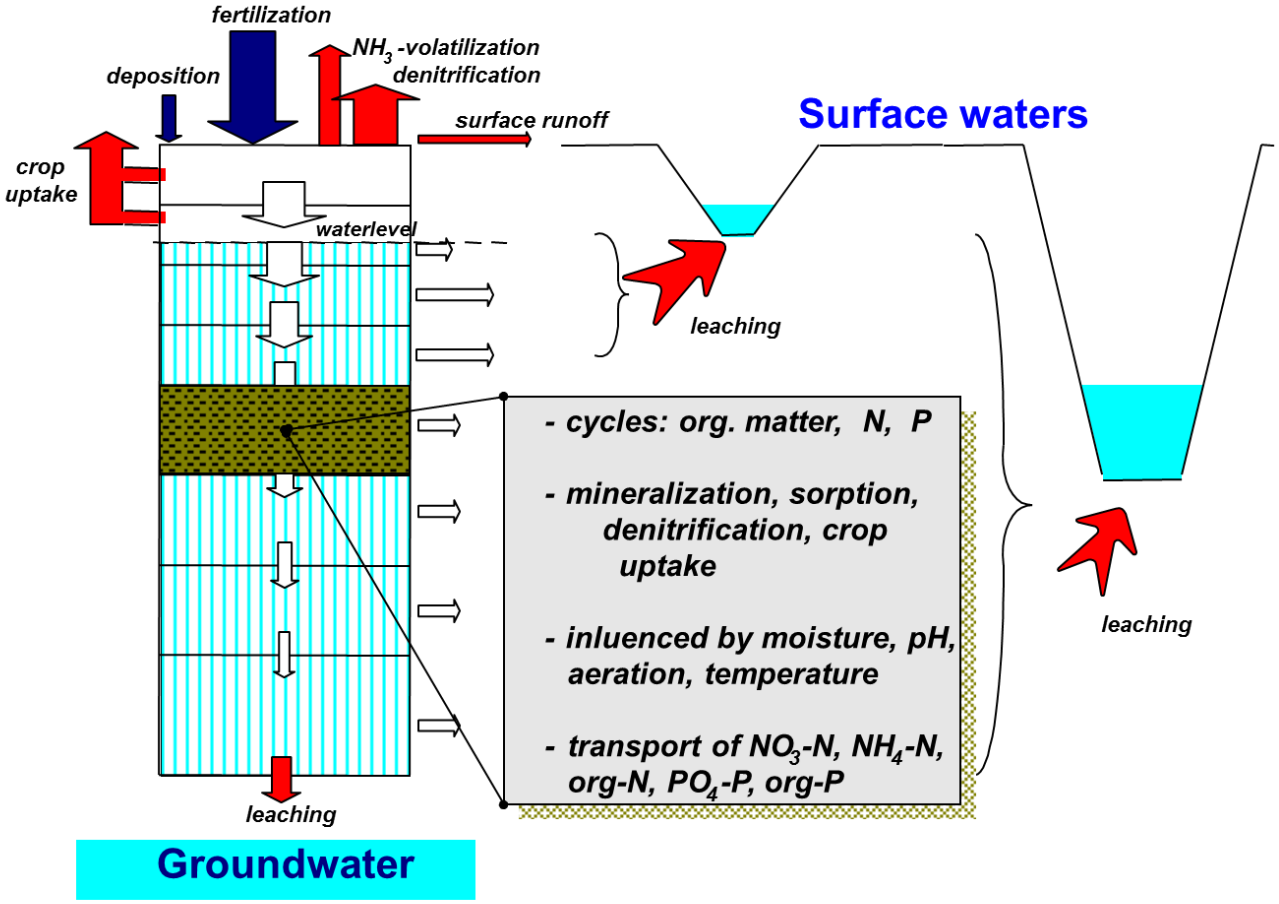


Figure 15: The overview of the nutrient cycle as considered in the model ANIMO.

Weka

Weka is data mining software. The tool is used to visualise graphs and to perform a linear regression analysis (Abernethy, 2010) (Witten, 2013). It outcomes are: a regression equation, the equation's correlation coefficient and the equation's confidence (measured by RMSE and RRSE). Weka neglects all the attributes that do not statistically contribute to the accuracy of the model, resulting in the most precise factors.

The regression in Weka is done with the leave-one-out 10 folds cross validation technique. The data set is divided in 10 equal parts. 1 Piece is taken out as a test set and is compared with the other 9 training sets. This is done for all pieces and the results are combined into average values. Finally, a final piece is taken and tested on the whole result (Witten, 2013).

Annex 4: t-Test for data input.

In Table 15, the deviation of the input data at a 95% confidence interval is calculated as a percentage of the mean. To achieve a confidence interval of 95%, the data set should be in between -0.05 and 0.05, to be sufficient.

Table 15: T-test executed for the data input. 0.05=5%.

	NDrain	Nsurplus	Nseep	Ndenitrification	Nrecovery	Wsurplus	Wdrainage	Wseep
Clay	0.116	0.026	0.445	0.022	0.254	0.013	0.076	0.661
Sand	0.089	0.030	0.444	0.035	0.153	0.009	0.106	-0.584
	PDrain	Psurplus	Pseep	Precovery	Wsurplus	Wdrainage	Wseep	
Clay	0.085	-0.170	0.421	-0.092	0.013	0.076	0.661	
Sand	0.178	2.987	0.459	-1.210	0.009	0.106	-0.584	

Annex 5: Data visualization

This annex shows the relations between factors in the nutrient cycle visually.

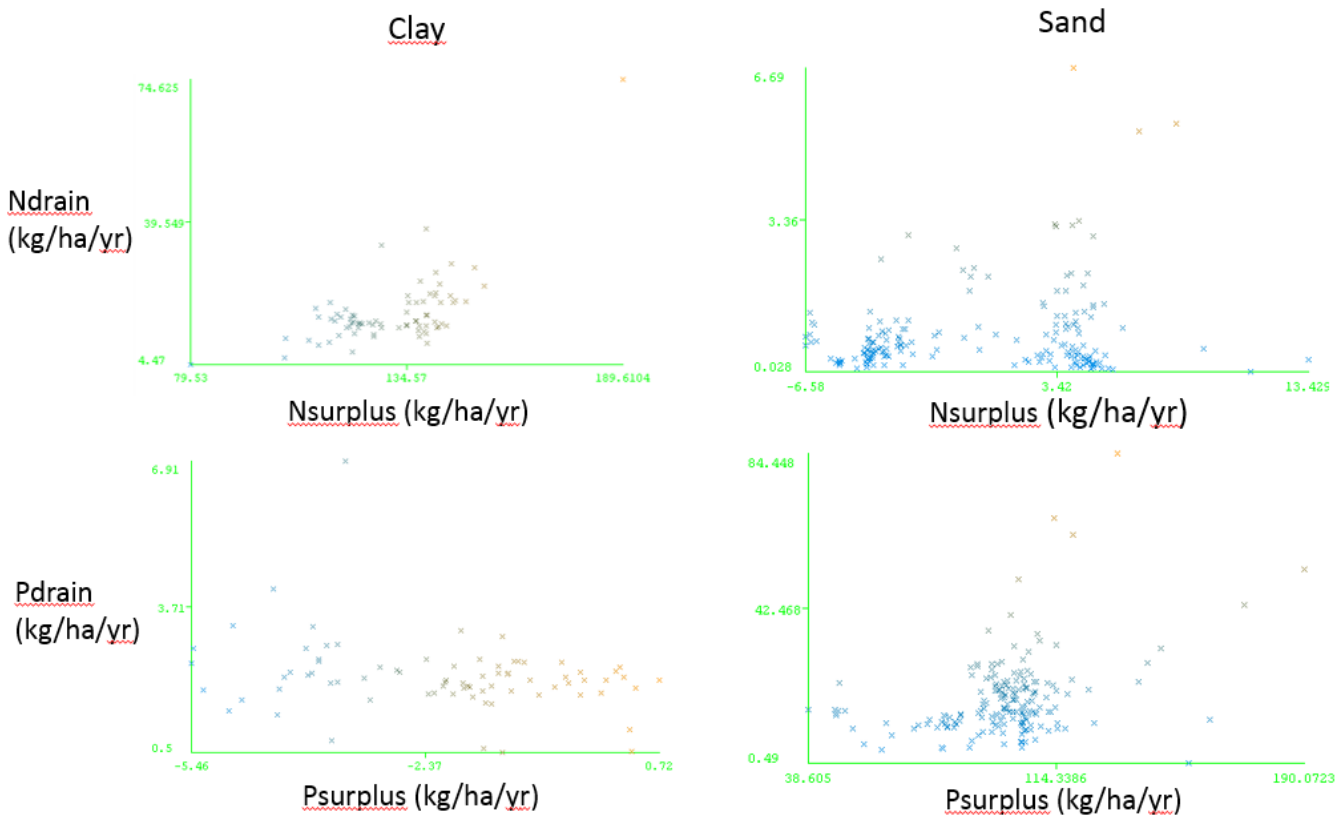


Figure 16: Surplus-Drainage graph. Top left: nitrogen in clay soil, top right: nitrogen in sand soil, bottom left: phosphorus in clay soil and bottom right: phosphorus in sand soil.

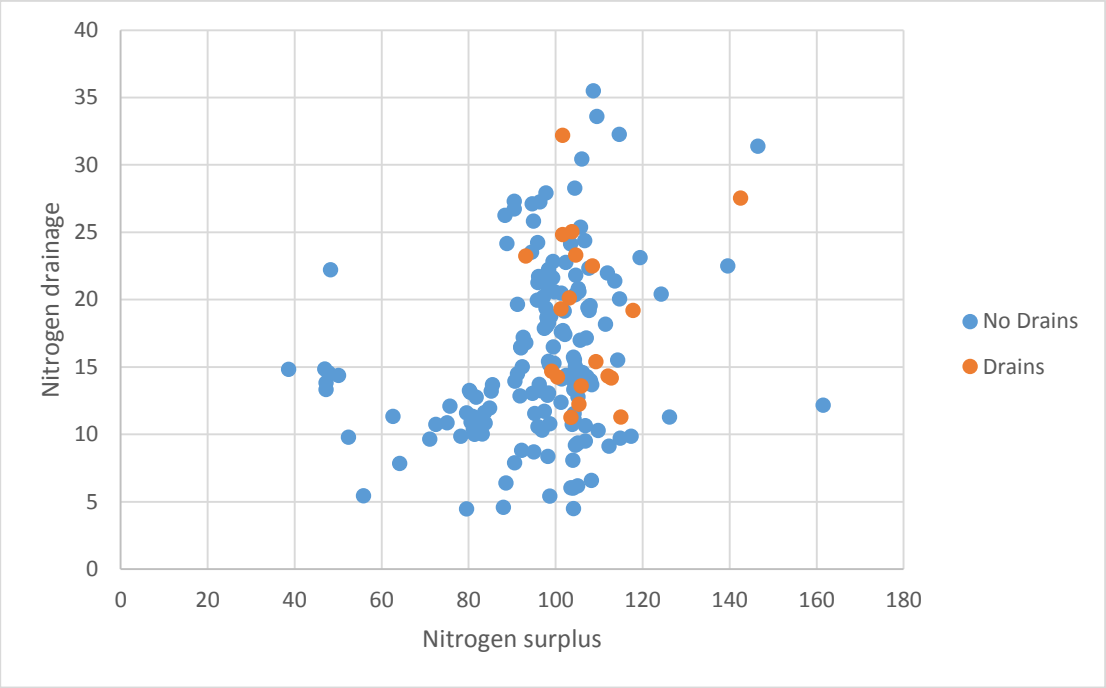


Figure 17: Presence of drains in clay.

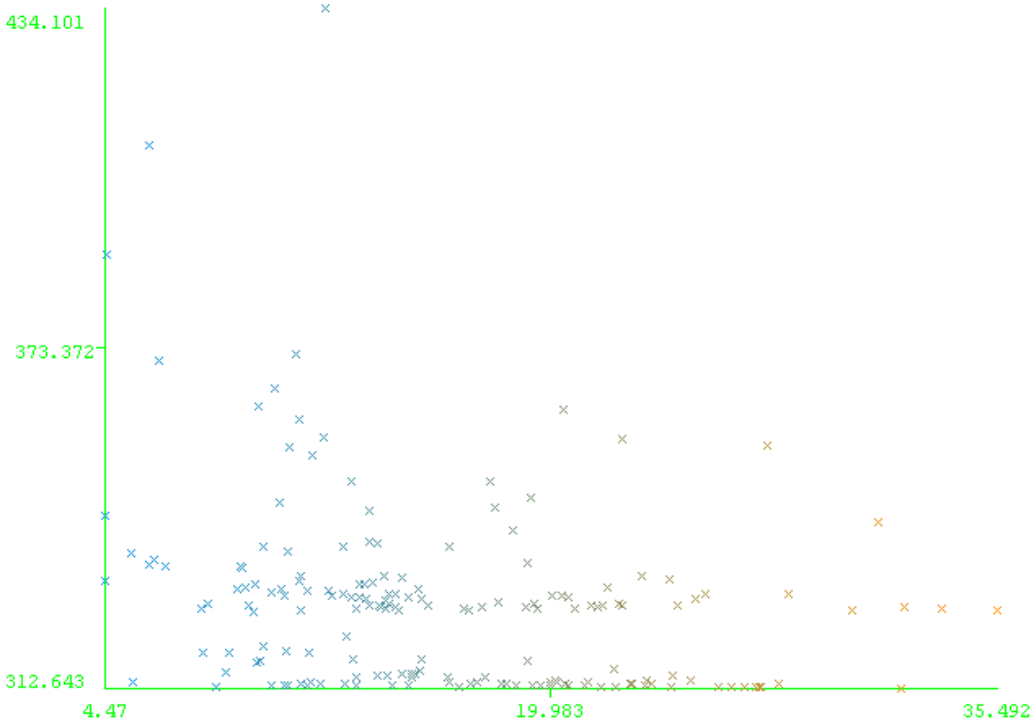


Figure 18: Ndrain (x)-Wsurplus (y) in clay.

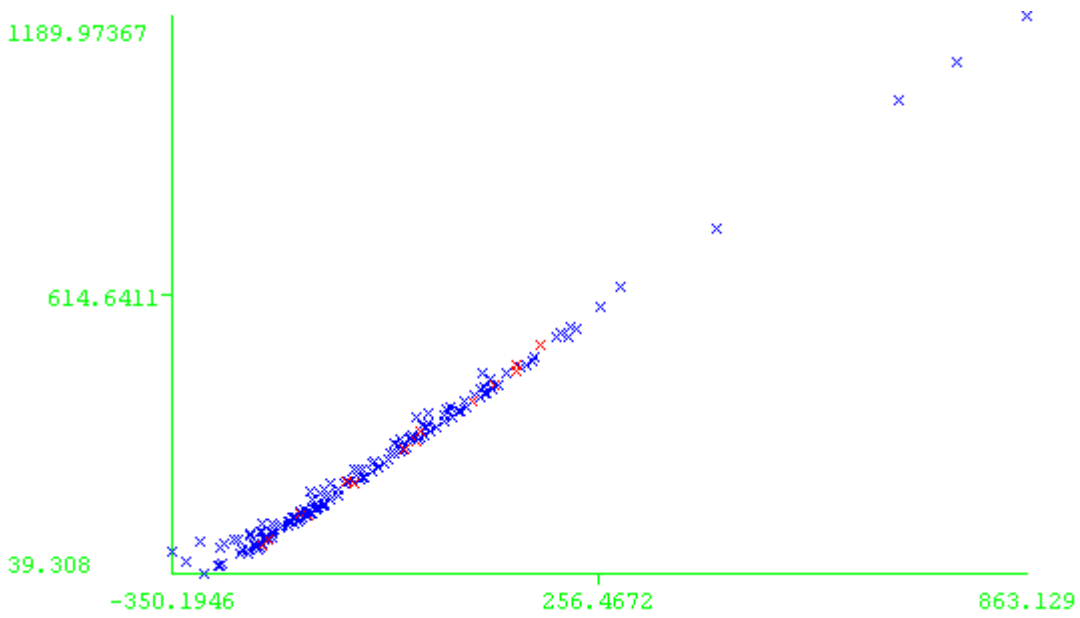


Figure 19: Wseep (x) - Wdrainage (y) in clay.

Annex 6: Correlation results

To examine how well factors are related to each other the correlation coefficients between factors are determined. Red indicates an incorrect correlation. Blue indicates a high correlation and therefore a high dependency between factors.

Clay

Table 16: Correlation between nitrogen factors. Red = incorrect relation. Blue = high dependency.

R	*NDrain*	*Nseep total*	*Ndenitrification*	*Nrecovery total*	*Nsurplus*	*Wsurplus*	*WDrainage*
Nseep total	0.896						
Ndenitrification	0.102	0.090					
Nrecovery total	0.509	0.426	-0.194				
Nsurplus	0.648	0.492	0.612	0.608			
Wsurplus	0.273	0.220	0.289	0.079	0.344		
WDrainage	0.872	0.866	0.045	0.651	0.620	0.162	
Wseep	0.841	0.843	0.001	0.647	0.575	0.010	0.988

Table 17: t-Value of the correlations. The t-value for the null hypothesis is 1.992. red = uncorrelated for NDrain or Nsurplus.

Calculated t	*NDrain*	*Nseep total*	*Ndenitrification*	*Nrecovery total*	*Nsurplus*	*Wsurplus*	*WDrainage*
Nseep total	17.471						
Ndenitrification	0.885	0.782					
Nrecovery total	5.125	4.081	-1.717				
Nsurplus	7.359	4.897	6.708	6.640			
Wsurplus	2.459	1.952	2.611	0.683	3.172		
WDrainage	15.422	14.957	0.391	7.421	6.846	1.426	
Wseep	13.481	13.571	0.009	7.353	6.090	0.085	56.037

Table 18: Correlation between phosphorus factors. Blue = high dependency.

R	*Wsurplus*	*WDrainage*	*Wseep*	*PDrain*	*Pseep total*	*Precovery total*
WDrainage	0.162					
Wseep	0.010	0.988				
PDrain	0.283	0.842	0.809			
Pseep total	0.195	0.857	0.838	0.698		
Precovery total	-0.394	0.025	0.086	-0.212	-0.045	
Psurplus	-0.220	-0.165	-0.133	-0.249	-0.109	0.905

Table 19: t-Value of the correlations. The t-value for the null hypothesis is 1.992. red = uncorrelated for PDrain or Psurplus.

calculated t	*Wsurplus*	*WDrainage*	*Wseep*	*PDrain*	*Pseep total*	*Precovery total*
WDrainage	1.426					
Wseep	0.085	56.037				
PDrain	2.555	13.494	11.923			
Pseep total	1.722	14.378	13.296	8.433		
Precovery total	-3.714	0.213	0.747	-1.882	-0.391	
Psurplus	-1.952	-1.447	-1.162	-2.222	-0.953	18.405

Sand

Table 20: Correlation between nitrogen factors. Blue = high dependency.

R	*NDrain*	*Nseep total*	*Ndenitrification*	*Nrecovery total*	*Nsurplus*	*Wsurplus*	*WDrainage*
Nseep total	0.535						
Ndenitrification	-0.201	-0.025					
Nrecovery total	0.117	0.090	-0.588				
Nsurplus	0.262	0.167	0.080	0.716			
Wsurplus	-0.279	-0.064	0.354	0.066	0.290		
WDrainage	0.713	0.792	-0.079	0.076	0.165	-0.201	
Wseep	0.723	0.780	-0.111	0.068	0.133	-0.292	0.996

Table 21: t-Value of the correlations. The t-value for the null hypothesis is 1.974. red = uncorrelated for NDrain or Nsurplus.

Calculated t	*NDrain*	*Nseep total*	*Ndenitrification*	*Nrecovery total*	*Nsurplus*	*Wsurplus*	*WDrainage*
Nseep total	8.353						
Ndenitrification	-2.707	-0.333					
Nrecovery total	1.547	1.187	-9.589				
Nsurplus	3.587	2.234	1.057	13.540			
Wsurplus	-3.828	-0.844	4.997	0.874	3.999		
WDrainage	13.418	17.119	-1.041	1.004	2.201	-2.708	
Wseep	13.801	16.415	-1.469	0.896	1.769	-4.028	140.123

Table 22: Correlation between phosphorus factors. Blue = high dependency.

R	*Wsurplus*	*WDrainage*	*Wseep*	*PDrain*	*Pseep total*	*Precovery total*
WDrainage	-0.201					
Wseep	-0.292	0.996				
PDrain	-0.030	0.867	0.849			
Pseep total	-0.066	0.820	0.807	0.756		
Precovery total	0.406	-0.073	-0.110	0.014	0.054	
Psurplus	0.441	-0.010	-0.052	0.119	0.099	0.986

Table 23: t-Value of the correlations. The t-value for the null hypothesis is 1.974. red = uncorrelated for PDrain or Psurplus.

calculated t	*Wsurplus*	*WDrainage*	*Wseep*	*PDrain*	*Pseep total*	*Precovery total*
WDrainage	-2.708					
Wseep	-4.028	140.123				
PDrain	-0.400	22.917	21.203			
Pseep total	-0.872	18.918	18.039	15.219		
Precovery total	5.853	-0.969	-1.465	0.182	0.712	
Psurplus	6.479	-0.136	-0.690	1.577	1.306	78.695

Annex 7: Calculated means in sand

To illustrate that the nominal values do differ from each other, the mean for each nominal values is calculated.

Table 24: Means: presence of drains.

Clay	drains=Yes	drains=No
Ndrain	17.131	18.590
Pdrain	2.066	2.342
Sand		
Ndrain	18.872	15.885
Pdrain	0.766	0.733

Table 25: Means: presence of irrigation.

Clay	Irrigation=Yes	Irrigation=No
Ndrain	17.245	10.813
Pdrain	1.144	2.190
Sand		
Ndrain	17.326	16.073
Pdrain	0.759	0.553

Table 26: Means: GT-class.

Clay	dry	middle	wet
Ndrain	12.642	14.695	20.406
Pdrain	1.586	2.010	2.443
Sand			
Ndrain	14.874	16.544	16.825
Pdrain	0.513	0.564	1.178

Table 27: Means: soil physical units.

Clay	A	B	C	D	E	
Ndrain	16.741	14.869	17.243	27.159	9.984	
Pdrain	2.034	1.985	2.582	2.724	1.566	
Sand	F	G	H	I	J	K
Ndrain	14.722	19.209	17.318	16.349	11.097	12.030
Pdrain	1.199	1.508	0.640	0.748	0.380	0.842

Annex 8: Results extra tests.

Extra tests, with datasets that consist of data with a specific characteristic, are done. These tests are compared with the tests that take into account the whole dataset.

Table 28: The parameters from the different obtained equations. The columns represent the different tests in Weka. The rows represent the different variables that are taken into account. Furthermore, the accuracy indicators of each formula are given.

	Characteristic tests (filtered dataset)							Normal test (whole dataset)	
TEST Ndrain=	NoDrains (nominal)	NoDrains (non-nominal)	F, H, I (nominal)	F, H, I (non-nominal)	NoDrains, no irrigation, F, H, I	Dry soil (nominal)	Dry soil (non-nominal)	nominal	non-nominal
Nsurplus	0.0464	0.1056	0.1307	0.1147	0.1154		0.1517	0.0490	0.1302
Nrecovery		-0.0330	-0.0645	-0.0540			-0.0877		-0.0501
Wsurplus		0.0580	-0.0666	-0.0818	-0.0574				-0.0683
Wseep	0.0262	0.0232	0.0252	0.0232	0.0247	0.0284	0.0294	0.0262	0.0225
soil physical unit=F, I, H	3.6215					5.3867		3.6032	
Soil physical unit=G	0.6750					5.3867		0.3608	
GT-class=wet	-1.1889		-1.8435		-1.5625			-1.4047	
Drains=Yes			2.4336					2.4902	
Irrigation=Yes	2.2255		3.2088			3.2561		2.5705	
β	10.2504	26.6920	28.4813	34.7346	26.6641	13.6780	5.3601	10.0471	28.1460
R	0.766	0.773	0.745	0.718	0.751	0.447	0.649	0.745	0.746
RMSE	4.133	3.184	4.123	4.290	4.068	5.912	4.625	4.278	4.248
RRSE	64.54%	13.19%	67.16%	61.70%	65.71%	96.47%	75.47%	66.65%	65.97%
n	156	156	138	138	113	44	44	175	175

Annex 9: Sample size

Rules of thumb are used to check if the amount of data used for the regression analysis are sufficient.

Table 29: Rules of thumb: size of data set.

Test		N=10k-15k	N = 100 + k
Nominal and numerical (Nominal test)	k=9	90 - 135	109
Only numerical (Non-nominal test)	k=5	50 - 75	105

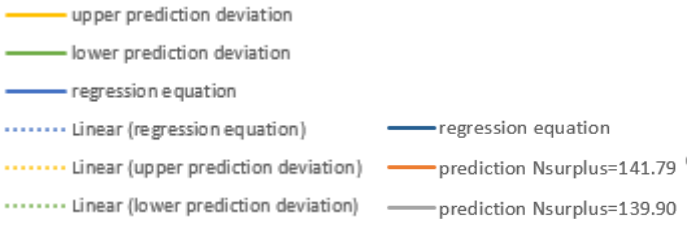
Annex 10: results

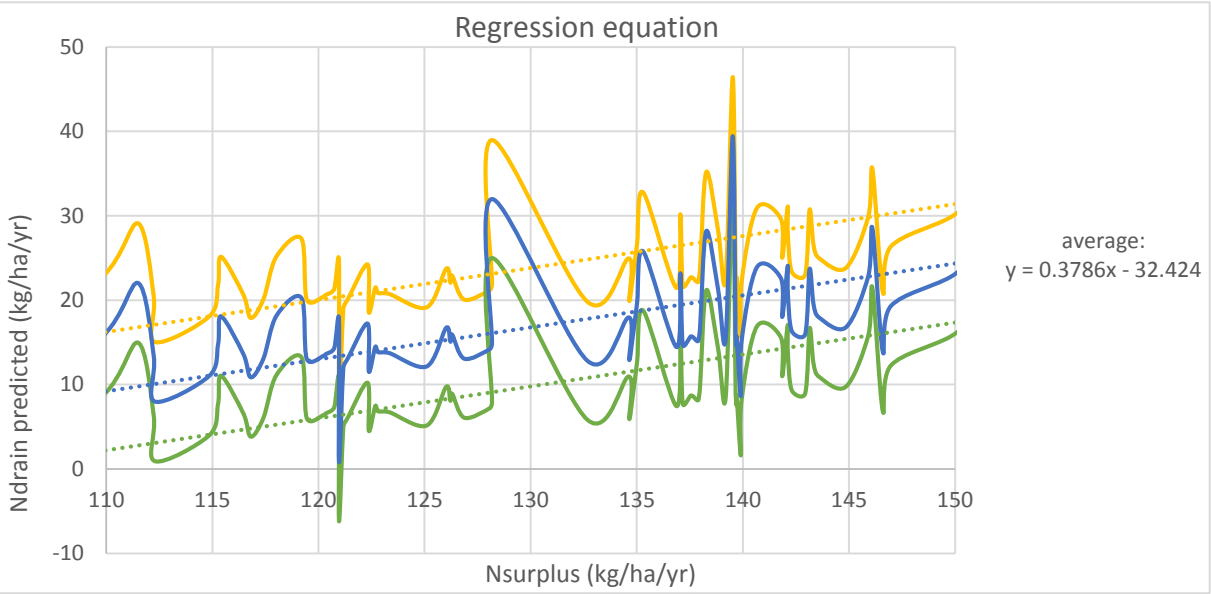
The results consist of the regression equation with its indicators: R, RMSE and RRSE. In the top right graph, the relation between the nutrient surplus and the nutrient drainage as is predicted by the regression equation is shown. In addition, this graph shows the accompanying prediction interval of the equation and the trend line. On the bottom right the predicted nutrient drainage is plotted against the real drainage from STONE. The bottom left represents a case study for a single farm or data set. In the case study, the effect of changing the nutrient surplus, while the other factors are assumed to be constant, is illustrated with a prediction line.

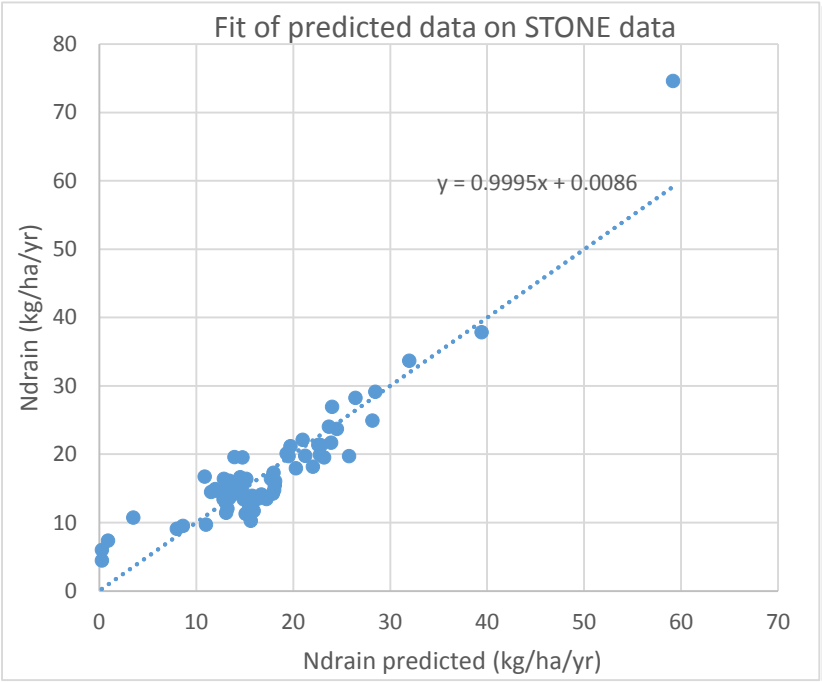
Annex 10.1: Nitrogen clay nominal

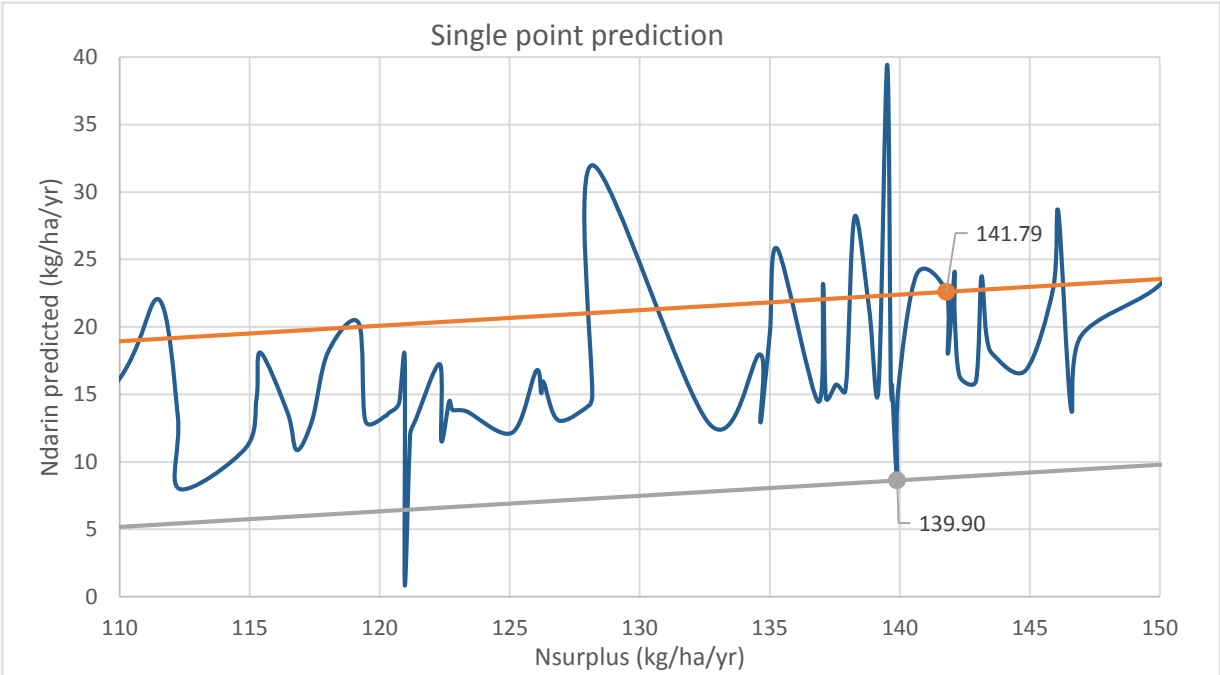
Regression equation:
Ndrain = 0.1152Nsurplus + 0.0507Wseep + 6.9316 if soil physical unit=D - 5.754 if Gt-class=Middle or Wet + 3.9555.
N = 77
R = 0.8175
RMSE = 5.0343
RRSE = 57.34%

Legend:





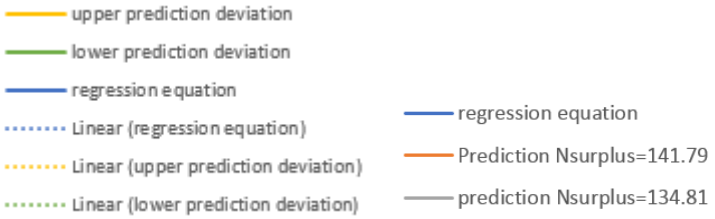


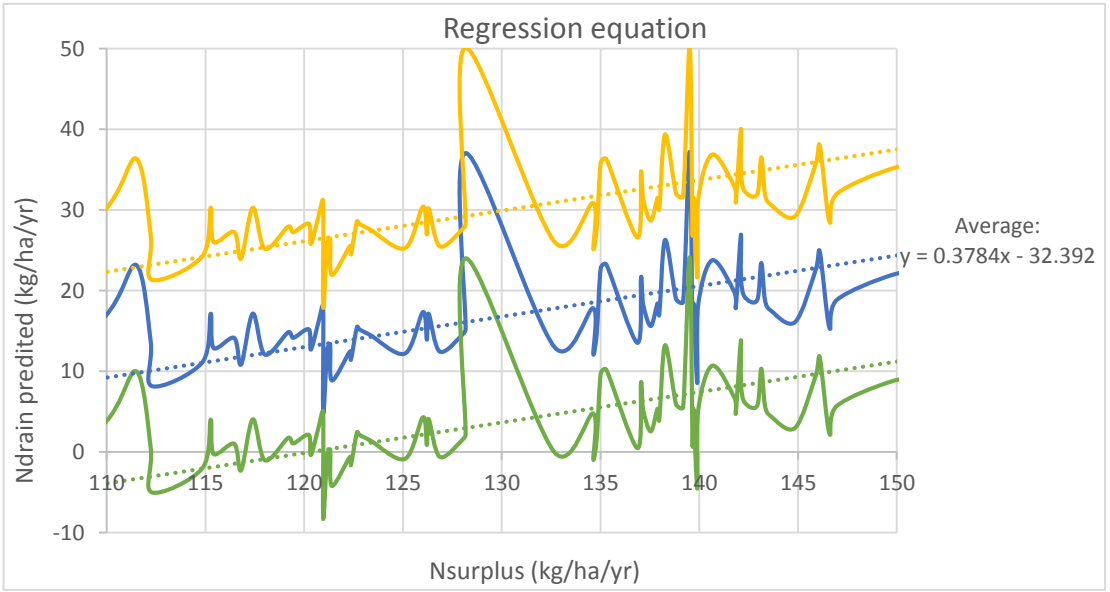


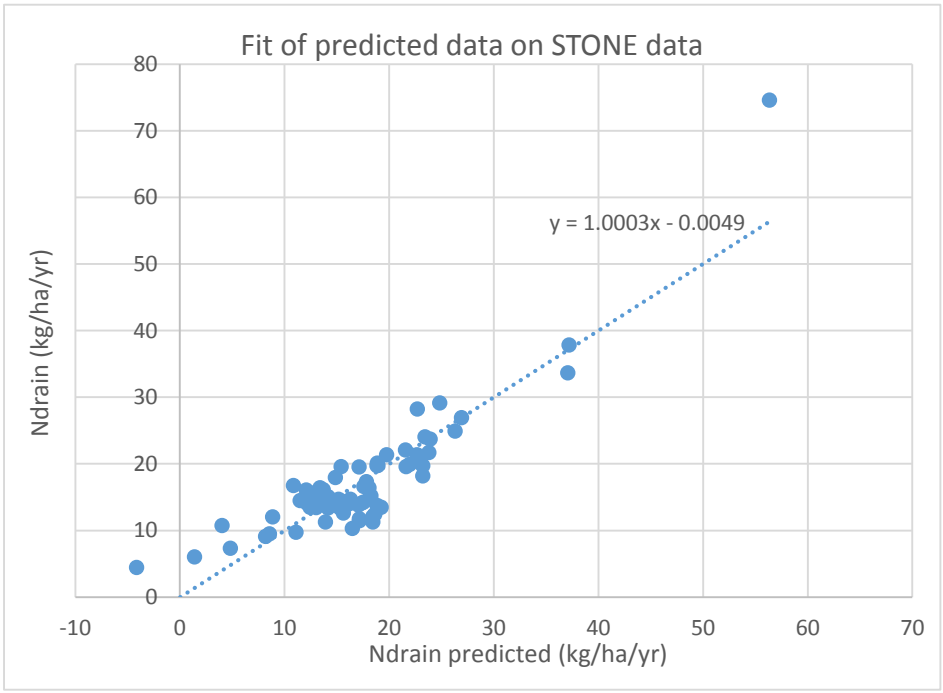
Annex 10.2: Nitrogen clay non-nominal

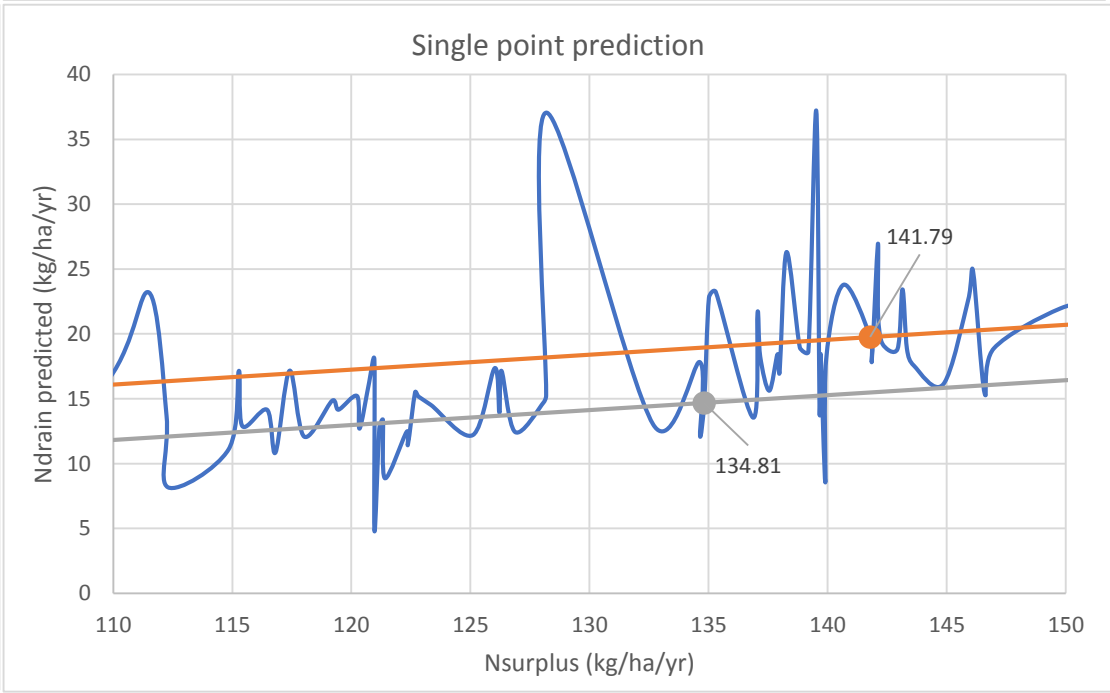
Regression equation:
Ndrain = 0.1152Nsurplus + 0.0569Wseep + 0.0927Wsurplus
- 0.1597Nrecovery - 31.1868.
N = 77
R = 0.838
RMSE = 3.0912
RRSE = 53.98%

Legend:





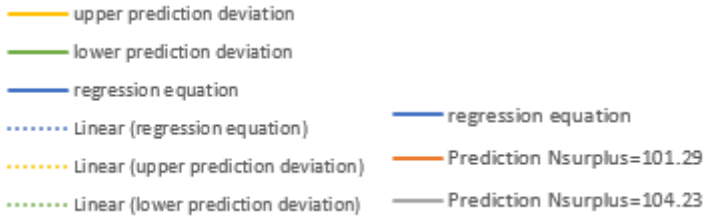


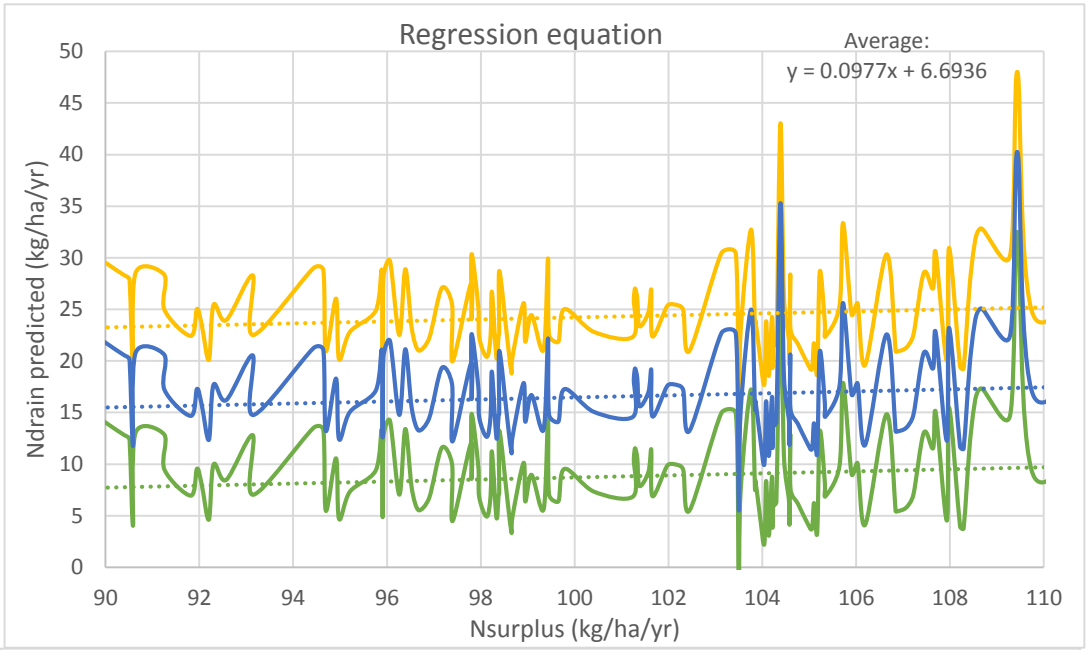


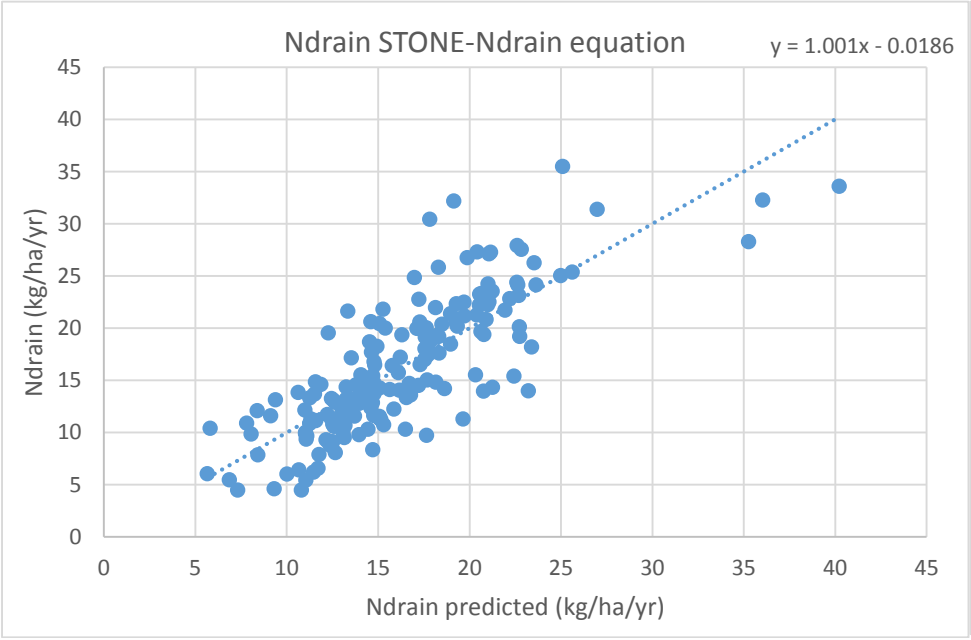
Annex 10.3: Nitrogen Sand nominal

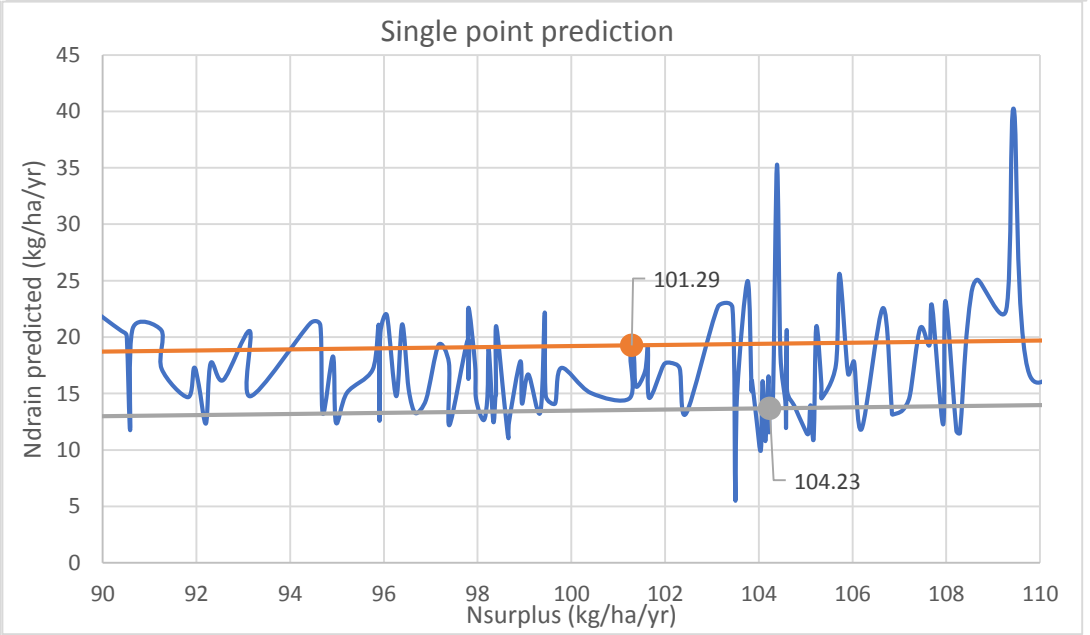
Regression equation:
Ndrain = 0.049Nsurplus + 0.0262Wseep + 3.6032 if soil physical unit=F,H or I + 0.3608 if soil physical unit=G – 1.4047 if GT-class= Wet + 2.4902 if Drains=Yes + 2.5705 if irrigation=yes + 10.0471
N = 175
R = 0.7453
RMSE = 4.2782
RRSE = 66.65%

Legend:





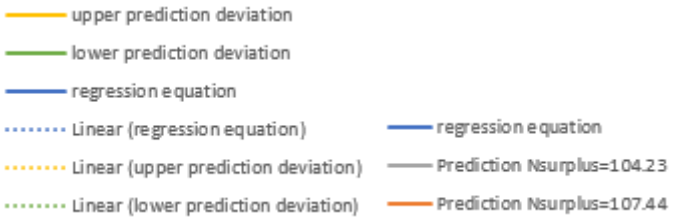


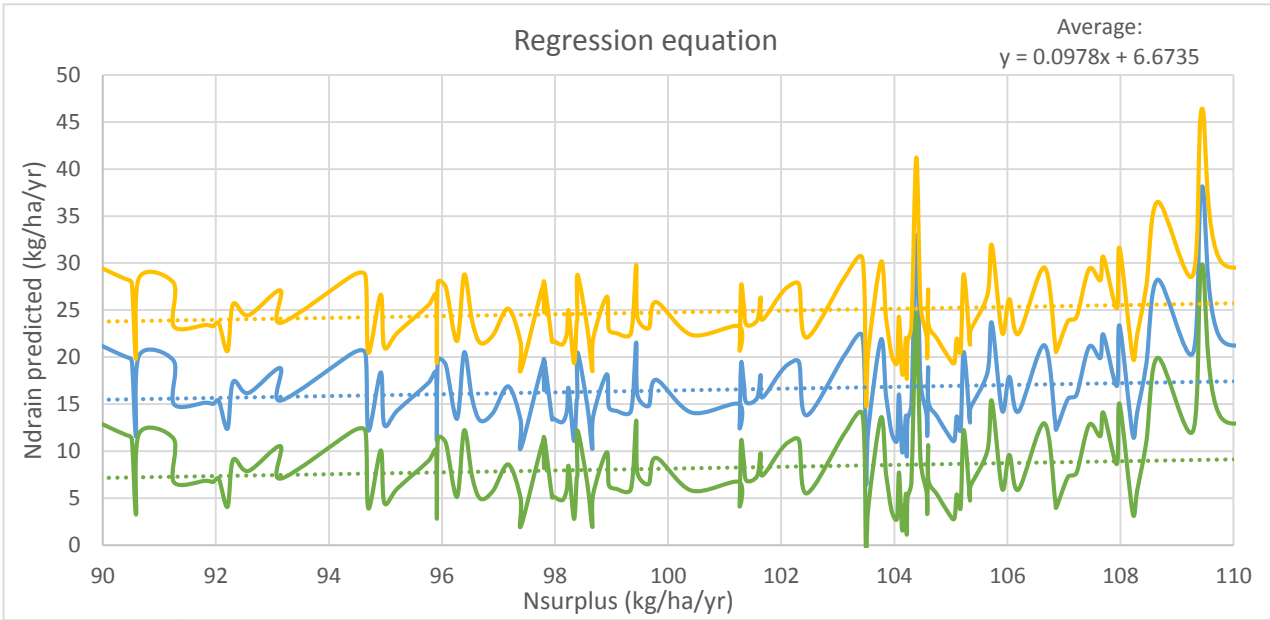


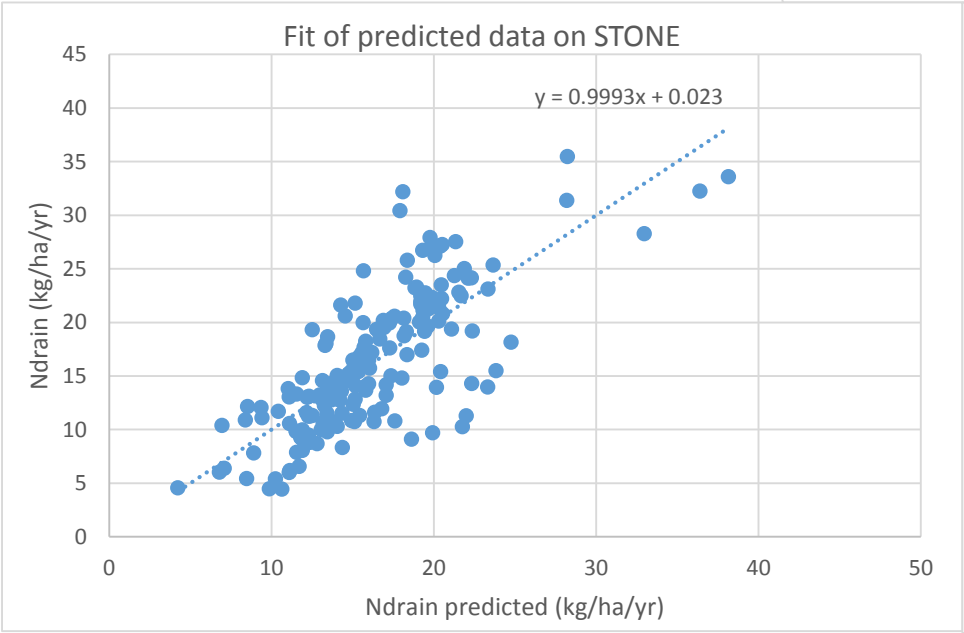
Annex 10.4: Nitrogen sand non-nominal

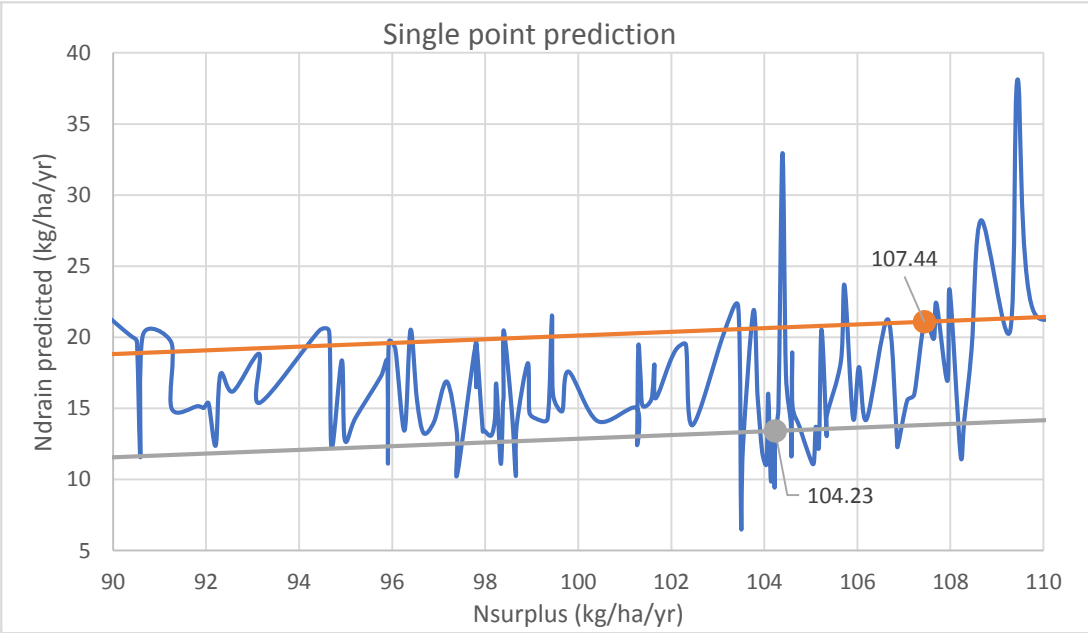
Equation:
Ndrain = 0.1302Nsurplus + 0.0225Wseep – 0.0683Wsurplus - 0.0501Nrecovery + 28.146.
N = 175
R = 0.7464
RMSE = 4.2484
RRSE = 65.97%

Legend:





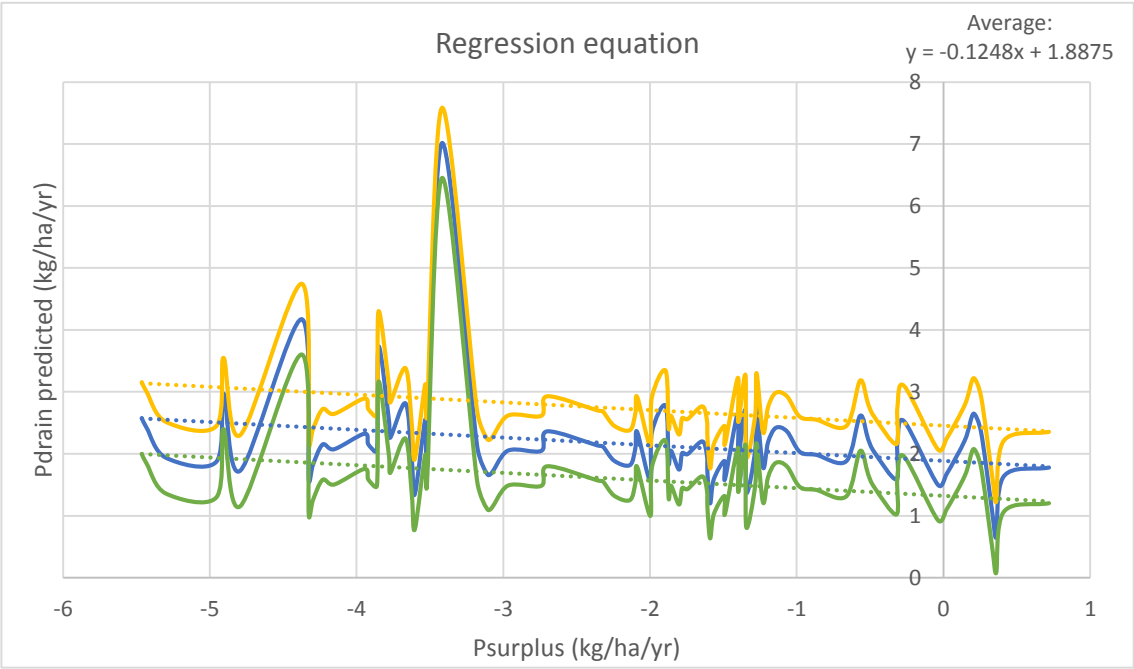


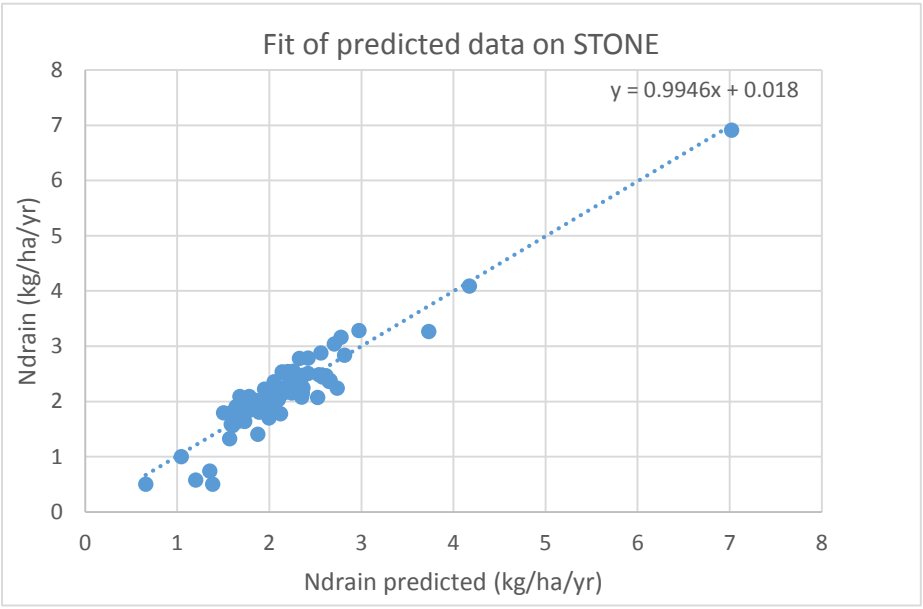


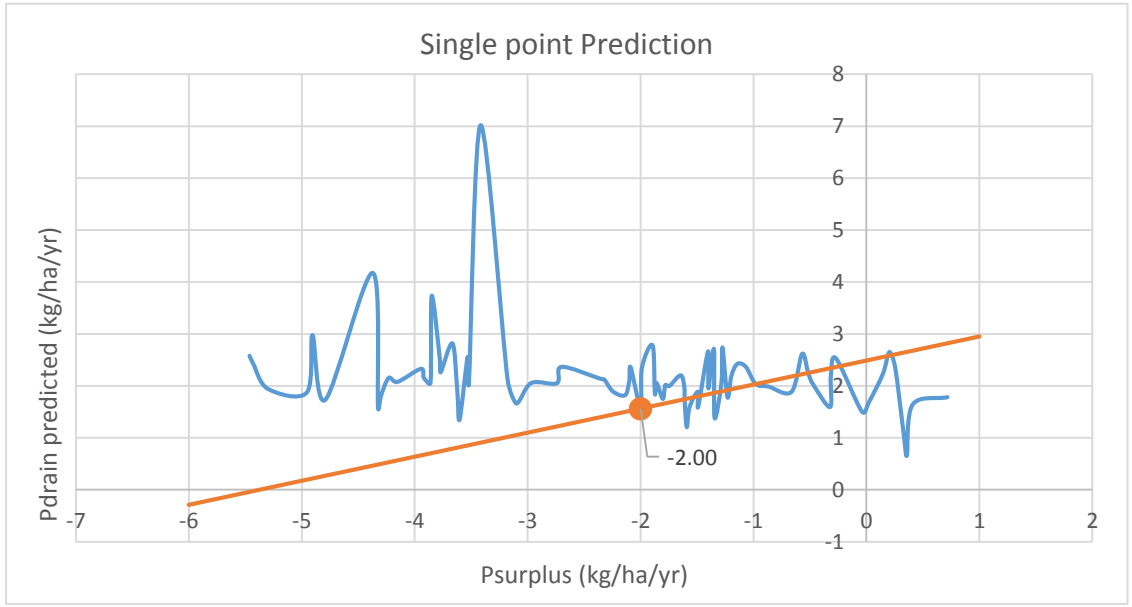
Annex 10.5: Phosphorus clay nominal

Equation:
Pdrain = 0.4634Psurplus + 0.0066Wseep – 0.004Wsurplus
- 0.5139Precovery + 0.2297 if soil physical unit=A or B + 0.7445 if soil physical unit=C + 0.284 if soil physical unit=D – 0.2969 if GT-class=wet + 0.376 if Drains=no +1.7009
N = 77
R = 0.8745
RMSE = 0.3986
RRSE = 48.74%

Legend:
upper prediction deviation
lower prediction deviation
regression equation
Linear (regression equation)
Linear (upper prediction deviation)
Linear (lower prediction deviation)
Prediction line
Prediction Psurplus=-1.998



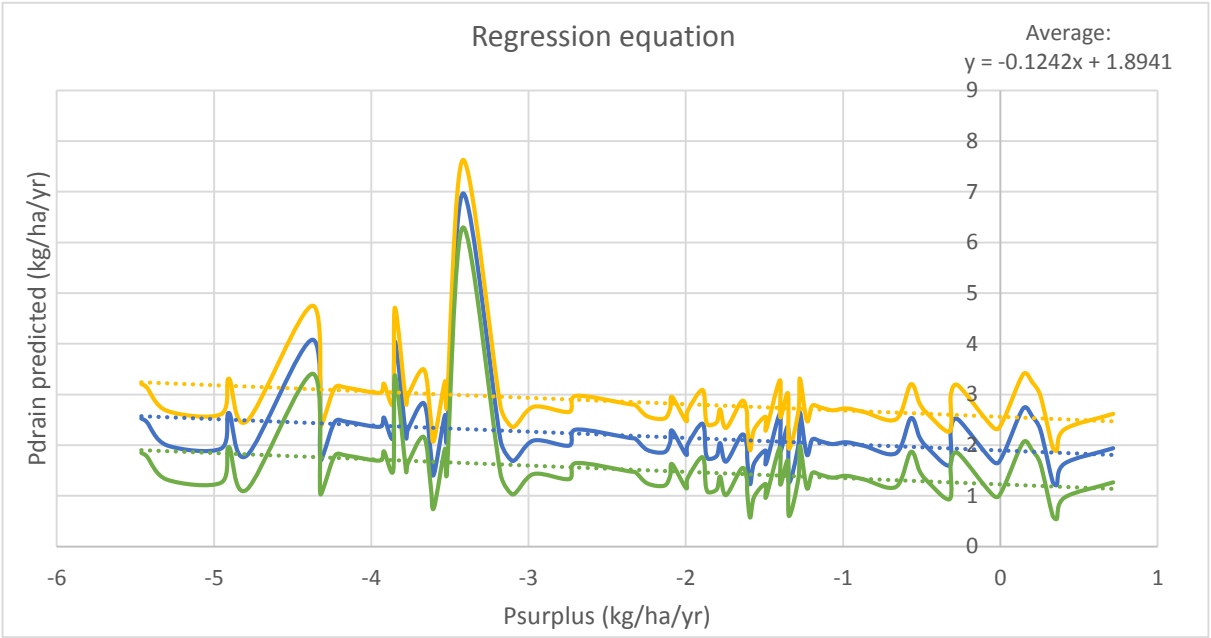


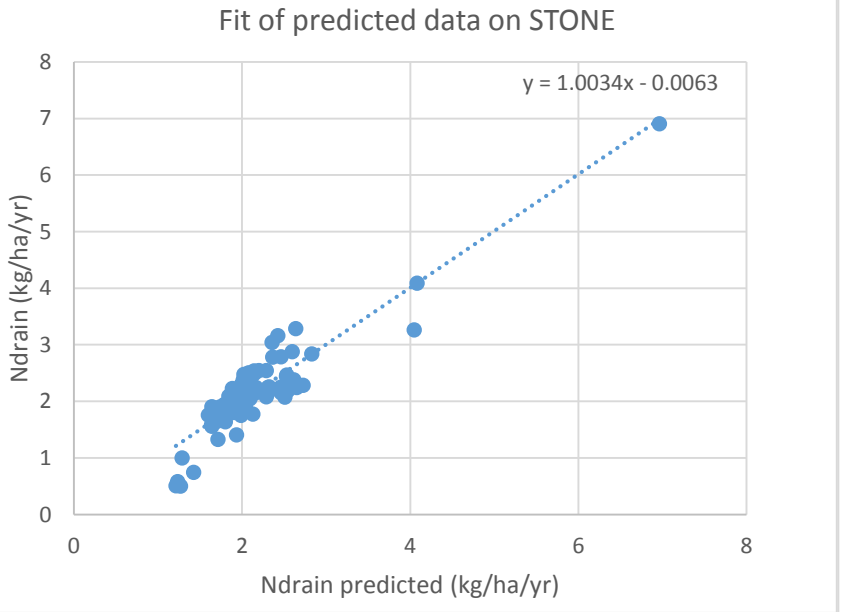


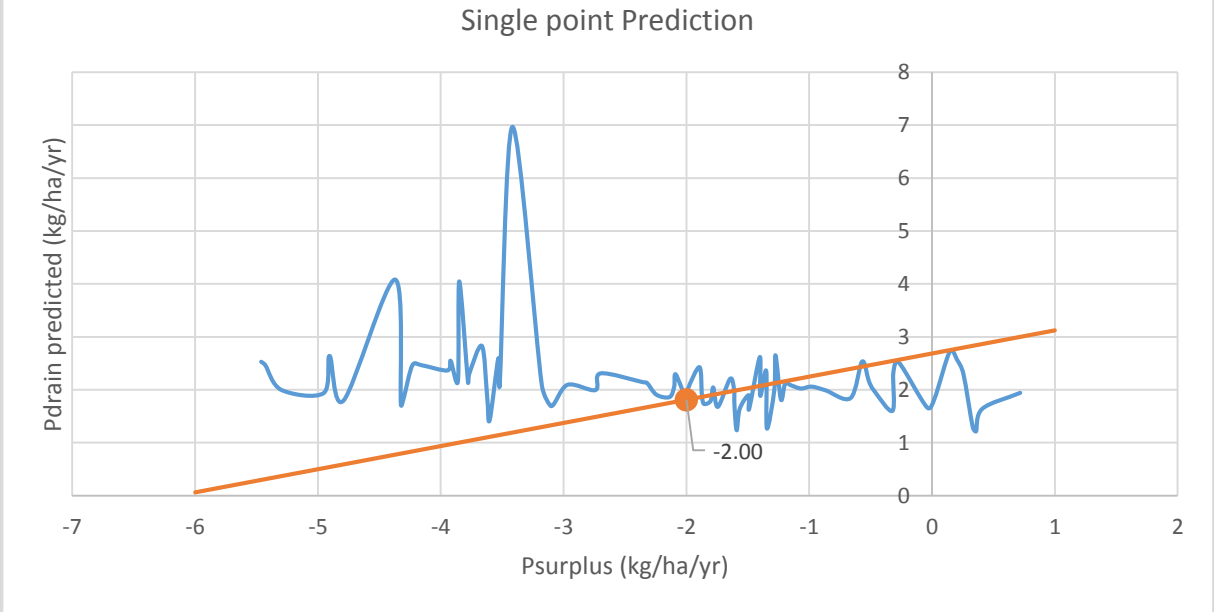
Annex 10.6: Phosphorus clay non-nominal

Equation:
Pdrain = 0.4373Psurplus + 0.0064Wseep - 0.4934Precovery + 0.6649.
N = 77
R = 0.8972
RMSE = 0.3538
RRSE = 43.27%

Legend:
upper prediction deviation
lower prediction deviation
regression equation
Linear (regression equation)
Linear (upper prediction deviation)
Linear (lower prediction deviation)
Prediction line
Prediction Psurplus=-1.998



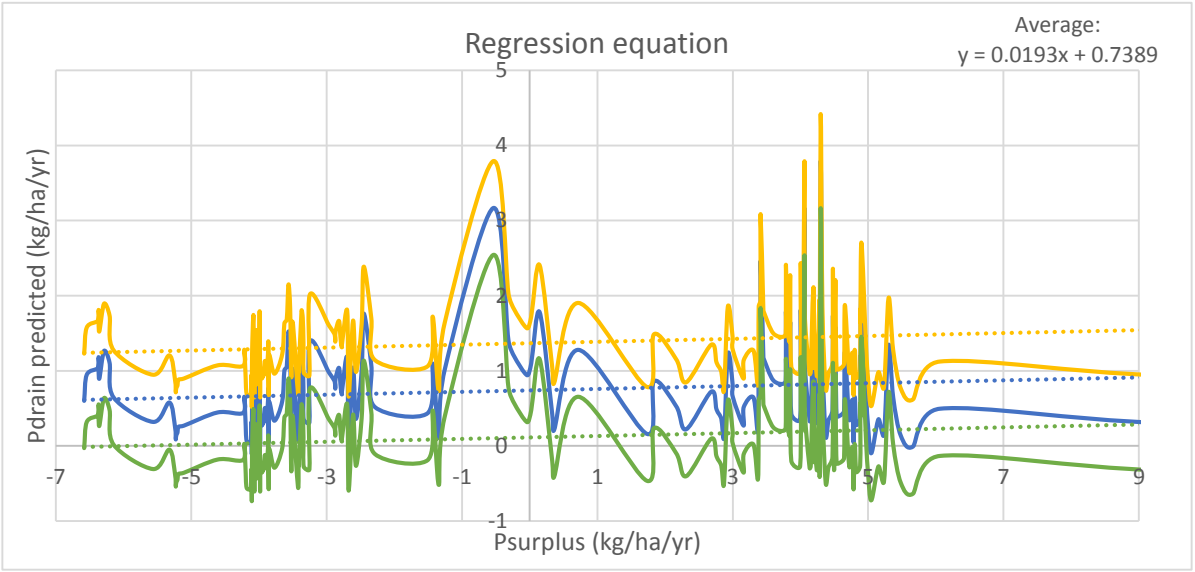


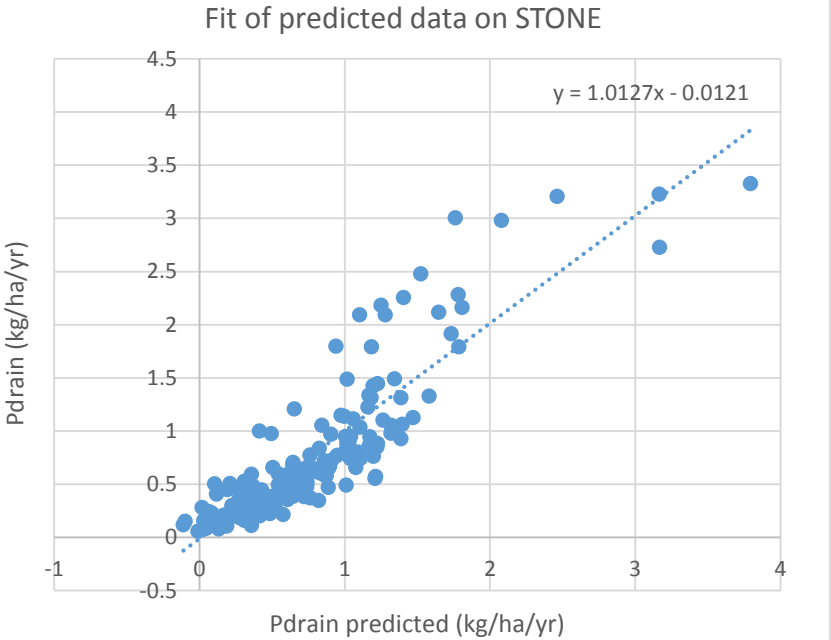


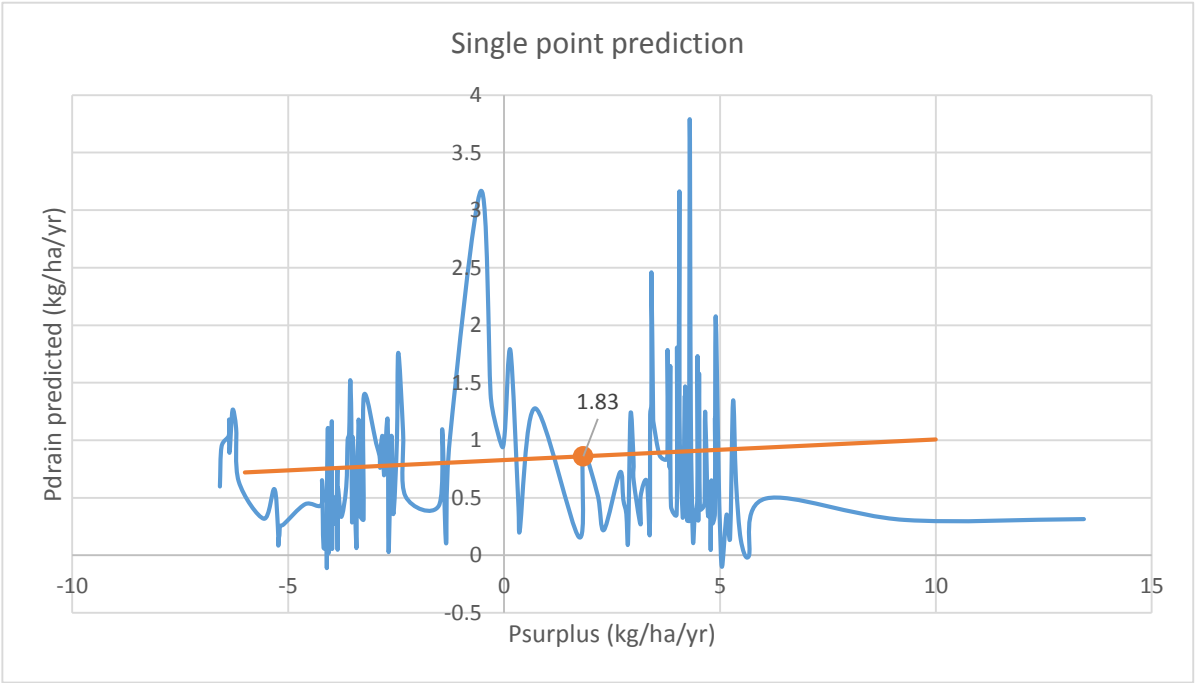
Annex 10.7: Phosphorus sand nominal

Equation:
Pdrain = 0.0178Psurplus + 0.003Wseep – 0.0862 if GT-class=middle + 0.2395 if GT-class=wet + 0.8872.
N = 175
R = 0.8696
RMSE = 0.3345
RRSE = 49.36%

Legend:
- upper prediction deviation
- lower prediction deviation
- regression equation
- Linear (regression equation)
- Linear (upper prediction deviation)
- Linear (lower prediction deviation)
- Prediction line
- prediction Psurplus=1.83



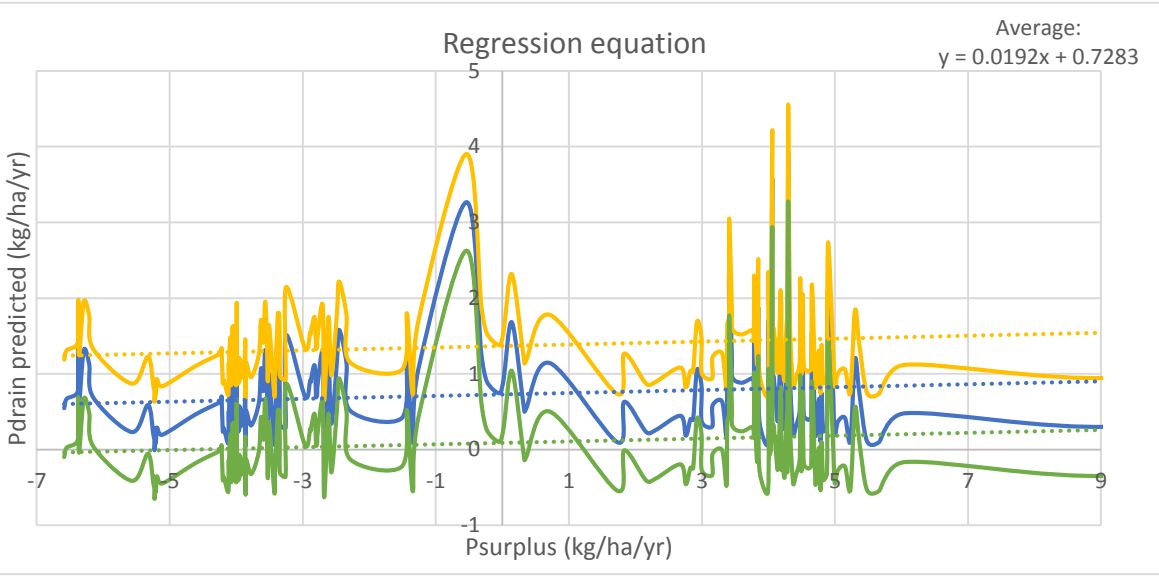


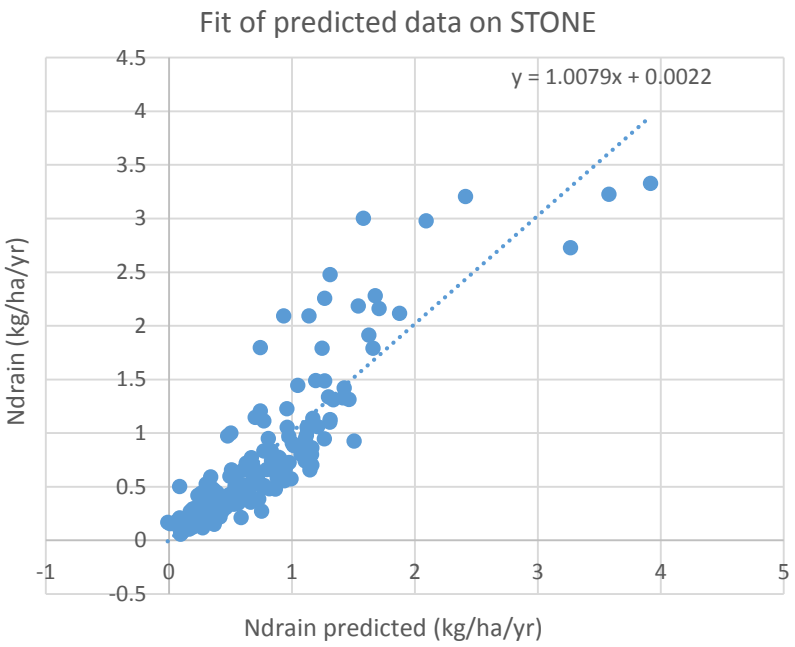


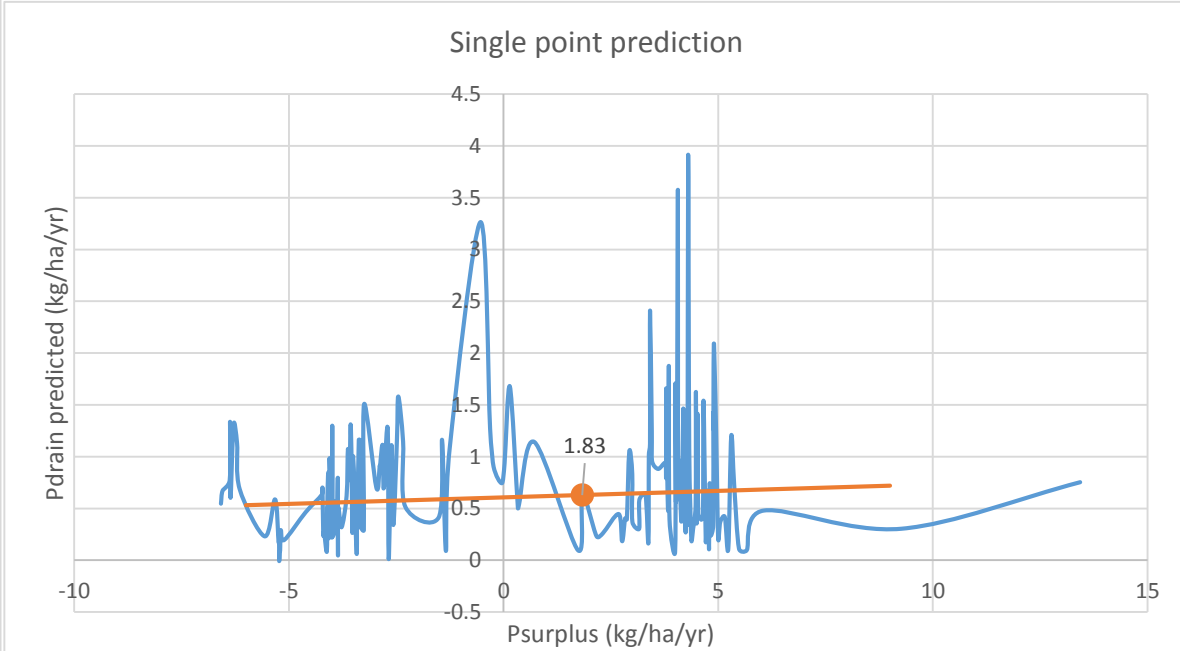
Annex 10.8: Phosphorus sand non-nominal

Equation:
Pdrain = 0.0125Psurplus + 0.0034Wseep + 0.0079 Wsurplus – 1.6541.
N = 175
R = 0.8737
RMSE = 0.3288
RRSE = 48.29%

Legend:
upper prediction deviation
lower prediction deviation
regression equation
Linear (regression equation)
Linear (upper prediction deviation)
Linear (lower prediction deviation)
Prediction line
prediction Psurplus=1.83







Annex 11: residual plots

To test the linearity of the regression equations, residual plots are made.
Residual = Yreal - Ypredicted.

Nitrogen

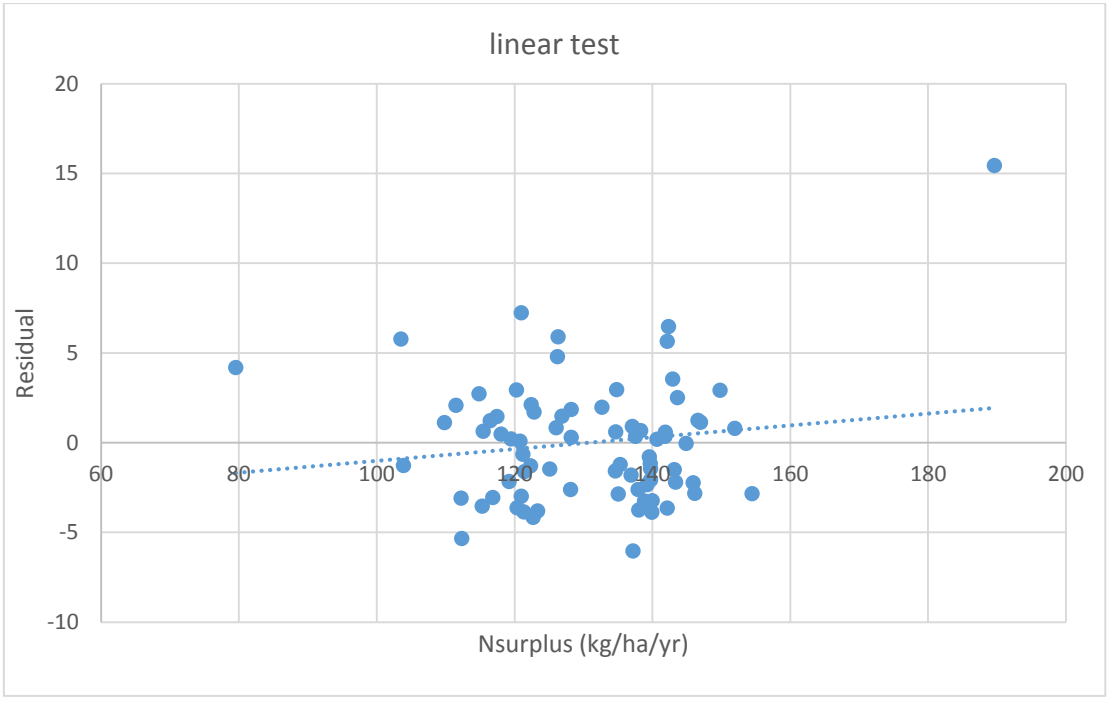


Figure 20: Linear test of the nominal equation of clay.

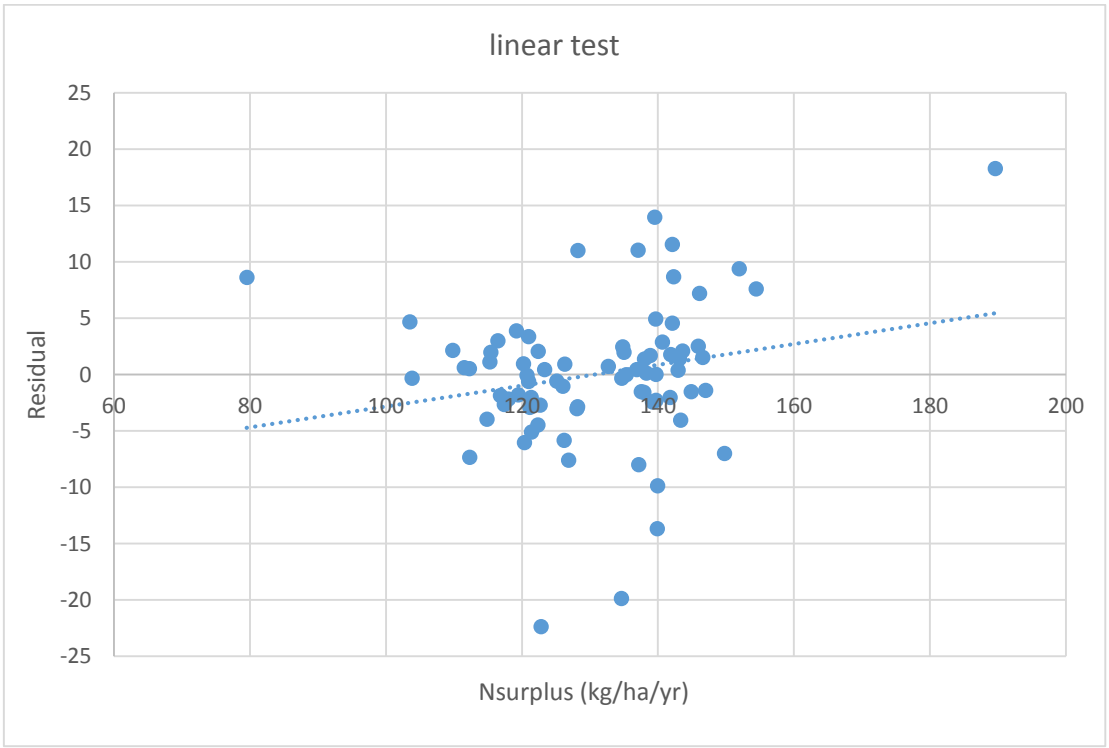


Figure 21: Linear test of the non-nominal equation of clay.

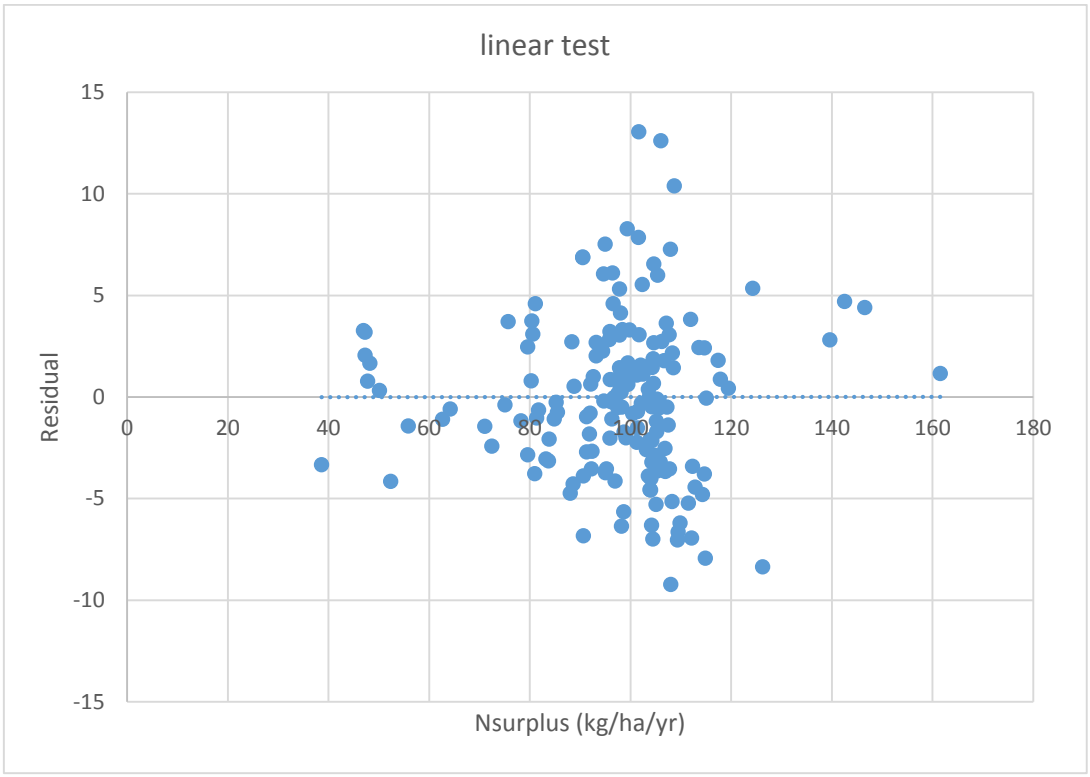


Figure 22: Linear test of the nominal equation of sand.

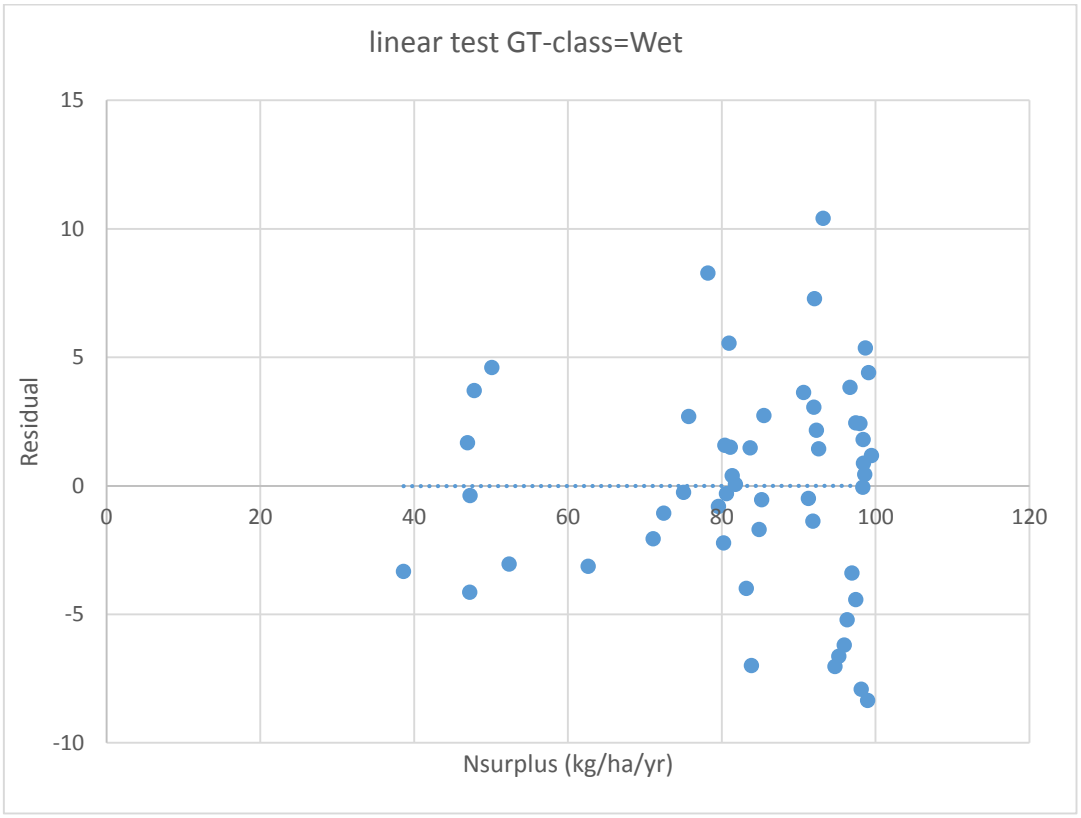


Figure 23: Linear test of the nominal equation of sand, with characteristic GT-class=Wet.

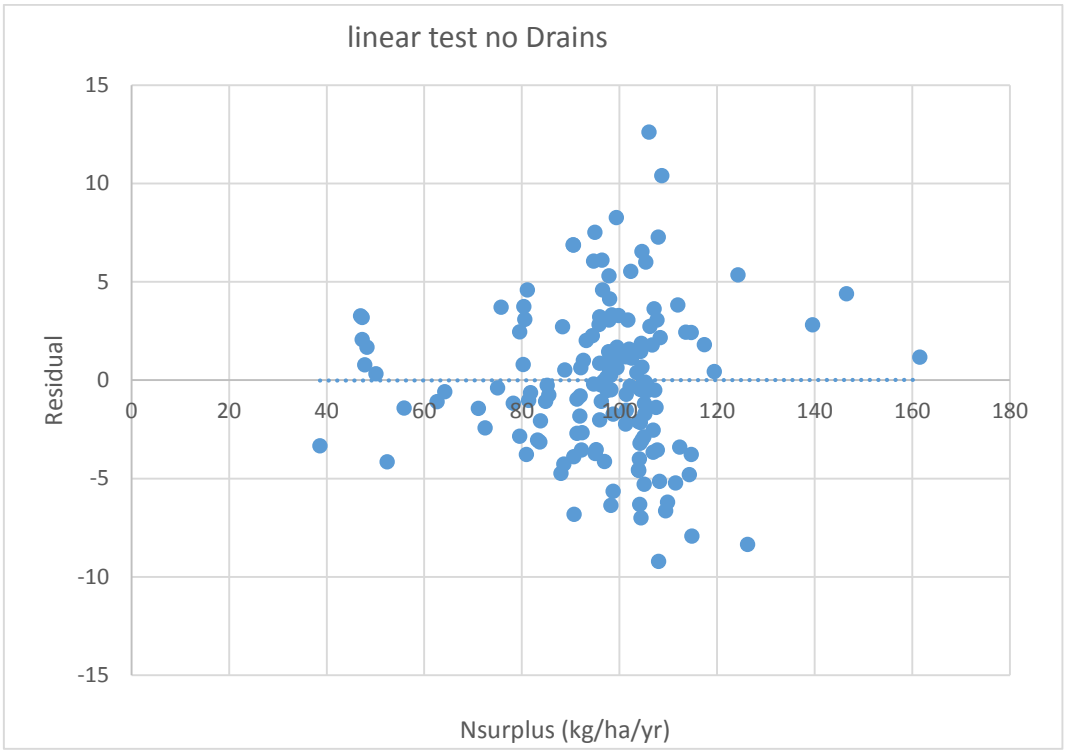


Figure 24: Linear test of the nominal equation of sand, with characteristic no drains present.

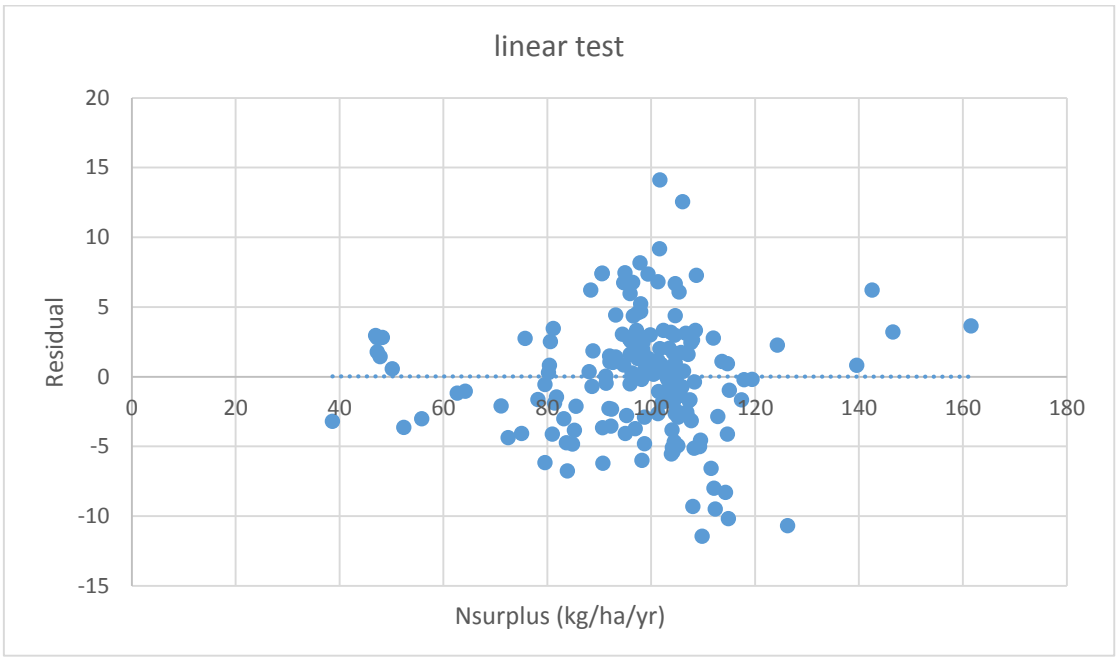


Figure 25: Linear test of the non-nominal equation of sand.

Phosphorus

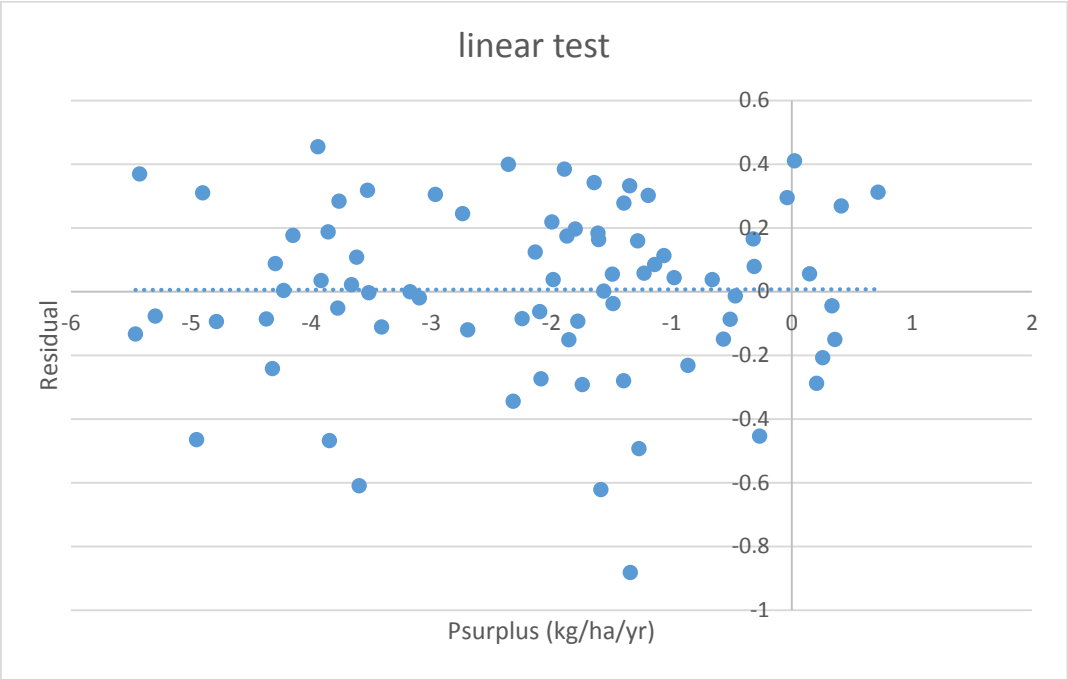


Figure 26: Linear test of the nominal equation of clay.

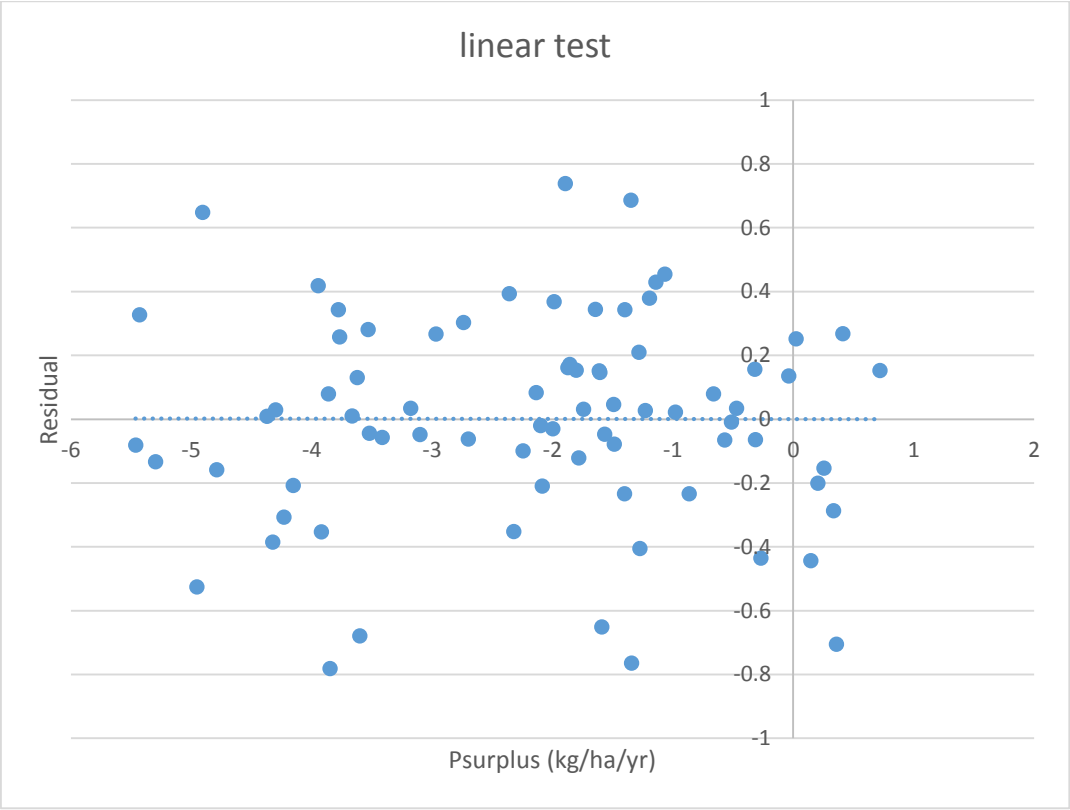


Figure 27: Linear test of the non-nominal equation of clay.

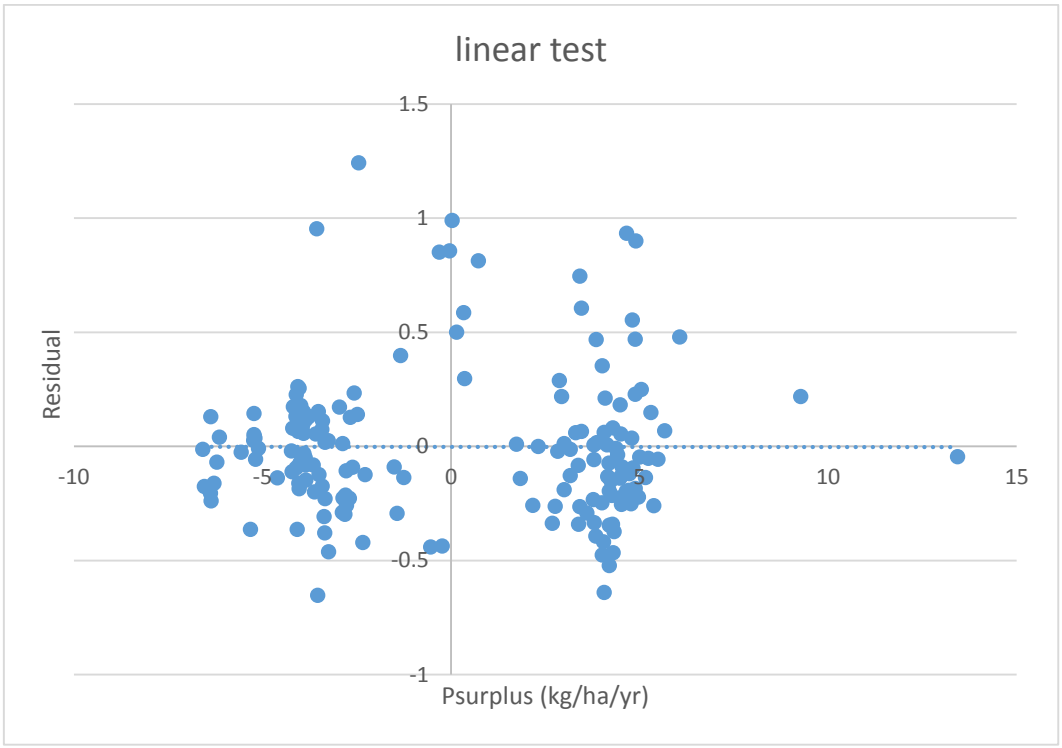


Figure 28: Linear test of the nominal equation of sand.

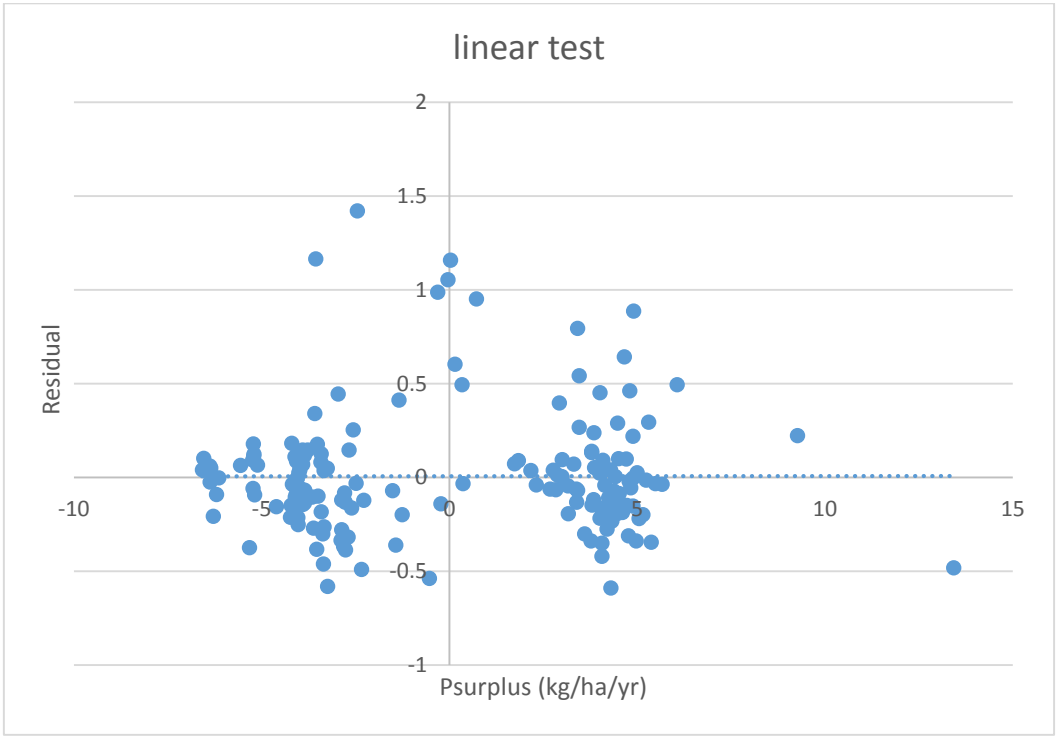


Figure 29: Linear test of the non-nominal equation of sand.

Annex 12: the effect of extreme values

This attachment shows the whole regression equation graphs for nitrogen. In the graphs, single point predictions are done. As is shown, the predictions made are within the boundaries of the graph for a certain interval. Outside this interval no clear boundaries are shown, due to the low amount of data points. Predictions made with starting points outside this interval (extreme values) or predictions with a reduction of the nutrient surplus >10% are therefore assumed to be invalid.

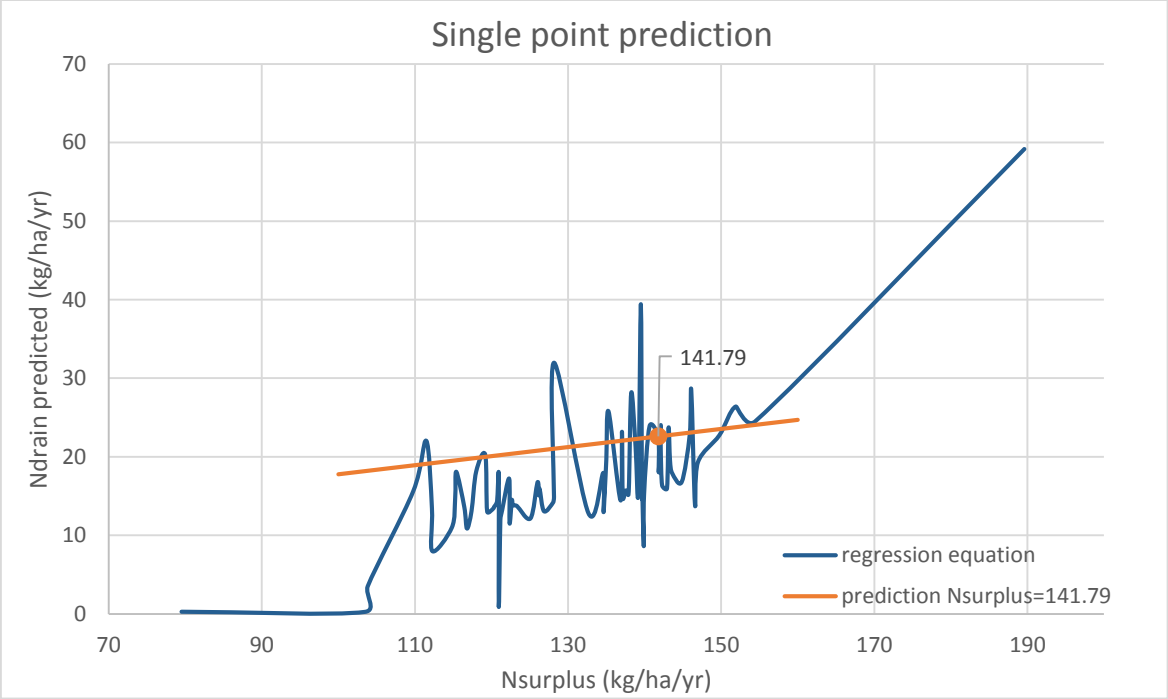


Figure 30: Single point prediction and the whole regression equation for clay.

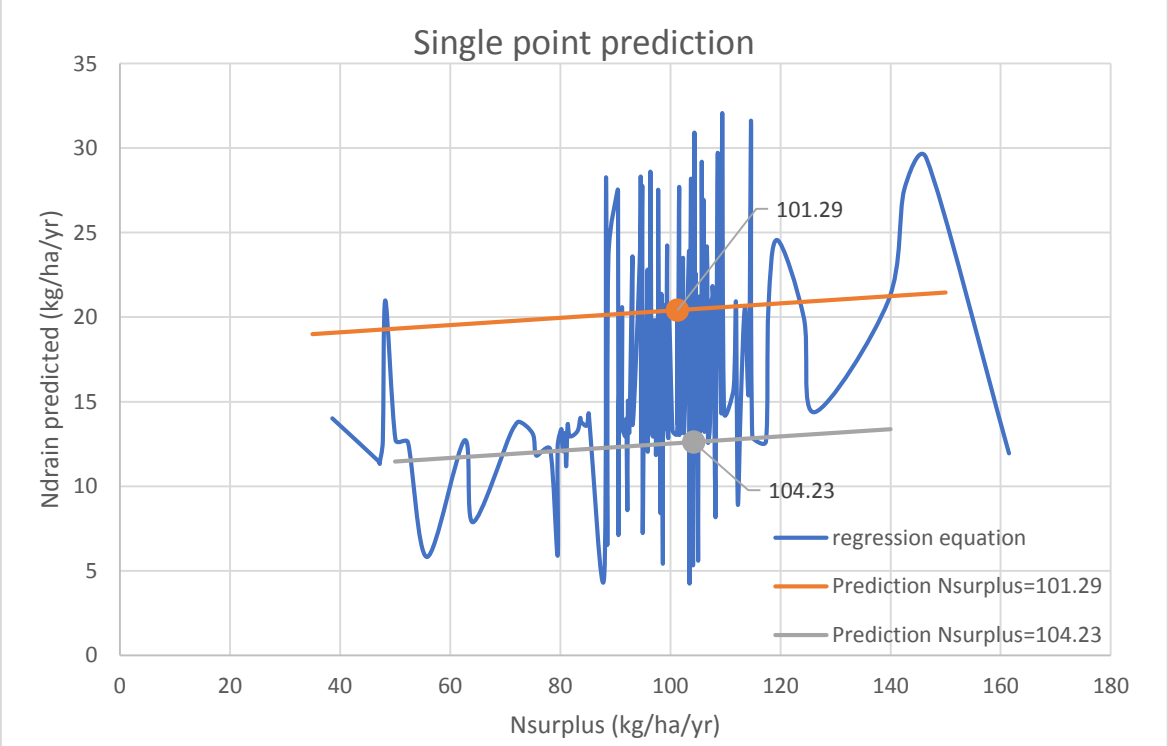


Figure 31: Single point prediction and the whole regression equation for sand.