Regression Kriging of nitrate levels in upper groundwater in Dutch sandy soils

An analysis at national and regional extents

GIMA Thesis report 2014-11/2014

Cor J. de Jong

Colofon

Cor J. de Jong
UU student number 3336034

Thesis supervisor dr.ir. G.B.M. Heuvelink, associate professor WUR
GIMA Professor prof.dr.ir. A.K. Bregt, WUR

Contact:

Cor de Jong
Email: cor.de.jong@rivm.nl

This thesis research has been conducted in partial fulfilment of the requirements for the MSc-GIMA programme

Abstract

Regression kriging of nitrate levels in upper groundwater in Dutch sandy soils
An analysis at national and regional extents

Sample data concerning nitrate concentrations in groundwater, were collected on farmland and in nature reserves on sandy soils in The Netherlands. Using Regression Kriging modelling, a geostatistical approach that exploits both the spatial variation in the sampled variable itself, and environmental information collected from covariate maps for the target predictor, it is possible to predict groundwater quality maps for the sandy soil regions in The Netherlands, and quantify the uncertainty in accompanying maps.
Maps were produced for four different sandy soil regions and three different years in 2007, 2008 and 2009. For a combination of the regions into a nationwide model for the three years maps were made as well. The most successful covariate to be found in the regression part was the groundwater table map. The differences between a regional approach and a combined nationwide approach were explored. The nationwide approach seemed to generate slightly better predictions in a more stable manner, although differences are not very pronounced.

Keywords:
Regression Kriging, geostatistics, interpolation, covariates, nitrate in groundwater, sandy soils

Acknowledgements
Thanks to my colleagues from RIVM for enabling this thesis research project. Dico Fraters for sparing me time in the already busy project itinerary, and for being patient. Arnold Dekkers for advice and assistance with the (geo-)statistics and adding a little boost in the progress occasionally. Finally, I want to express my gratitude to my supervisor Gerard Heuvelink for guiding this thesis and for always remaining optimistic and patient during the time it took.

Contents

1 Introduction

1.1 Problem definition and context

In the Netherlands, as well as in other countries, governmental agencies deploy large-scale measurement campaigns in order to investigate and evaluate environmental quality. This thesis focuses primarily on a method to predict groundwater quality based on samples from these campaigns, together with mapped variables to ultimately yield maps with nitrate content in groundwater.

Decades ago, emission of pollutants into the air, soil and water occurred widespread and was common practice. Awareness of the severe negative effects of this practice on the environment have led to governments starting up large scale environmental research and set-up of measurement campaigns. Air quality was one of the first fields of interest, soon followed by soil and water investigation. Primarily, the focus was on human health effects, but soon realization came that whole ecosystems were threatened by deteriorating quality of biotic and abiotic components and long lasting effects of certain changes and the accumulation of often persistent substances.
On the countryside, exceeding intensification of agricultural land use, industrial emissions of pollutants and the widespread use of fertilizers have led to deteriorated environments. In the Netherlands, where both agriculture and further industrialization were stimulated after the second world war, farm intensity was increased and for instance cattle breeding and dairy farming were successfully growing into a major export producing industry. Regarding water quality, the effect was most apparent on surface water. Algal bloom, loss of species diversity and finally, imbalanced ecosystems are but a few examples of these effects. Since groundwater is the main source for fresh water, water companies have to increase their efforts to make groundwater suitable for drinking water (Pebesma, 1996). Protection is therefore necessary. Nitrate in drinking water is considered as a contaminant in large quantities and therefore a WHO health standard is set at 50 mg/l (WHO, 1998). To stop increasing nutrient levels, legislation was enforced in order to reduce input of artificial fertilizers and animal manure, thus decreasing their negative effects on the environment. To ensure that this new legislation would improve environmental quality, monitoring of the current state and future developments was necessary. To register and follow environmental quality in time, several monitoring programs were put in place. In the Netherlands, but equally in neighbouring countries having similar intensified agriculture, like Denmark, groundwater is monitored in yearly measuring campaigns. In the Netherlands, however, groundwater is easily accessible at most locations at a depth of 1-6 meters, due to the small differences in elevation as well as the almost complete absence of rocky and impenetrable soil layers. Therefore, in the Netherlands, samples can simply be taken from the groundwater layer close to the soil surface. In other countries deeper groundwater layers have to be sampled or stream-area monitoring takes place (Fraters et al, 2005).

Measuring the actual quality in situ is a proven technology. Sampling and measuring involve large budgets for travel, measuring and sampling equipment and finally laboratory analyses and still not every desired location can be sampled. Considering a certain variability that seems to differ per location and, for instance, land use or soil type, it becomes clear that modelling could provide further information, based on data gathered thus far (Reijnders et al, 2004). It is therefore necessary to establish relationships with easily available environmental factors, like soil type, precipitation, land use etcetera. Variations in local quality levels indicated that a relation existed between soil variables and human induced environmental factors such as the type of land use and the intensity of that land use (CCRX, 1995); (RIVM, 2002).

When modelling groundwater quality, the spatial and temporal variation of groundwater quality variables are summarized with a limited set of mathematical equations (Pebesma, 1996). This will, because of the large complexity, always be a simplification of reality.

At RIVM, the Dutch National Institute for Public Health and the Environment, data are available from monitoring networks intended for measuring the effect of national emission policies. These data are gathered in various compartments like air, soil, and groundwater. The Minerals Policy Monitoring Network (LMM) was founded in cooperation with the Agricultural Economics Institute (LEI-WUR), in order to deliver information on the effectiveness of policy with regard to reduction of fertilizers and manure application. From this network, samples are collected at farms in the upper meter of groundwater. This upper meter is considered as being below the root zone and thus 'out of reach' for crops and other vegetation. In the field, nitrate concentrations in each of the individual samples are determined with a quick field check method. The illustration (Figure 1) shows all samples that were collected between 2007 and 2009. Some locations were visited repeatedly during this timeframe, others only once.

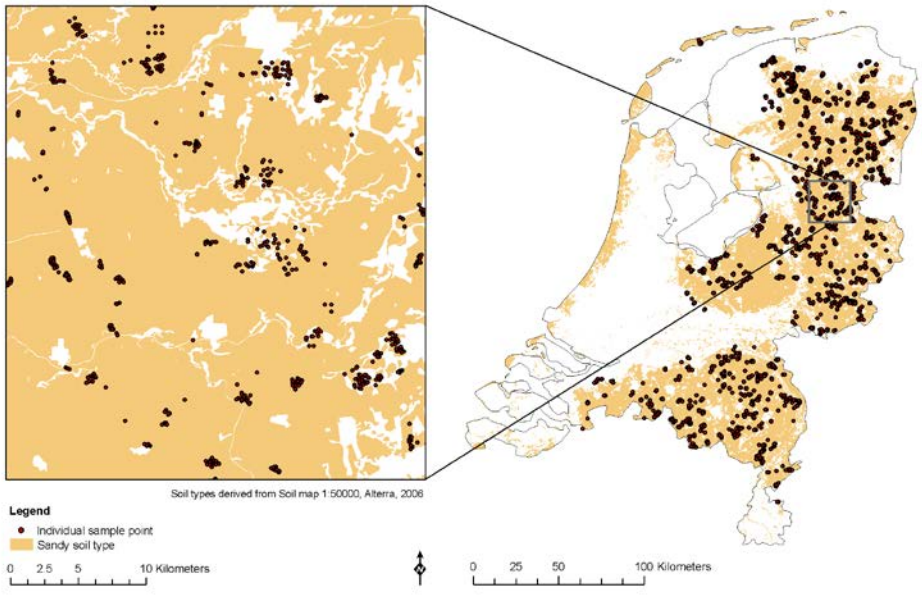


Figure 1. Spatial distribution of sample points for groundwater from the years 2007-2009 at national scale, and with inset detail of sample pattern

Results from past campaigns have shown that policy urgency is larger in some areas in The Netherlands (Fraters et al., 2005). In the sandy soil regions, exceedances of nitrate groundwater standards have been reported repeatedly and occur more frequent than in other soil types. Highly intensive animal farming is present here more than in other parts of The Netherlands, causing higher nitrogen surpluses, and sandy soils are more vulnerable to nitrate leaching (Boumans et al, 2008).

In the past, the RIVM has made use of interpolation techniques for the prediction of groundwater quality from shallow or deep groundwater wells (Pebesma, 1996), (Boumans et al, 2004, 2008) and distribution of pesticides and metals in soils using data from surveying networks (Tiktak, 1999).

1.2 Spatial Interpolation

Spatial interpolation deals with predicting values for locations that have unknown values. Measured values can be used to predict, or interpolate, values at locations that were not visited for sampling. Various authors have listed the possible methods available for interpolation, for instance Mitas & Mitasova (1999) and Harris et al (2010). Knotters et al (2010) provide an extensive review of available interpolation techniques and the degree of uncertainty associated with these methods.

In general, there are two accepted approaches to spatial interpolation. The first method uses deterministic techniques in which only the information from the point observations themselves are used. Examples are direct interpolation techniques like inverse distance weighting, or trend surface estimation.

The second approach also relies on regression of auxiliary information, or covariates, gathered for the target variable (such as regression analysis combined with kriging). These are geostatistical interpolation techniques, better suited to account for spatial variation, and capable of quantifying the interpolation errors. The estimation of a propagated total error for the final outcome of prediction maps is however a subject by itself and is not discussed in this report. Covariates are often available as inexpensive maps, perhaps originally intended for other purposes. A digital elevation model may explain quite well how water (and dissolved components) flows, but was maybe intended at first hand for cartographic purposes.

Depending on the available data, Hengl et al (2007) advocate the combination of these two into so-called hybrid interpolation. This is known as Regression Kriging (RK). In another paper, Hengl et al (2004) describe a framework for Regression Kriging that forms the basis for the research in this report. A limitation of RK is the greater complexity than other more straightforward techniques like ordinary kriging, which in some cases might lead to worse results (Goovaerts, 1999).

Both interpolation and regression techniques have been used before with similar data sets but the hybrid combination of both is not used very frequently yet for nitrate in groundwater. Some cases exist however, e.g. in the USA (Gotway & Hartford, 1996) and Portugal (Stigter et al, 2008). Also in Florida, USA, regression kriging was used for prediction of soil-nitrate nitrogen (Lamsal et al, 2009). Other approaches were also used to predict the nitrate content in groundwater, for instance Sonneveld et al (2010), use regression models only, whereas Woodard et al (2010) use a Bayesian method of interpolation. It can be concluded that many approaches exist and have been applied already, also in predicting water quality, but not all combinations have been seen. Regression Kriging as a combination of using a widely accepted interpolation technique like ordinary kriging, strengthened with the already available knowledge in ancillary maps of environmental factors can add some level of detail.

A variety of maps containing explanatory information is available. Groundwater-tables, soil maps and the amount of nitrogen used per year are promising candidates. Regression Kriging may also reveal some new use of field data for RIVM, which were collected for another purpose (informing local participants). By predicting levels and establishing the accuracy of the predictions, the outcome of this study can possibly be used to gain insight in detecting areas where, and in which degree set standards or legal limits are exceeded.

1.3 Research objectives

The main objective of this research is to predict nitrate levels at unsampled locations in upper groundwater in sandy soils in the Netherlands using Regression Kriging, and to assess the accuracy of these predictions. To achieve this goal, the sampling locations illustrated in Figure 1 are utilized, in combination with the knowledge from ancillary maps in order to predict nitrate levels for the sandy soil region as a whole. The emphasis is on the practical application of established regression kriging methods and achieving results with them.

1.4 Research questions

1. Which environmental covariates are related to groundwater quality and are useful in a spatial interpolation?
2. How can the covariates, determined in (1), be used in a regression model that predicts the groundwater quality from the covariate information?
3. How can the regression model be combined with kriging in the case of nitrate levels in upper groundwater and how accurate are the regression kriging results?
4. What are the differences between resulting maps for three consecutive years when the same methodology is applied, and can these differences be explained?
5. Will the model, when constructed at two extents (national and regional), differ in structure and accuracy, and can these differences be explained?

Ad 1. This will result in a list of the most selected covariates for, in this case, nitrate concentration in groundwater.
Ad 2. As will be demonstrated
Ad 3. As will be demonstrated. Results will be discussed by means of summary tables and uncertainty maps.
Ad 4. When the data from different years are used for prediction on the same region, do the linear models change? How? And what are the differences in the map outputs?
Ad 5. Does the change of extent lead to effects on predictions and prediction errors and do these occur at other extents too? Can these differences be explained by (knowledge of) regional circumstances?

1.5 Report outline

This report is structured as follows. Chapter 1 is a general introduction, stating the research problem and research questions. In the second chapter the data set is described and the study area is characterized. Chapter 3 summarizes the research methods, whereas in Chapter 4 the results for one region are presented in detail, for three different years. The general outcomes of all model parameters are given after that but since they are a repetition, much is placed in the appendices. In Chapters 5 and 6, the results are discussed after which conclusions are presented. Further recommendations are also made in Chapter 6. In the Appendices finally, more details are available in the covariate descriptions. Map results for regions untreated in the main text are also given in the appendices, as well as the developed scripts with R-code.

2 Study Area and Exploratory Data Analysis

2.1 Study area characterization and data description

The research area is limited to the sandy region of the Netherlands, in particular agricultural terrain in The Netherlands. This limitation is set by the availability of samples and the methodology of sample collection in the data set. Most samples were taken on agricultural parcels. The available data set concerns the years 2006-2009. In order to also cover natural terrain like forest and heath lands on sandy soils, data from an acidification survey study are also available for the same period to include natural terrain and increase the coverage of data. In this data set, only nitrate concentration levels are available at the individual point locations. Therefore, this research is targeted at nitrate occurrence. Other regions, having dominantly non-sandy soils (like clay or peat) are not part of the research.

Figure 2. Boundaries of the sandy soil areas in The Netherlands. N=north, E=east, C=centre and S=south. Wadden Islands (in fat grey) belong to, but are excluded from the North area.

In the monitoring programmes of the RIVM, four regions were defined in which sandy soils are the dominant soil type. Figure 2 shows the geographic orientation of those regions. The delineation of these regions was based on soil material occurrence, land use and socio-economic reasons (Fraters et al, 1998). This simplification is mainly for administrative and organizational purposes. Since these regions are not homogeneously consisting of sand only, other soil types may occur as well when sampled. Moreover, sandy soils may also be present in smaller quantities in other regions. In these areas no samples were taken, notably the dune areas at the western coastline. This is mainly caused by the absence of agricultural activity but also by the limitations of the monitoring program to Pleistocene soils (Fraters et al, 1998). The latter ones obviously are not taken into account. Some characterizations of the four sandy soil regions are given in Table 1. The four described regions are the target areas for the interpolation process. Grid dimensions for the regions are listed in Table 2.

North

Sandy soil region consisting of the province of Drenthe and large parts of Friesland and Groningen provinces. During the Saale Ice Age, large ice sheets locally shaped landforms (phenomena like 'De Hondsrug' and 'Fries-Drents Plateau'). Glacial till has influenced the subsurface. In this region, land use in the western half is dominated by dairy farming while in the eastern half cattle breeding and arable farming can be found, mostly low intensity type farming. In the eastern half of the region, sandy soils alternate with degraded peat soils. From the four distinguished regions, this is the most 'humic' one, having abundant areas with peaty sub layers and former moorland. Nature occurs sparsely and is spread evenly over the region. Only a small proportion of this area is urban. In general, the nitrate concentrations found here are typically low to medium, depending on the farm type and intensity. The Wadden islands administratively are part of this region but are excluded from the interpolation because of lack of samples, isolated position and different soil genesis.

East

This region is covering most of the provinces of Overijssel and Gelderland. Agriculture here in general can be classified by having a high intensity and most often consists of dairy farming and cattle breeding. Diverse in soil type, loam and clay areas occur along with the dominant sandy soil type. Nitrate levels found here are among the highest, being second only to the South region. Most dominant land use type is grassland. Denser urban areas are present in and around the cities of Almelo, Hengelo and Enschede. There is some increase in elevation towards the German border.

Centre

Agriculture is typically consisting of high intensity types like hog-breeding or poultry farms. Almost no arable farming occurs here, while in the northern part of this region dairy farming on grassland is the main land use type. Large contiguous areas with only nature exist (Veluwe, Utrechtse Heuvelrug) with some elevation differences, that were shapes by ice sheet movements, dating back to the Saale glacial period, like in the North Region. The nature share contributes around 44% of the total area. Overall, this region is quite consistently sandy. Only a few minor urban areas exist in this region.

South

Covering most of the province of Noord-Brabant and partly the province of Limburg. This region has a mixed and more intensive type of agriculture, consisting of dairy farming, cattle breeding (mainly hogs, fowl, cows) and arable farming. The highest values of nitrate in groundwater can be found in this region. Nature areas are mostly concentrated (notably De Peel, Loonse- en Drunense Duinen). Along the river Meuse and its tributaries, thick banks of river clay have accumulated. Some loam and peat areas exist in the centre of this region. Elevation increases towards the south. From the four regions, this one has the largest amount of urban area.

Table 1. Areal and land use characteristics of the four regions, based on reclassification of LGN6 (Alterra). Actual sand percentage based on Alterra's simplified soil map 2006

Region	Area (hectares)	agriculture %	nature %	other %	actual sand %
North	535298	70.4	17.4	12.2	69.5
East	375321	70.5	16.8	12.7	74.1
Centre	227162	37.4	44.2	18.4	80.1
South	569005	57.9	21.7	20.4	76.2
Total for the Sandy Soils	1706786	61.7	22.3	15.9	74.2

The amount of surface covered by both agriculture and nature originate from the LGN6-land use classification. "Nature" here also includes forests, although these could be production forests. During the years of the study period, slight changes in land use have occurred, but these are assumed to be minimal and concern mostly within-class changes. The actual soil type, upon which these land use types are situated, will not always be sand as other soil types may occur within the large sandy soil regions. In Table 1 the percentages of actual sandy soil-based samples are given per region in the last column. The Centre region can be described as the most sandy (comprising over 80%), while the North region has the lowest sand percentage (69.5%). The dimensions of the model regions are presented in Table 2. Region south is by far the largest region.

Table 2. Regional grid dimensions in number of 25x25m cells.

Grid dimensions	North	East	Centre	South
rows (y)	3806	3431	2802	3566
columns (x)	3731	2740	2578	5582

2.2 Sampling setup – monitoring networks

The dataset consists of a combination of data from two monitoring networks in the Netherlands, TMV and LMM. TMV – TrendMeetnet Verzuring in Dutch – stands for Acidification Trend Monitoring Network. The network records the effect of atmospheric deposition of acidifying and eutrophicating substances from the atmosphere on the groundwater. In an evaluation of the monitoring network (De Goffauet al, 2009) TMV is described as follows: "*The TMV was established in 1989 and is administered by the RIVM. The network monitors the quality of the top first meter of groundwater under natural areas (forest and heather land) with sandy soils. The groundwater under these areas is not affected by any other notable acidifying and eutrophicating substances and, in addition, sandy soils have a limited capability to neutralize the impacts of acidification. For these reasons, the impact of atmospheric deposition are most clearly detected under natural terrains with sandy soils. In other monitoring networks, the effects of atmospheric deposition are difficult or impossible to distinguish from other sources of pollution. In agricultural areas, for example, the impacts of fertilizer application on groundwater quality eclipse those of other sources of pollution."*

In other words, since no fertilizer is applied on natural areas, the expectation is that the relation with for instance nitrate levels in groundwater, is strong for atmospheric emissions

(Boumans et al., 2004). The representativeness of TMV is described in De Goffau et al (2009) and Masselink & De Goffau (2010). Within TMV, 150 locations, spread over nature reserves on sandy soils are visited bi-annually. At each of these 150 locations, 10 samples are taken, usually in a straight line with 50 meters between samples. Practical limitations to the program are, that the locations had to be owned by Staatsbosbeheer (Dutch Forestry Commisision) and that groundwater was accessible within 6 meters below soil surface.

In the other monitoring network, LMM or Minerals Policy Monitoring Programme, the focus is on agricultural practice. Here the application of fertilizer and animal manure, amongst others, is the leading factor for the level of nitrate in groundwater. In the report for 2007-2010 the setup is described for all soil regions in the Netherlands (De Goffau et al, 2012). The backgrounds of the monitoring programmes are explained in (Fraters et al., 1998). The aim of the LMM-network is described as follows: "*The objectives of LMM are monitoring the water quality on farms and explaining the results in relation to agricultural practice on those farms. Up to 2006, the results of the LMM were primarily used to assess the effectiveness of Dutch agricultural mineral policies. This network monitors the impacts associated with the EU derogation, adjudicated to the Netherlands, for the permissible amounts of nitrogen from manure on grassland farms. Since 2006, the number of farms monitored has increased considerably. Secondly, the network now consists of a stationary group of farms. Prior to that, monitoring was done on a revolving group of farms from the total number of participating farms. Thirdly, the sampling frequency for water quality monitoring has gone up. Finally, the interest in the quality of surface water has gradually increased; at the onset, LMM focused largely on groundwater, water from drains and soil moisture.*"

Since the LMM is limited in the number of samples taken, and objectives are aimed at national goals, a stratified sampling strategy is applied. The research population is confined to the most important farm and land use types. Regarding the representativeness of samples in sandy agricultural areas, Buis et al (2012) and De Goffau et al (2012) describe the selection process for participating farms in LMM and its sub programmes (being the other non-sandy soil regions and specific farm type programs), as well as the statistical support for the necessary number of farms. To give some idea, if the total population consists of farms with certain characteristics (minimal land surface size of 10 hectares, specific economic boundaries, and geographic representation), in 2010 nearly 2 percent of that population was sampled (around 300 of the potential 18000 farms) (de Goffau et al., 2012).

The sampling design was focused at farm level, in which all land in use by a farm is sampled proportionally. When the same farm is visited again in another year, approximately the same locations are sampled, lest the farm area has not changed too much. These semi-permanent sampling locations are considered easier than permanent wells, since they are quick to install and remove and do not interfere with agricultural management practices. They are also easily adaptable to the varying groundwater tables.

Note that the field measurements that are used for this thesis, are not the official data that RIVM reports. The official data consist only of analyses in a certified laboratory, where conditions are controlled and measurements can be assured to comply with international standards. The field measurements correlate with the lab measurements at a high rate (near 1:1) but can only be compared at farm level, since the groundwater samples that were collected in the field are analysed only after they are mixed for each farm (the 16 samples are reduced to one sample), thus limiting analysis costs. In practice so far, the field measurements have been labelled as 'indicative', in order to provide participating farms with a general impression of nitrate levels that were found under their farmland. The same holds for the samples from TMV, where the 10 field samples are mixed into one

sample per location and then analysed for many more compounds in the laboratory. The field data for all individual point samples have not been used for any official interpolation study before.

2.3 Sampling and the sample data set

During the years 2007-2009 a total of 18 438 samples were collected. Samples were taken at an average depth of 2-5 meters below the soil surface, but always in the upper meter of groundwater. Each sample was taken following Standard Operating Procedures (documented in De Goffau et al (2012), and Masselink & de Goffau, 2010)). The sampling, in short, takes place as follows: manually, a hole is augered until the local phreatic aquifer (the groundwater) is reached. A tube is then inserted, by which a water sample is taken from the first meter of the groundwater layer, pumped up by a peristaltic pump. The water is filtered with a 0.45 µm in-line filter to prevent small particles from clouding and influencing the sample. After some measurements, the tube is removed and the hole is closed again. This is repeated at each farm for 16 locations, and at each nature site for 10 locations.

At each location, the nitrate content of the groundwater was determined in situ, using a simple portable Nitrachek 404 reflectometer. A determined volume of the sample is used on a test strip, which has one minute to react before it is placed in the reflectometer. This field method is fast but reasonably accurate, enhanced by a temperature correction and a batch verification of the test strips afterwards (Vissenberg, 1994). The standard measuring range is from 4 - 440 mg/l, where values below this range will yield a 'low' sample rating, while 'high' values indicate that samples need to be diluted first before repeated measurement. The device precision is 1 mg/l. Maximum values found in the field during the research years range to as high as 1800 mg of NO_3^- per litre.

The nitrate field measurement is based on a chemical reaction, generating a colour change, which is then registered. This reaction is temperature sensitive and therefore all measurements afterwards get corrected for the temperature at the time of sampling. For this purpose, temperature is one of the parameters which are also monitored while sampling. The test strips are produced in a batch, each having a specific production date. The batch is also tested under lab conditions and results are used to correct the found concentrations when needed. Other parameters include pH, electrical conductivity, groundwater depth and GPS recorded x,y-coordinates. The water samples later are mixed and analysed in a laboratory for determination of many more chemical environmental parameters.

The time necessary to acquire one sample, varies with conditions encountered in the field like accessibility, actual soil type, local depth of groundwater and presence of loamy particles (thus increasing filter time). Under average circumstances, the time needed to collect the sample ranges between 15 minutes and one hour per sample.

The farms in the sandy regions are sampled during the summer period, roughly ranging from May until September. This has some practical reasons, the main ones being that the water can still be related to agricultural practices and that the water is already out of reach for crops at this depth and season of the year. However, samples that are taken in nature reserves, generally are taken in winter period, when retrieving samples is easier (higher groundwater tables) and less disturbance to plants and wildlife occurs. In Figure 3 this structure can be seen where there are two peaks, one in the start of the year, the other around midyear. In 2009 the winter peak is almost absent because of a break year in sampling. The sampling date is not taken into account for this study; only the year in which the sample is collected. Sampling procedures are the same, but the sampling strategy is slightly different for natural areas: instead of 16 area-distributed samples per

farm, 10 samples are usually taken in a transect of 500 m. This is because of the size of most nature reserves.

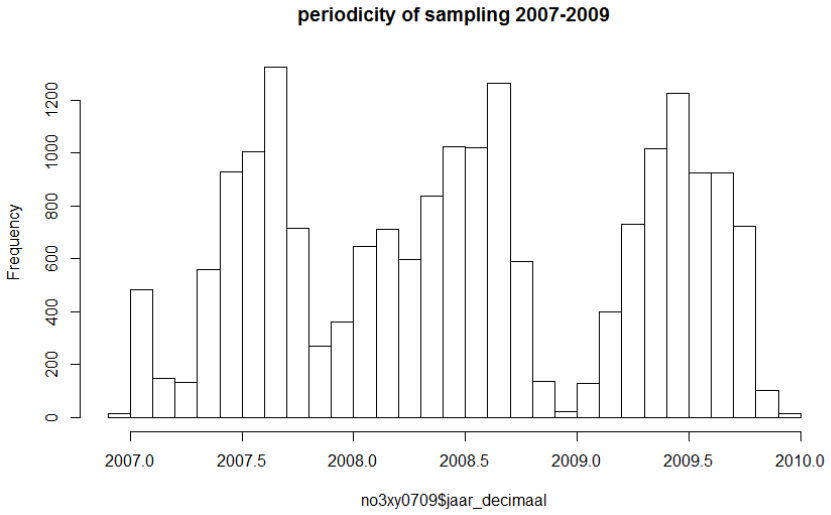


Figure 3. Sampling distribution during the research years 2007 - 2009

The actual percentage of sampling points on a sandy soil type is given in the last column of Table 3. This table also classifies the sampling points in terms of agriculture and nature[2] based sample. Since these data are derived from the 1:50 000 soil map, reality may differ somewhat.

Table 3. Distribution of sample points from 2007-2009 over the regions and classified as either nature or agriculture. Last columns indicates number (and %) of sample points with actual sandy soil type.

Region	**Number of Sampling Points**	**# of SP in *Agriculture*[1]**	**# of SP in *Nature*[2]**	**# of SP in *Other*[3]**	**# of SP at actual sandy soil[4]**	**% samples on sand**
North	6734	6234	487	13	4581	68.0
East	3980	3764	203	13	3084	77.5
Centre	1234	1132	97	5	772	62.6
South	6490	6024	462	4	5949	91.7
Total	18438	17154	1249	35	14386	78.0

The attributes of the final sample data set for this research are x,y-coordinates, year and sample date, depth of groundwater table, and the NO_3-concentration in mg/l. Corrections for temperature and batch number lead to concentrations below the precision of the field device (4 mg/l), as well as treating missing values with a common measure (0.5 * lower limit of detection). That last measure was rarely used since field procedures require double or even triple measurement of a sample. The final results are the averaged values. In Table 4 some standard characteristics are given. Further analysis is presented in Section 2.4. The total number of measurements per region were presented already in Table 3, whereas the number of samples per year for each region is listed in Appendix IVa. We assume there are no differences in sampling methods, and that obvious errors have been removed by scripts that database managers previously applied. These scripts check,

[1] As classified in LGN6 'Monitoring' class Agriculture
[2] As classified in LGN6 'Monitoring' class Nature + Forest
[3] LGN 'Monitoring' classes other than Agriculture or Nature + Forest. Presumably these points are actually situated in either agriculture or nature, but the LGN6 resolution of 25m forces these into neighbouring cells
[4] Taken from the simplified Alterra soil map, 2006

amongst others, for valid ranges of data, and whether required attributes are missing. After data retrieval from the central database, data is considered as valid.

Table 4. mean, median, min and max properties for NO_3-samples (in mg/l).

avgno3 / year	North	East	Centre	South	All regions
mean					
2007	57.61	71.70	35.61	116.80	76.73
2008	41.01	61.31	30.69	141.50	84.78
2009	33.90	46.86	25.18	132.50	76.73
median					
2007	12.67	47.32	9.58	74.24	35.65
2008	8.20	27.68	6.86	97.12	29.30
2009	7.63	14.31	6.71	80.41	20.05
minimum					
2007	1.13	1.13	1.13	1.02	1.02
2008	1.13	1.13	1.13	1.13	1.13
2009	1.13	4.25	1.13	1.13	1.13
maximum					
2007	585.60	559.80	256.91	982.50	982.50
2008	1176.0	1222.0	308.40	1725.00	1725.00
2009	677.30	530.90	298.79	1232.00	1232.00

From the first quick scan, it appears that the highest values for NO_3 are found in the South region, followed by East, North and with the lowest range of values, the Centre region. All regions show a decline during the sampling years, except for the South region, which by its size in samples is heavily influencing average total numbers.

2.4 Exploratory Data Analysis

All field data are stored in a central database. After retrieval, the dataset containing the original location measurements and field analyses needs to be examined for its distribution and disturbing features. Next, only records containing valid x,y- locations are used in the statistical modelling and prediction process. One check is to print all measurement locations to check whether they are located within the defined sandy regions. If not, they are removed. For kriging purposes, duplicate x,y-locations need to be checked, in order to prevent singularity issues (Pebesma, 2004). Duplicate locations share the same coordinates (based on one decimal digit), making it impossible to apply interpolation. Therefore the choice is made to delete each second record that has duplicate coordinates. This way, 11 duplicates (on a total of 18 960) are removed.

NO_3-data distribution

The data set with field sampling results contains around 19 000 point samples, distributed over the three research years 2007, 2008 and 2009. In Table 4 the data were tabulated, while the boxplot graph per year in Figure 4 adds some more information on the data distribution per year for all regions combined.

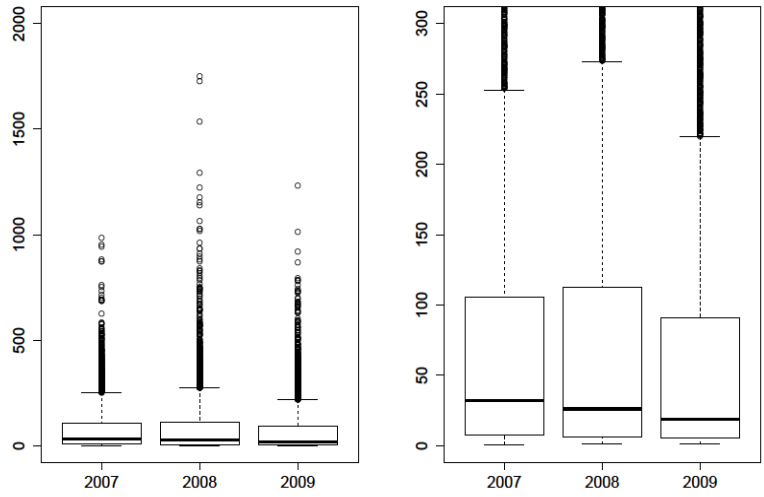


Figure 4. Nitrate measurements for 2007-2009, all regions combined. Left: All data. Right: same plot, zoomed in at bulk data area

Arranging the data by sampled year already shows a few interesting features. The 'box' of the boxplot represents 50% of all measurements. The median value of each measurement year is well below the mean, as can be read in Table 4. The lowest possible value is zero, as negative concentrations are not possible, but in the data values lower than 1.02 does not occur. Apparently, the majority of the data are relatively low values (around 100 mg/l or lower) but there are also very high values. The values above the 'whisker' indicator are known as 'outliers'. These measurements are considered valid ('possible', not an error) and usually can be explained, for instance by sampling an old concentrated animal manure spot or cattle dropping from previous years. These high values are known to occur on arable farms as well, where excess fertilizer can infiltrate easily after harvesting. Unless there are reasons to reject certain values, e.g. an obvious error was made, they should remain in the dataset. To distinguish between accidental high values and actual agricultural practice is not always possible. These outliers are thus kept in the dataset. The distribution of the data, when we zoom in by region, can be seen in Figure 5 and 6.

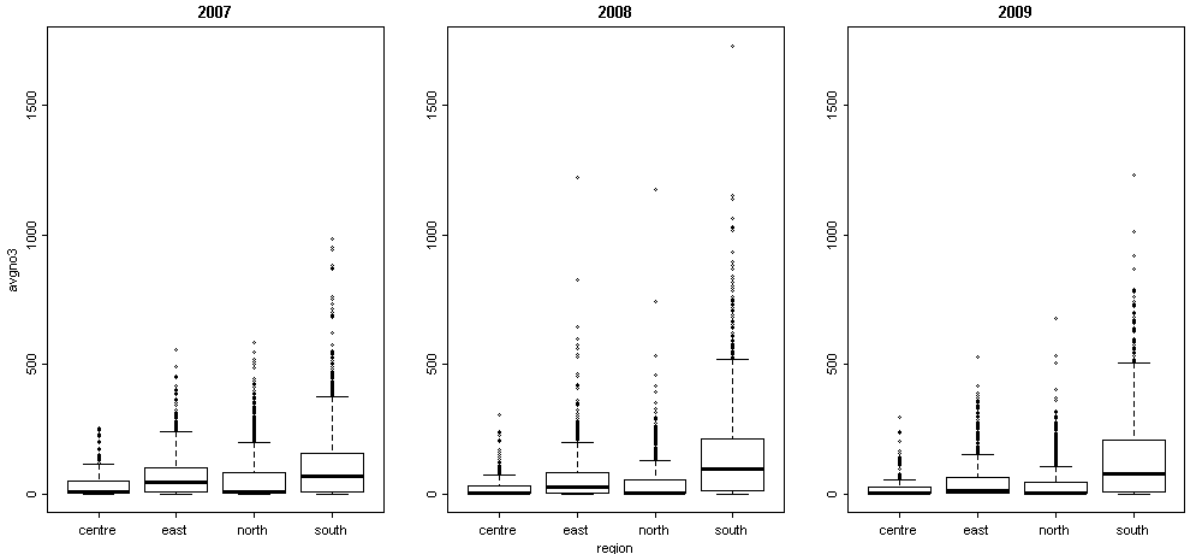


Figure 5. Nitrate measurements per region for the years 2007-2009 (see below for more detail).

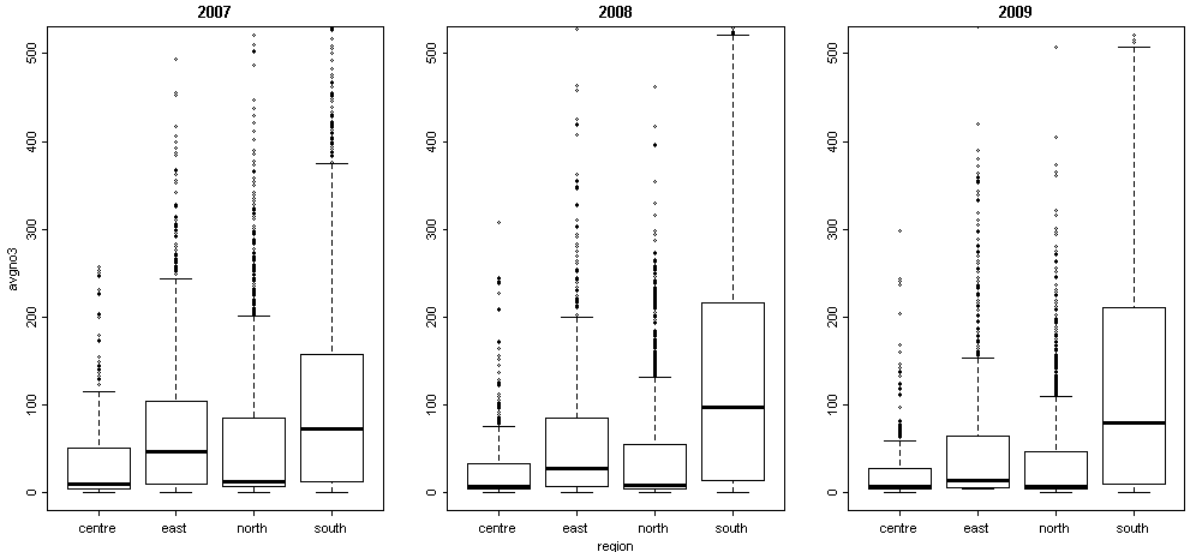


Figure 6. Nitrate measurements as in Figure 5, displaying only values maximized at 500 mg/l.

As was clear from Table 4, the highest values are found in the south region, marked by the outliers. Overall, this region presents far higher concentrations of NO_3 than the region with the lowest levels, the centre region. Judging only by nitrate levels in the regions, it seems that concentrations decrease from 2007 onwards, but attention is drawn again to the south region, where an increase can be seen in 2008. In the following year levels are decreasing again. Since a large share of samples is from the south, this is influencing the average of all regions combined.

Combined data distribution

From the previous boxplots, it was clear already that there were many values in the lower levels, and only few in the high levels, resulting in a long tail to the right (Figure 7, right). This distribution is inherent to measuring natural phenomena. It can however not be considered as a normal distribution. For statistical modelling it therefore makes sense to transform the target variable to as close to a normal distribution as possible. All further analyses should be done on the transformed data (Webster & Oliver, 2007). In Section 3.1 the transformation of the data is considered.

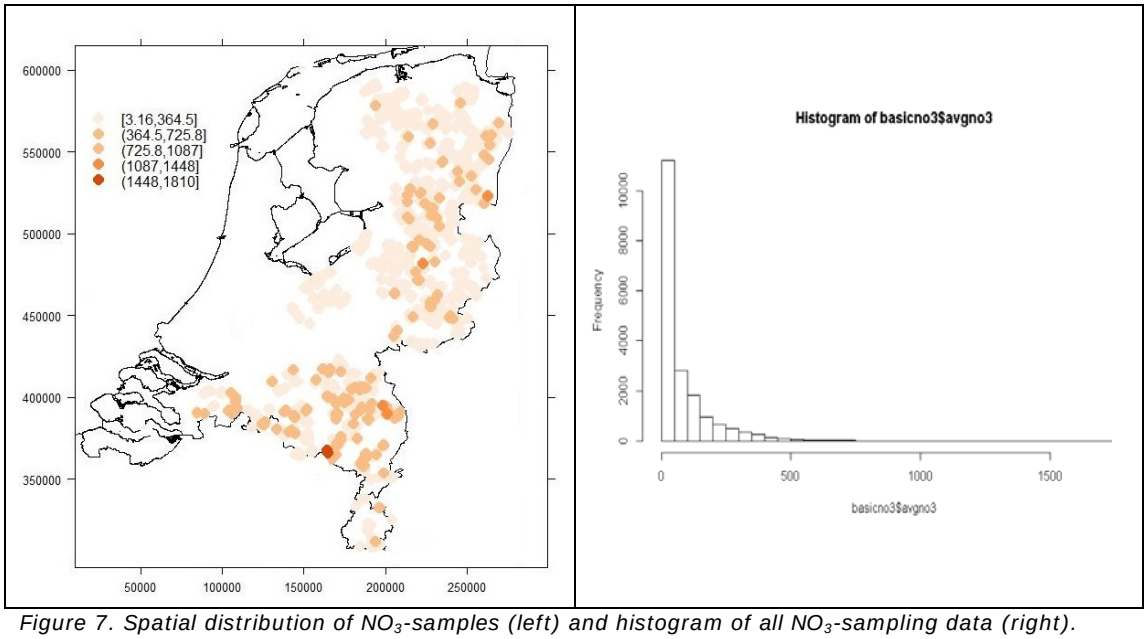


Figure 7. Spatial distribution of NO_3-samples (left) and histogram of all NO_3-sampling data (right).

The spatial distribution of the sampling points (Figure 7, left) is difficult to display at national scale and needs to be interpreted with care. As the symbols are proportional, the high values seem to take more importance. It only gives an impression of where higher values can be found.

2.5 Covariate selection

Many natural and human-induced phenomena have been mapped and are available as input as covariate variables at various scales. Some of these maps have promising explanatory properties with regard to nitrate, as will be investigated in the next paragraphs. Groundwater tables for instance, indicating at which depth range the phreatic water levels are situated, hold a close relationship with nitrogen reduction speed and thus may explain the spatial pattern in the nitrate levels. However, not all maps are suitable in the modelling phase as the explanatory value may be too little or may not be present sufficiently enough in the measured data set that is available. Selected maps are listed below and in Table 5, and more background information on these maps can be found in the appendices.

In general, the usefulness of a covariate map revolves around two aspects:

(1) Strength of relationship with the variable of interest
(2) Availability

All covariates need to be independent as well. Each of the selected maps is initially examined for its explanatory value, first by charting correlation with the data set in a scatterplot or boxplots. This determines whether to use the variable in a second step, the regression modelling. To keep the procedure alike for all regions and years, the same covariates are offered in each regression model. This is described in Section 3.1.

Strength of relationship

Regarding nitrate presence and behaviour in soil and groundwater, one can think of a number of potential explanatory variables. This includes for instance, soil factors (reactive parameters; e.g. organic carbon (Rivett, Buss, Morgan, Smith, & Bemment, 2008)), meteorology (dilution and transport effects), hydrology (mixing and transport parameters) and land use management (fertilizer application). In other studies, many parameters have already been mentioned, like soil type, land use and geohydrological conditions as stipulated by (Pebesma, 1996) and (Stigter et al., 2008). (Gotway & Hartford, 1996) use harvest yield data to predict soil nitrate. Nitrogen surplus on farms is also often mentioned (Sonneveld et al., 2010) and (Boumans et al., 2008), as well as livestock density and deposition from atmosphere (Bonten et al., 2009). Sometimes relations seem to exist but cannot be explained well. A correlation might then be used, but may be based on side-effects of another phenomenon. Caution is always necessary when applying such correlations.

Availability

One prerequisite is that these variables need to be available in map form for all the areas targeted for interpolation, as stated by Knotter et al (2010), while another requirement is that the detail of these variables is sufficient to add explanatory power. In most of the accompanying map documentation, information is available on how these maps were constructed and how accurate the represented variables are. Such metadata is important when interpreting the results of the model fitting (Hengl, 2009). This might later be used to judge the reliability of predictions.

Unfortunately, not all desired parameters are available in sufficient detail or completeness. Meteorological parameters like precipitation and evapotranspiration are available at many weather stations, but need interpolation themselves in order to provide a full cover map. This leads to other complications in accuracy and methodology, and therefore these data are not part of this study. A map showing the mineral pyrite content (related to nitrate decomposition) also was not available to cover all of the sandy areas.

The origin or composition of some of the other covariate maps may lead to complications too, since these are produced by a model or have been derived from various other maps. This may lead to a set of covariates that are more or less related to the same properties. Other parameters, for instance livestock data, are available from national inventories, local studies or field scale methods, but not directly suitable for the regions as required in this research. Some of these are too unspecific or aggregated to be useful.

Maps typically harbour uncertainties and errors in them, both systemic and methodological. From the selected maps listed in Section 2.6, in Appendix V the original uncertainties are given, when published. Sometimes these are unknown.

Examples of covariates

Organic matter is important for binding nutrients and soil moisture, for the soil structure and stability but also for rootabilty and accessibility to soil organisms. Decomposition of organic matter liberates nutrients for vegetation (De Vries, 1999). Maps are available for different soil depths, since variation in organic matter occurs not only per location but also from top soil to the deepest available layer of 1.20 meter.

Nitrogen-application. These maps are a combination of nitrogen in fertilizer and nitrogen from animal manure. The map data are modeled distributions of farm based manure production data, together with applied fertilizer in the format of so-called stone-plots. This is a characterization of 6505 individual plots, each having a unique combination of soil type, land use and regional differences. Minimum size is 250 x 250 m, but larger cells exist as well. The files are available for four years, stone5 (from 2005), stone6 (from 2006, etc.) to stone8. The data for 2009 were not available at the start of this study.

From the data description provided in the file documentation, many auxiliary data share a common base, for instance the soil map is the basis for many other products, while the interpretation of satellite images for land use map LGN6 also uses the statistical map BBG06 as reference data. Top10nl topographic maps were used to delineate satellite data classes. Therefore, borders and classes from one map may be present in another map as well. This might lead to so called spatial dependence or multicollinearity. However, the stepwise linear regression method selects only those covariates yielding the most significant values.

Logical choices have to be made about the validity of a given data set for a certain year: most of the data sets were produced before the first year of the modelling, so before 2007, or at least the data contained in the maps were recorded before the time of interest of this research. These maps can simply be applied for each year. Only for the maps NHx (atmospheric N-deposition), which dates from 2010, and the maps with the combined fertilizer/manure application (stone07-stone09) this is not entirely true. These have a clear link with a certain year.

The atmospheric nitrogen emission of NHx-map is based on data from 2010, but this map is provided with a so-called scale factor for the previous years, 2007-2009. This scale factor is a linear derivation and therefore the resulting maps are not different in exact

pattern but only in magnitude. This means that the NHx-map is essentially the same for each year to the model, and feeding it as covariate just once is sufficient.

2.6 Available covariate maps

Covariate data are usually divided in two types, being continuous data and categorical data. The first kind is a measurable phenomenon that follows a certain range of measurement in a defined unit, for instance elevation in meters. Typically, these are grid-style maps. Next to that, there are categorical data maps, defining where each map unit begins and ends. These are usually vector-style maps (Burrough & McDonnell, 1998).

Each of the candidates is addressed shortly in the following two Sections. More detailed descriptions of the possible covariates, and further references for them, are listed in appendix V. The complete list of auxiliary data is described in more detail in the annexes (Va, Vb).

Covariate data preparation

Most covariate maps need some treatment in order to be useful as input. Before the models can be constructed, the covariate map data need to be in similar projection, alignment and preferably, resolution. In order to assign each future grid cell a residual covariate value, it is necessary to make grid maps which have exactly the same cell size and which are originating exactly at the same point. Preparation took place by means of a Modelbuilder model in ArcGIS (see appendix III for details). In order to calculate with regional extents, this model also cuts out four regional data sets aimed at the individual regions North, Centre, East and South.

For categorical variables, often a processing step is necessary, as they need to be converted from vector maps to grid maps, with 25 x 25 m resolution. Next these maps must often be reclassified, as some of the data classes may not be significant enough during the model selection phase. The land use maps `bbg06` and `lgn6` for instance, contain 30+ classes, not all of them appearing relevant to nitrate levels. Some authors therefore create maps for each single class or derive a single aspect like a slope class or proximity to certain objects. The non-significant classes can then be rejected by the model. Reclassification and grouping are therefore relevant options to eliminate the surplus of classes. For instance, in a land use map one could merge all forest types as nature in one class and all arable crops together as arable farming.
Reclassification of categorical covariates in the end has not pursued all maps, leaving a refinement step for follow-up research. Only the groundwater table map and the soil maps were reduced since some specific classes hardly occur.

An overview of all the covariate data that are used in this thesis is presented in Table 5. The covariates will be used by the abbreviated name, as listed, throughout this thesis. Descriptions and maps of the covariates are available in Appendices Va and Vb.

Table 5. List of available covariates

#	abbreviation	type	name	data description # appendix V
1	ahn	n	elevation model	9
2	bbg06	f	statistical land use	1
3	draf	f	water discharge by drainage - superficial	8
4	dront	f	water discharge by drainage - profound	8
5	geom	f	geomorphology	2
6	gronds	f	simplified soil types	3
7	gt06	f	groundwater tables 2006	6
8	kwel2	n	seepage/upwelling (vertical velocity mm/day)	10
9	laf	f	distance to discharge by ditch	8
10	lgn6	f	satellite land use map	7
11	lont	f	distance to profound discharge	8
12	nhx	n	nitrogen emission	11
13	om05	n	organic matter 0-5cm	12
14	om10	n	organic matter 5-10cm	12
15	om25	n	organic matter 10-25cm	12
16	om40	n	organic matter 25-40cm	12
17	om60	n	organic matter 40-60cm	12
18	om80	n	organic matter 60-80cm	12
19	om100	n	organic matter 80-100cm	12
20	om120	n	organic matter 100-120cm	12
21	pawn	f	hydraulic soil property districts	5
22	slaf	f	distance to discharge by ditch - superficial	8
23	slont	f	distance to discharge by ditch – profound	8
24	stone5	n	fertilizer and manure application 2005	13
25	stone6	n	fertilizer and manure application 2006	13
26	stone7	n	fertilizer and manure application 2007	13
27	stone8	n	fertilizer and manure application 2008	13
28	vds	f	soil map RIVM classification	4

n = numerical or continuous value map (15)
f = factor or categorical map (13)

3 Statistical methodology

3.1 Regression Kriging modelling

The most basic form of kriging is called ordinary kriging. When we add the relationship between the target and covariate environmental variables at sample locations, and apply this to predicting values using kriging at unsampled locations we get Regression Kriging. In this way the spatial process is decomposed into a mean and residual process. The first step of regression-kriging analysis thus is to build a regression model by using the explanatory grid maps (Hengl, 2009). The kriging residuals are found by using the residuals from the regression model as input for the kriging process. Adding up the mean and residual components finally, results in the regression kriging prediction. In a simple form, this can be written as:

$$z(s) = m(s) + \varepsilon'(s) \qquad \text{(eq.1, Hengl, 2009)}$$

With $z(s)$ being the value of a phenomenon at location s, $m(s)$ being the mean component at s, and $\varepsilon'(s)$ stands for the residual component including the spatial noise. The mean component is also known as the regression component, and is expressed as $m(s) = \beta_0 + \beta_1 X_1(s) + \cdots + \beta_m X_m(s)$ where X_i(s) are the explanatory variables and the β_i are the regression coefficients.

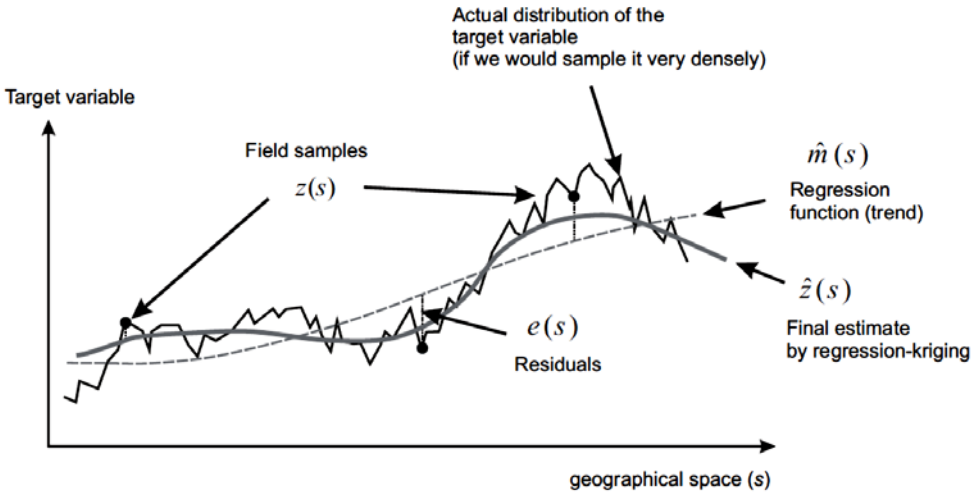


Figure 8. A schematic representation of regression kriging using a cross section (reproduced from Hengl, 2009)

The fundamentals and theoretical backgrounds of this approach are described in various sources, for instance in (Hengl et al., 2007; Hengl et al., 2004; Knotters et al., 2010; Webster & Oliver, 2007). The process of refining the prediction in two steps (trend estimation and kriging) is illustrated in Figure 8, where the result of the mean component, only regression, is visible as a dashed line $\hat{m}$(s), and the sum of trend+kriging is the curving thick line $\hat{z}$(s). This should approach the actual distribution better than either just a trend surface or just a simple interpolation.

A schematic representation of the regression kriging approach is depicted in Figure 9. To perform the regression kriging, the following steps are necessary (Hengl et al., 2004):

1. Determination of linear model(s) of the variable, and determine the residuals
2. Use the residuals for generation of a variogram to quantify spatial correlation
3. Application of the regression model at all unobserved locations
4. Kriging of residuals at the same locations
5. Addition of both predictions (steps 3 and 4)

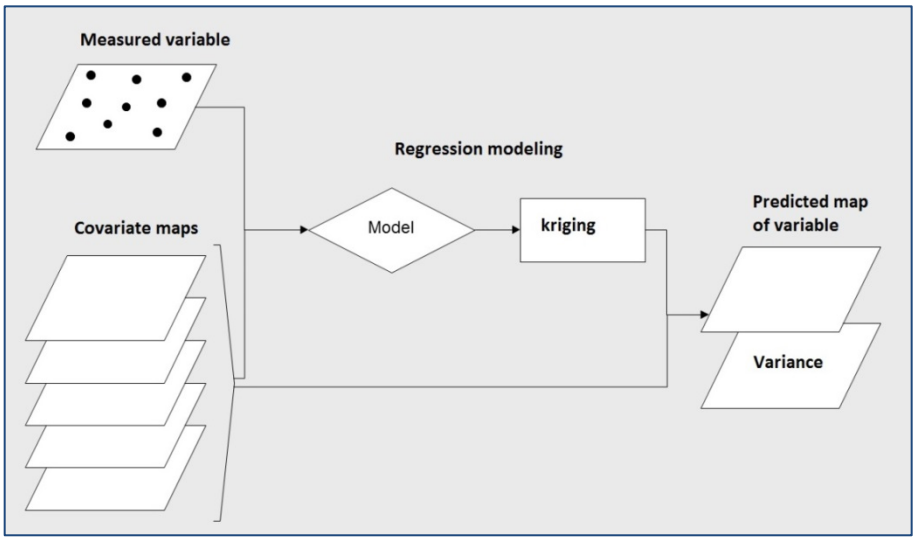


Figure 9. Regression kriging schema (after Kempen, 2011).

The linear modelling of the relationship between the dependent and explanatory variables is quite empirical. The model selection determines which covariable is important or not, even when we do not have process knowledge supporting this. It is not necessary to know all these relations, as long as there is a significant correlation. Once the covariates have been selected, their explanatory strength is determined by using (stepwise) multiple linear regression (MLR) analysis. For each covariate this leads to a coefficient value, describing its predictive strength, and whether this is a positive or negative relationship. With the combination of these values for all covariate maps, a trend surface is constructed. This regression prediction is in fact the calculation for each target cell from each input cell from all covariates times the coefficient value. The amount of correlation is expressed by the R^2 in the regression equation.

To enable this, the covariate data first need to be processed by overlaying the sample locations with the covariate data layers. In this way a matrix of covariate values for each sample point is constructed. This matrix may still hold several 'NA' or missing values due to the fact that some maps do not have coverage where others do. An example of this is the absence of information on organic matter in urban areas. Since the linear models cannot be constructed properly when some covariate data are missing, these sample points are discarded altogether. The resulting data matrix is therefore complete for all remaining measurement data points.

The second step in which the covariate data are needed is the model prediction phase of the mean surface values. First, a prediction mask is made. This prediction mask is the selection of grid cells for which covariate data is available and only contains the coordinates of valid cells. Next, the regression mean values are calculated by predicting the regression model for every grid cell that is in the prediction mask (see example in Figure 11). In the residual kriging phase this prediction grid is used again as a mask for the kriging prediction.

Regression modeling

The regression part in the regression kriging consists of constructing a linear model with a selection of those covariates that make the best contribution to explaining the (transformed) nitrate level. Since we have different years of measurements in the field data set, and at the same time discern different regions in the final grid map we want to predict to, several options are possible.

First, for each year in the measured dataset (2007, 2008 or 2009) a model can be applied. This will render a prediction map per year, at national scale (combining regions (n)orth, (e)ast, (c)entre and (s)outh). There is only a distinction in time, using just the data from the selected year. This implies three different models. All variables are up for selection. Regions for nationwide models have to be calculated separately to speed up and control the prediction process with the available computing means. The total number of model runs is as follows: number of years: 3, number of regions: 4. This results in 3 * 4 = 12 separate calculations of a map region-year combination, finally giving three different maps.

2007: 07n, 07e, 07c and 07s → all.regions.2007
2008: 08n, 08e, 08c and 08s → all.regions.2008
2009: 09n, 09e, 09c and 09s → all.regions.2009

Secondly, for each region a separate model can be constructed, based on the best 'fit' by selecting only those map variables most significant to the region. Predictions are then made for each region, based on each available year. This yields 12 different models, 3 years, 4 regions. Since the models are adapted to fit each region and year dataset specifically, they cannot be applied to the other regions, thus resulting in 4 * 3 = 12 region-year combinations.

north: 2007, 2008, 2009
centre: 2007, 2008, 2009
east: 2007, 2008, 2009
south: 2007, 2008, 2009

Not all these combinations will be presented completely in the main text, since many steps are repetitions. The methods followed in the research are illustrated for one region, presented in Chapter 4, for three different years. In the Appendices (IVa and IVb), the results for the other regions and years can be found.

For each year and region a subset is selected, using additive modeling with multiple linear regression and the Akaike criterion[5]. The lowest score for the Akaike value determines the outcome of the stepwise model selection. For the regression models, no interactions have been set. A selected variable can only contribute positively or negatively. Possible interactions between variables were not investigated to restrict the time allotted to the modelling phase.

The linear model regression formula can be written as follows:

$$10\log no_3(s) = \beta_0 + \beta_1 \times cov.1(s) + \beta_2 \times cov.2(s) + \ldots + \beta_n \times cov.n(s) + residual\,(s) \qquad \text{(eq.2)}$$

where βi stands for the regression coefficients assigned by the model, and *cov.* for the covariates. Regarding the automatic selection for all models using stepwise regression, sometimes categorical variables are selected while only one level is indicated as significant. Besides that, sometimes a selection seems irrelevant, since it does not have a significance at all. Although it is often considered good practice to accept only levels of p < 0.1, this

[5] Akaike Information Criterion (AIC): a value indicating the goodness of fit; see for instance (Webster & Oliver, 2007) and http://en.wikipedia.org/wiki/Akaike_information_criterion

rule would lead to the dismissal of many categorical covariates. Manual selection might often lead to fewer variables, but possibly the inclusion of one covariable strengthens the effect of another one. The model however with the lowest AIC value determines whether to incorporate a variable or not. With every added covariate the explaining power increases somewhat, generating a higher (adjusted) R^2. It does not automatically mean the model is better but this is accepted in order to automate the selection process as much as possible and constructing every model using the same method.

The regression formula and parameter coefficients are subsequently used to predict a regression surface. Adding up this regression prediction surface to the kriging prediction, ultimately yields the regression kriging prediction. The linear model also delivers the regression residuals (measured minus predicted values). With these residuals a *variogram* is then modeled, after which prediction by simple kriging delivers a second surface. This will also give us the kriging variance.

Variogram and experimental variogram

The residuals from the linear regression now can be modeled to display the spatial variability. When the distance between two point pairs increases, the variance increases, meaning that the similarity is decreasing. The experimental variogram is a plot of the semivariance against the distance between sampling points. The variogram is the fitted line that best describes the function connecting the dots from the experimental variogram. The following parameters are often used to describe variograms (see also Figure 10):

Nugget: the (positive) intercept on the ordinate axis
Sill: the value where a constant maximum or asymptote value is reached
Range: the distance where the model (approximately) reaches the Sill

The variogram can be constructed, using (Webster & Oliver, 2007):

$$\gamma(h) = \frac{1}{2}E[\{z(s_i) - z(s_i + h)\}^2] \qquad \text{(eq. 3)}$$

Where $\gamma(h)$ is the semivariance, $z(s_i)$ stands for the value of a target at a sampled location and $z(s_i + h)$ is the value at location $s_i + h$. The equation changes when the variogram is fitted to a variogram model.

First, an experimental variogram is constructed using the residual data from the previous modeling phase. The variogram is then fitted through the data both by `gstat` and by adjusting the variogram manually. In this way, for every regional RK-model the most optimal fitting variogram model has to be selected, resulting in a variogram formula including function, nugget, sill and range. The variogram terms are defined above and illustrated in Figure 10.

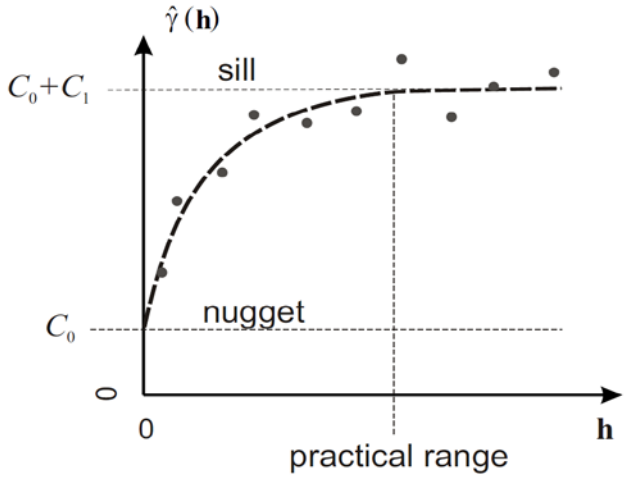


Figure 10. Basic estimated and fitted (exponential) variogram with terms (reproduced from Hengl, 2009).

As stated by Webster and Oliver (Webster & Oliver, 2007), fitting models can be difficult and fitting by eye can be unreliable. The automated fitting procedure in `gstat` allows an initial setting of parameters and then fits the best possible curve based on them. The resulting fit needs a visual check before commencing with the next phase.

The theory and background of kriging are not discussed or explained here. The basic formulas for regression kriging are given below (Hengl, 2004):

First the residuals are calculated using linear regression, the prediction yields a weighted average of all input covariables:

$$\hat{z}(s_o) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0); \qquad q_0(s_0) \equiv 1, \qquad \text{(eq. 4)}$$

$$\hat{z}(s_o) = \hat{m}(s_0) + \hat{e}(s_0) \qquad \text{(eq. 5)}$$

Where $\hat{m}(s_0)$ is the fitted drift and $\hat{e}(s_0)$ the interpolated model.

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} \lambda_i \cdot e(s_i); \qquad \text{(eq. 6)}$$

Where $\hat{\beta}_k$ stand for the estimated drift model coefficients, $q_k(s_0)$ being the k^{th} external covariate at location s_0 and $e(s_i)$ the residual at measurement location s_i , and λ_i are the kriging weights.
Ordinary kriging turns to Simple kriging when the mean is known. This will be the case as the known mean of (the results from the linear prediction) the residuals is zero. This kriging prediction is then added to the result of the regression modeling.

Some excellent readings on the backgrounds of kriging are available by Webster and Oliver (2007), Isaaks and Srivastava (1990) and also in Burrough & McDonnel (1998) a good overview is presented. In this thesis only the methods are applied, which are conveniently available for use in `gstat` (Pebesma, 2004).

Back transformation of the results

After adding up the MLR prediction and kriging prediction products, the results are still in the transformed `10log` format. Since this is not comparable with the original sample data, back transformation to the original (mg/l) units of NO_3 is necessary. For this, the calculated variance of the kriging process is needed (Webster & Oliver, 2007):

Back transformed predicted $NO_3 = 10\text{^}(\text{predicted transformed } NO_3 + 0.5\sigma^2)$ (eq. 7)

Where σ means the standard kriging deviation. The resulting prediction can be exported to, and presented using a GIS. Graphical capabilities of **R** are limited for larger SpatialGridDataFrames. The R-coding for the regional RK model is given for the south.2008 model in appendix I. A script was made for each separate region and year, in which certain region specific modifications were applied. Since the south region is the largest in grid size, it demanded the most computing resources; at least 16 Gb of RAM computer memory was required. More efficient use of memory by optimizing the R code is however very well possible.

Figure 11. Example of a prediction mask for the centre region, 2009. Grey indicates valid cell locations for predicting, white means no prediction will take place. For other years this may slightly differ.

3.2 Time comparison within a region

Within the sample data set, values are distinguished by year of sampling. The mean value per region varies between years, as was illustrated in Table 4. The prediction will also have different appearance when sample data for different years are the basis for prediction. Discriminating between years leads to a change in covariate strength, since model structure depends on the relation between the available set of covariates. The stepwise method judges whether a parameter should be dropped or selected in the model. Apart from the variogram fitting, the kriging part of the RK-approach remains the same.

Using the regression kriging models, values are predicted for the years 2007, 2008 and 2009. In every year-region combination, data are available for at least 348 points (centre, 2009) to a maximum of 2523 point samples (south, 2008). Each year-region combination will have its own uncertainty map (addressed in Section 3.5), giving the range of values in which the predicted value most likely will fall.

Differences will be explained visually and when possible quantitatively. When possible, the geographical distribution of selected covariates will be used to explain the differences.

3.3 Extent comparison

By looking at two different extents, national and regional, an effect can be studied in the output that will be generated. Models can use a wider span of data and are expected to behave differently upon this in contrast to smaller areas with less variation in offered covariates. Moreover, the composition of models is subject to change as well and parameters that were important in regional models need not occur in 'national' models, and vice versa.

The cell size in this report is set to remain constant at 25 x 25 meters. Therefore, the variation between extents is the amount of surface that is included in the analysis. The dataset allows for the investigation of regional behaviour of the interpolation process because of its large size. Like in Section 3.2, changing the extent may influence the selection of covariates.

3.4 Cross validation of regression kriging results

In order to be able to judge the goodness of the interpolation method, the prediction results can be compared with the original dataset. The goodness is evaluating whether the method of achieving results is robust. One way of doing this, is by sub-setting the dataset. A certain amount of the original data is then set aside (called the 'hold-out') and is not used for the prediction modelling. Later this partial data set can be used to compare with the predicted values. This implies that the hold-out is not available during the modelling phase. When the available data are already sparse, this has serious limitations for the modelling itself. An alternative method is cross-validation, where all data can be used for modelling and later are available as well for the validation.

One particular method is called k-fold cross validation, where the 'k' stands for the number of folds one wants to apply. Each fold is a set of data kept apart from the analysis, repeating for the number of folds. In this way, the influence of including or excluding sample points can be investigated, thus establishing the robustness of the method.

A special type of k-fold cross validation is where the repetition of analyses (k) is equal to the number of data. This is called 'leave one out' cross validation, for the analysis is repeated for once for every sample in the dataset, omitting the sample value itself. This leave one out cross validation is used throughout this thesis, where k equals the number of observations in each linear model.

Resulting is a prediction for every observation, made by using the same variogram model settings as for the normal regression kriging prediction. The degree in which the cross validation predictions resemble the observations is then a measure for the goodness of the prediction method. This can be calculated by using the mean squared normalized error or 'zscore' as follows: when the variogram is correct, the computed variance of the zscore needs to be near 1, the mean zscore should approach 0 or be very small. In contrast to standard residuals, the zscore takes into account the kriging variance as it is a standardized residual (Bivand et al, 2008).

To aid further in the assessment of prediction results, additional parameters can be calculated from the cross-validation output, like the mean prediction error (MPE), root mean square prediction error (RMSPE) and average kriging standard error (AKSE).

$$MPE = \frac{1}{N}\sum_{x=1}^{N}(z_{(x)} - z'_{(x)}) \qquad \text{(eq. 8)}$$

$$RMSPE = \sqrt{\frac{1}{N}\sum_{x=1}^{N}[(z_{(x)} - z'_{(x)})]^2} \quad \text{(eq. 9)}$$

Where *N* stands for the number of pairs of observed and predicted values, $z_{(x)}$ is the observed value at location x, and $z'_{(x)}$ is the predicted value by ordinary kriging at location z.

$$AKSE = \sqrt{\frac{1}{N}\sum_{x=1}^{N}[\sigma(x)]^2} \quad \text{(eq. 10)}$$

Where *x* is a location, and $\sigma(x)$ is the prediction standard error for location *x* .

MPE indicates whether a prediction is biased and should be close to zero. RMSPE and AKSE are a measure of precision and have to be more or less equal.

The cv-procedure only accounts for the kriging part, since the input are the residuals from the linear modelling phase. The k-fold cross validation is available in the gstat-package (Pebesma, 2004). MPE and AKSE are calculated using the outcomes of R-packages gstat (*krige.cv()*-procedures) and geospt (the *criterio()*-command).

3.5 Uncertainties

Presenting only maps with prediction outcomes can be misleading when no information is shared on the uncertainty of that outcome. The range of values where an outcome can be found is usually defined as the confidence interval. This interval has an accepted certainty, for instance 95%. The assumption is that the Normal Distribution is valid for the predictions. We can then take the range between the 2.5th and 97.5th percentiles of the normal distribution. The lower and upper boundaries can be found by using the kriging standard deviation σ. The range between these boundaries can be presented as a map. The 95%-confidentiality map can provide some assistance in interpreting the meaning of a regression kriging outcome. In order to read the maps in a meaningful way, they need to be back transformed to the original concentration levels.

For 95% confidence levels, the following values are valid for each cell:

lower boundary : (0.025)percentile of prediction = (prediction - 1.96σ)
upper boundary: (0.975)percentile of prediction = (prediction + 1.96σ)

the 95% confidence interval (the “width”) around the prediction is then defined as:

(prediction - 1.96*σ, prediction + 1.96*σ)

Where ‘σ’ stands for the kriging standard deviation.

As these values are still in 10log-scale, they need to be back transformed to the original value scale. This is a simple 10^{UB} for the upper boundary and 10^{LB} for the lower boundary limit. The 95%width is then the back transformed (UB – LB).

The results of this bandwidth can be presented in a map, which should always accompany the regression kriging prediction.

3.6 Legend color choice

Nitrate values in groundwater are subject to standard levels. A health standard exist for drinking water of 50 mg NO_3/l (WHO, 1998). This standard is reflected in the color choice for the prediction maps: units below 50 are considered 'safe' and therefor are colored in greenish colors, whereas all legend classes above 50 are seen as exceeding the limit, these are colored red or brown. The middle class ranging from 50-100 mg/l is here considered as an intermediate class and therefore has a yellow-orange color. Due to uncertainties, the confidence we have that maps with these predicted classes are true, is not unlimited. It is still possible that a certain area with this or that class can be lower or higher than the map shows. This is where the uncertainty maps fit in.

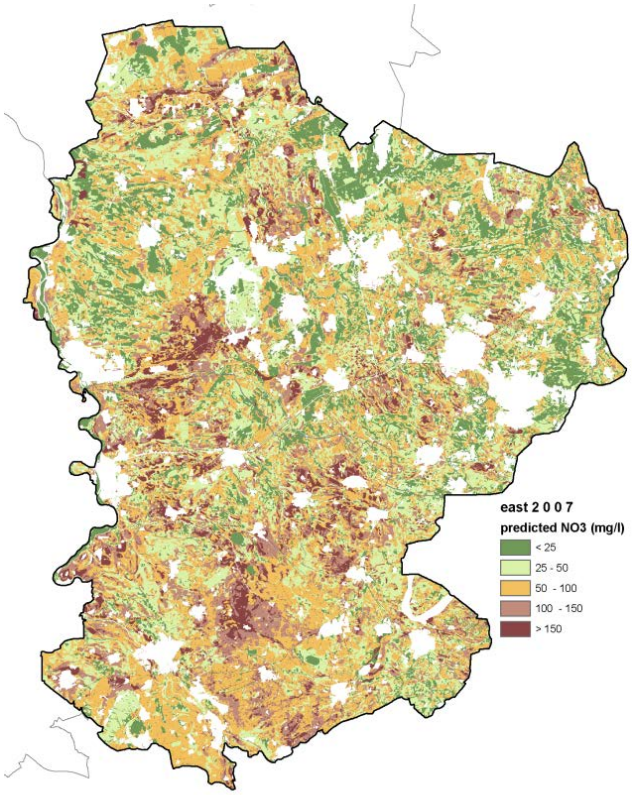


Figure 12. Explanation of legend colour choice, related to the WHO standard for drinking water of 50 mg NO_3 per litre. Map is an example.

The colors for maps with uncertainties and other statistical properties like variances, are different from the predicted value maps. With these maps, colors do not have a warning meaning. For maps without standards, and without the need to define certain class or quantitative intervals, 'High-Low' legends are used.

4 Results

4.1 One elaborated result: region south 2008

To present the results for all models is much of the same; identical steps are performed over and over again. To limit repetition, only for the region south and year 2008 the results are elaborated. The results of the other 11 region-year combinations are listed in appendices IVa and IVb. General outcomes of all predictions are presented in Sections 4.2 and 4.3.

The R-script for this procedure is available for this region in appendix I. The script is more elaborate than the following example here shows. We skip details like cleaning up and dealing with memory issues. Since the south region is the largest grid of all four regions, it was the most difficult one, with regard to the memory use. 16 Gb of RAM memory was barely sufficient to run the model in the current state. Adaptations to the coding may enable easier processing in the future.

4.1.1 Linear model for south.2008

The following description treats the script in appendix I for region south in 2008. The general steps are explained here, while in the script some annotation is available as well.

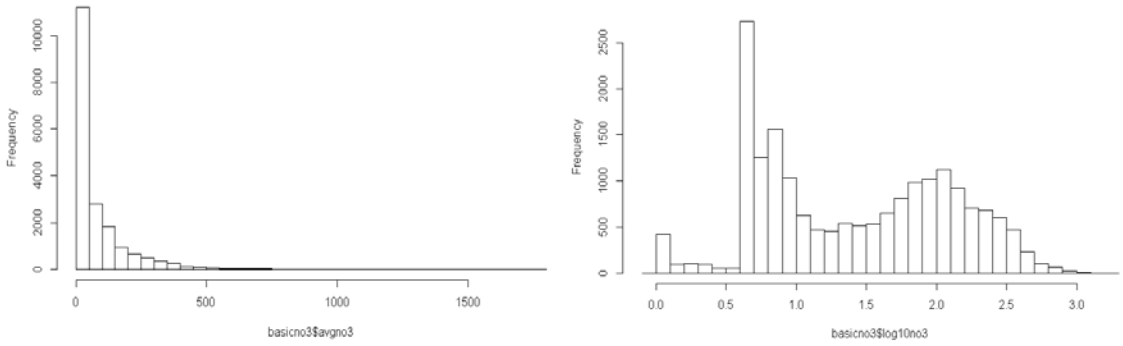


Figure 13. Measurement data before (left) and after transformation (right).

First, the sample data file is loaded and screened for duplicates. Since the target variable is required to have a normal distribution, transformation is necessary (Burrough & McDonnell, 1998; Webster & Oliver, 2007).

The target variable has many observations in the low range of measurement values, and few with very high values. Therefore, a transformation is needed to convert the variable to a more suitable distribution. The sampled NO_3 is transformed using 10log transformation (Figure 13). Note that the result is still not normally distributed. This is not a problem, since it is only the residuals from the linear regression that we are interested in. The dataset is split into the three consecutive years (2007-2009).

Using stepwise multiple linear regression, the linear model for region south in 2008 is found, composed of the following covariates:

```
om05 + om10 + om40 + om60 + gt06 + stone5 + stone6 + stone7 + nhx + bbg06 +
kwel2 + lgn6 + geom + slaf + draf
```

Each of the selected covariates now has a β-value, but since the factor variables `gt06`, `bbg06`, `lgn6`, `geom`, `slaf` and `draf` have many categories, the listing of these and their

significance can be found in appendix VI (verbose model summary). The model is based on 2523 data points and has an R^2 of `0.594` and an adjusted R^2 of `0.584`.

Next, some diagnostics can be plotted (Figure 14 and Figure 16).

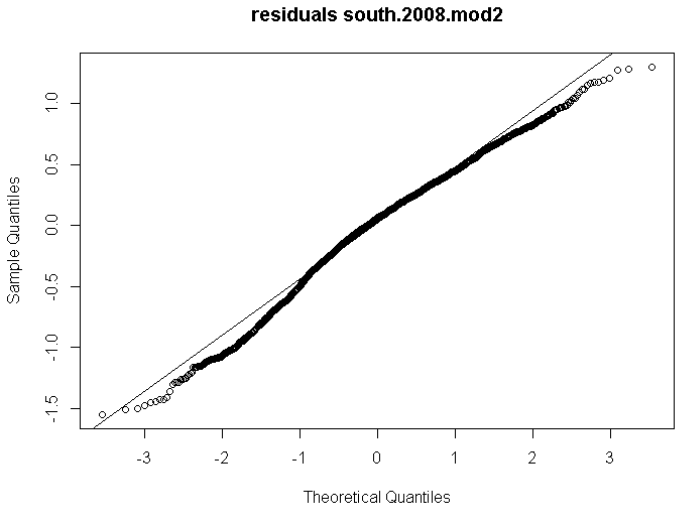


Figure 14. QQ-plot(left) for the outcomes of the `south.2008` model

The QQ-plot (Figure 14) indicates that the residuals have what can be considered a 'normal' distribution: they follow the straight line pretty well overall, are more dense in the middle and have fewer points at larger and smaller observations.

In Figure 15 (left), the histogram of the (10log) NO_3-values is given. Notice that the distribution is almost bimodal, and still not very symmetrical even after transformation. The right part of this figure shows the frequency histogram for the residuals of the linear model prediction `south.2008`. Now there is a symmetrical, almost normal distribution which can also be concluded from Figure 14. This means that one of the prerequisites for ordinary kriging (a normal distribution) is now fulfilled, and we can continue with the kriging calculation.

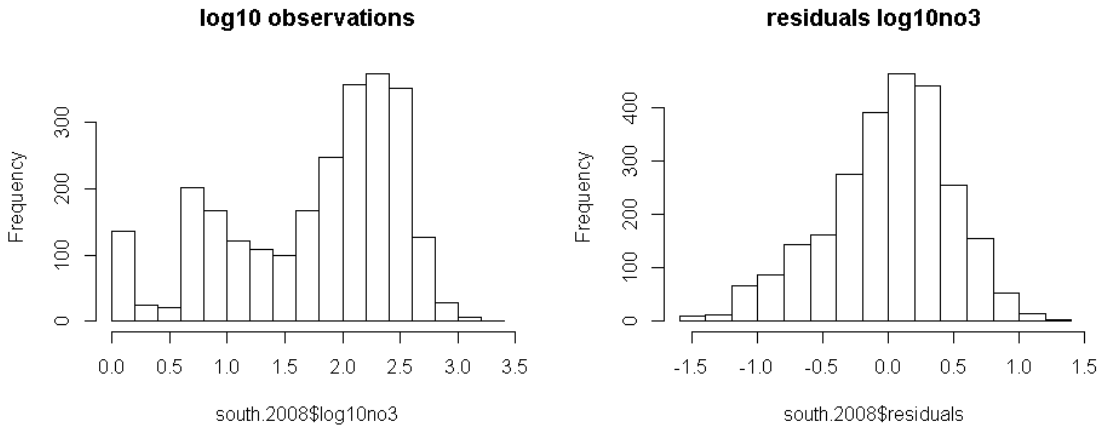


Figure 15. Histograms of (transformed) observations (left), and predicted residuals (right).

In the predicted values against the observed values-plot (Figure 16), two separate horizontal lines are visible, around 0 and around 0,75. This can be traced back to the two separate monitoring networks, described in Section 2.1. When the origin of the sample data is added as a differing colour, this becomes more evident.

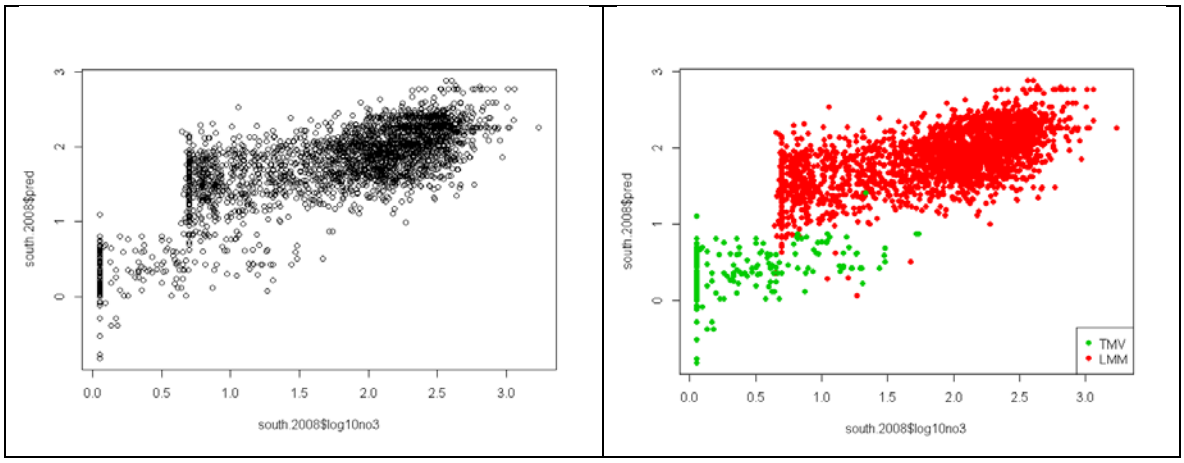


Figure 16. Plot of predictions vs (transformed) observations for south.2008. Left: all data, black and right: arranged by monitoring network. Green=TMV, red=LMM.

In the script x,y-coordinates are not considered as variables, but the influence on the model composition was examined for every model. It turned out that the south model had a (very) small increase with decreasing y-parameter, suggesting that target values increase when going south. However, the fit of the model got worse with the inclusion of y, so it was left out altogether.

Next, the model can be used to predict the residual values for unsampled locations. For this purpose, a prediction mask was constructed:

`Original covariate data → split into two separate dataframes:`

1. `df.ON : all valid locations on which prediction should take place`
2. `df.ON.NA: the remaining non-valid locations, no prediction`

For each of the selected covariate grid layers, this process is repeated until all 'NA' values are processed. The resulting locations in the `df.ON` dataframe are now used to predict with the linear model, after which the remaining grid cells with 'NA' are united again with the prediction data. After re-ordering of x and y coordinates, the first (regression) prediction surface is ready. For the same locations, now the kriging procedure is started. This is described in the next paragraph.

4.1.2 Experimental variogram for south.2008

Once the regression prediction has been performed, the variogram for the resulting residuals from the sample data can be modelled, using `gstat`. This concerns the transformed data. First, some initial settings are tried, by choosing a variogram model and fitting the appearance visually. Next, automated settings are applied by using `fit.variogram`. This results in Figure 17, where an exponential model is fitted. Note that the displayed range of the variogram is limited in the figure, but that various settings were used to check the behaviour at longer distances (eg.5 km), at which no difference was found in the fit of the variogram. The numbers at the data points are the actual point pairs available at the corresponding distances.

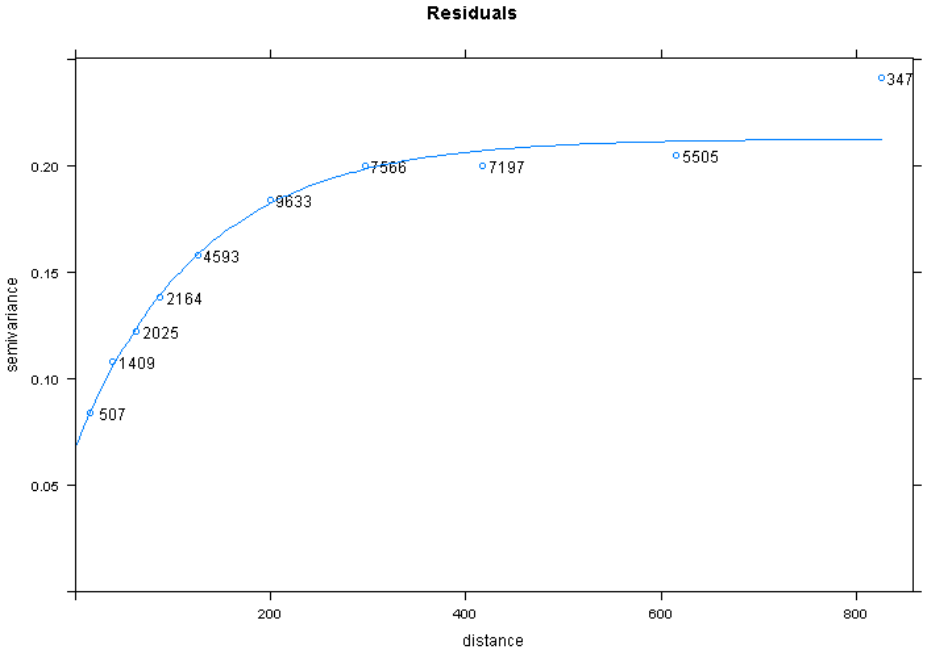


Figure 17. Experimental and fitted variogram for south.2008. Parameters are Vgm.res= 0.068 Nug(0) + 0.145 Exp(128.05). Numbers in the plot indicate the number of point pairs at each point.

At this stage, the simple kriging of the modelled residuals begins. Using the same locations from the first prediction surface as a prediction mask, the fitted variogram is input in the kriging equation. Now for each of the valid target cells, the kriging equation calculates the kriging predictions, based on the residuals from the sample data. The number of points to consider for the interpolation is set at (nmax=) 100. The outcome of the kriging predictions can now be added to those from the linear regression prediction, resulting in Figure 18 (top). The kriging variance is produced together with the kriging operation and is shown in the bottom part of the same figure.

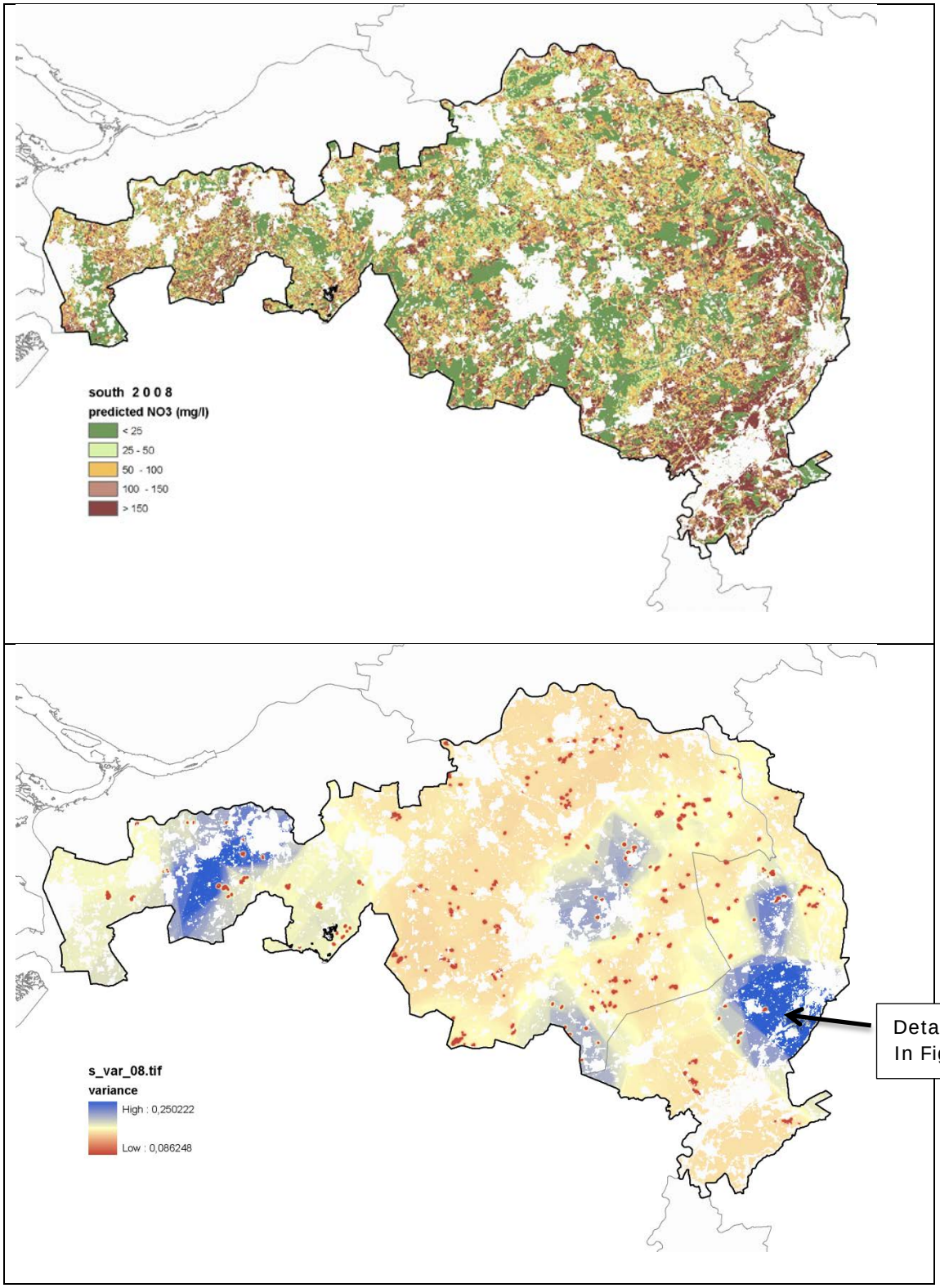


Figure 18. Regression kriging results (top) and (untransformed) variance (bottom) of the residuals from the south.2008 model. The red "dots" indicate where the original measurements were taken.

The variance results from Figure 18 have some artefacts. It can be expected that values close to the locations where point samples were taken, have lower variances. However, the blue coloured regions appear very strange here, especially when other sparsely sampled regions do not have this blue but yellow and orange colours, indicating a lower variance

value. Though in the prediction map the blue areas correspond somewhat with higher predictions (brownred, >150 mg/l), this is not reversely so for the other regions. A possible explanation can be found in the configuration of, or the number of points considered for the kriging prediction (100). When zoomed in, the point cloud at the detail inset reveals that this is a genuine hotspot. There are 96 points located here, each having high nitrate values (range 200-300 mg/l). These 96 points can account for 96/16 = 4 farms bordering each other, but unfortunately the amount of time to investigate this effect is too limited. In Figure 19 this phenomenon is enlarged, showing the scale at which the variance is increasing, just around a cluster of sample points. The cell dimensions in the image are 25 x 25 m. Recall from the variogram in Figure 17 that the practical range is around 500 meters. This corresponds with the decrease of the variance seen in the image. Similar point clusters appear in the other blue regions of Figure 18. It does not explain, however that the other regions with yellowish colours in Figure 18 have smaller variances, when there are no points located.

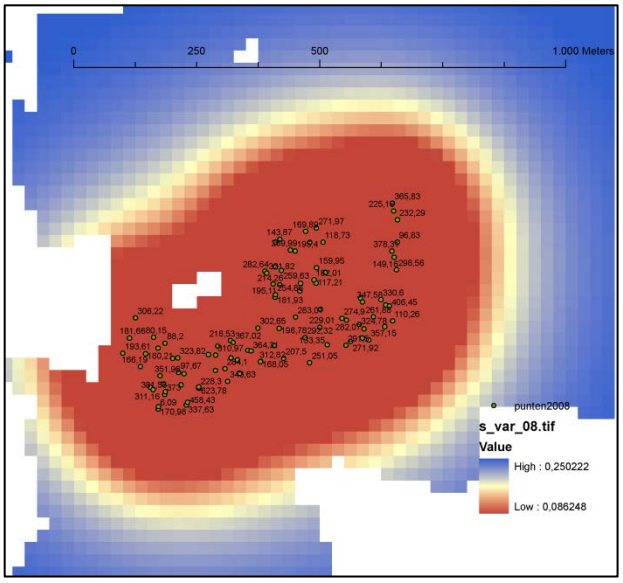


Figure 19. Detail of Figure 18, showing the range of variance around sample points (in green, with NO_3-value).

For the variance received from ordinary kriging, no back transformation is possible in the original units (mg NO_3 per litre) since the true spatial mean μ is unknown (μ was estimated in the prediction). Therefore Figure 18 shows the variance in 10log(no3). Clearly visible in the variance plot are the zones around the sampled locations, having the lowest variance. This points at the predominantly local effects of the kriging operation, which can also be determined from the relatively short range of the variogram in Figure 17.

4.2 Generated regression models for all regions and years

Models for twelve year-region combinations were constructed by means of stepwise multiple linear regression analysis. The model composition is shortly presented per region in Sections 4.2.1 to 4.2.6. For each chosen model the coefficient of determination, R^2 and R^2_a (adjusted R^2) are listed. R^2-values indicate the amount of variance explained by the model, R^2_a is the R^2 value that has been adjusted for the number of variables that were used in the equation (Hengl et al., 2004).

In `north.2008` and `north.2009` models, the `gt06` covariabele was dropped initially by the stepwise regression model selection function. Since these two models are the only ones without automatic inclusion and the `gt06` map serves another purpose as well (basic grid map for predicting values), a manual addition to the model was decided on. By using an additional argument in the stepwise regression selection (`scope = list(lower = ~ gt06)`) it was made sure this resulted in balanced models. Highlighted covariates are appearing in each of the three year-region models, per region.

4.2.1 Linear models for region North

Table 6. Model summary for region north. Highlighted variables occur in all three models.

Model name	Parameters	#data points	R^2	R^2_a
north.2007	om60 + om100 + om120 + gt06 + bbg06 + gronds + kwel2 + pawn + lgn6 + laf + slont	1983	0.493	0.473
north.2008	om60 + om100 + gt06* + ahn + bbg06 + kwel2 + pawn + lgn6 + lont + slaf	2167	0.504	0.488
north.2009	stone7 + stone8 + gt06* + ahn + bbg06 + kwel2 + lgn6 + laf + slont + vds	1979	0.390	0.372

*forced inclusion

In the north models `gt06`, `bbg06`, `kwel2` and `lgn6` occur every year. `ahn` and `pawn` occur twice. Two hydrological parameters are present in each year, but not the same. Looking at both the composition and the coefficients of determination R^2 and R^2-adjusted, for 2007 and 2008 are quite similar. In these years almost 50% of the variance was explained by the models. Looking at the gt06 covariate, the significance is high (for almost all categories in gt06) when looking at the forced instances, and almost not significant in 2007, when the variable was added by the stepwise method itself. The significance of covariates can be found in the verbose model listings, appendix VI.

4.2.2 Linear models for region East

Table 7. Model summary for region east. Highlighted variables occur in all three models.

Model name	Parameters	#data points	R^2	R^2_a
east.2007	om10 + om40 + om80 + om120 + gt06 + bbg06 + kwel2 + geom + draf + slont	1364	0.331	0.310
east.2008	om10 + om60 + om80 + om100 + gt06 + stone6 + stone8 + lgn6 + draf	1180	0.307	0.288
east.2009	om25 + om80 + om100 + gt06 + ahn + kwel2 + lgn6 + draf	1249	0.266	0.248

In the models constructed for the east region, only om80, gt06 and draf occur in each year model. Organic matter is present from different depth layers in all years. There seems to be a sequence from 2007 to 2009: 2007 and 2008 are alike, so do 2008 and 2009 have similarities, but not so much for 2007 and 2009. R^2 parameters are among the lowest, compared with the models for the other regions.

4.2.3 *Linear models for region Centre*

Table 8. Model summary for region centre. Highlighted variables occur in all three models.

Model name	Parameters	#data points	R^2	R^2_a
centre.2007	gt06 + stone5 + stone6 + bbg06 + geom + vds	452	0.584	0.558
centre.2008	gt06 + geom + nhx	379	0.452	0.426
centre.2009	om10 + gt06 + stone5 + stone6 + stone8 + bbg06 + laf + vds	348	0.529	0.489

The models in the central sand regions are somewhat different from those constructed for the other regions. The 2008-model has only three parameters left after the stepwise regression. Since one of them (geom) is not selected in the last model, only gt06 occurs in all three of them. There is a variety of parameters in use in the three different years and none of the models look really similar. This region shows one of the highest R^2/adj R^2 values of all models (2007), but this is probably linked to the compactness of the sample cloud, with little extremities. The Nitrogen-addition maps (stone) have a strong presence, save in 2008. This could be due to the contrast in land use in this region, which is also visible in the stone maps.

4.2.4 *Linear models for region South*

Table 9. Model summary for region south. Highlighted variables occur in all three models.

Model name	Parameters	#data points	R^2	R^2_a
south.2007	om05 + om40 + om60 + gt06 + nhx + ahn + bbg06 + kwel2 + lgn6 + geom	1603	0.493	0.481
south.2008	om05 + om10 + om40 + om60 + gt06 + stone5 + stone6 + stone7 + nhx + bbg06 + kwel2 + lgn6 + geom + slaf + draf	2523	0.594	0.584
south.2009	om05 + om10 + om25 + om40 + om60 + gt06 + nhx + kwel2 + pawn + lgn6 + geom + slaf + slont	2137	0.417	0.398

The coefficient values and signs (+ or -) for each variable can be found in the appendix (VI, verbose model summary) listing, because of the number of classes in the categorical variables these are too long to present here. In the three models for the south region that were defined by the stepwise regression, several common parameters occur. These are highlighted. From the 8 possible organic matter map variables, three are always present (om05, om40 and om60). Other parameters that were present in all three models were gt06, nhx, kwel2, lgn6 and geom. In the model for 2008, the variables for added nitrogen(fertilizer and manure) stone5, stone6 and stone7 are included, where they are absent in the other two models. The hydrological parameter slaf is present in two models, just like bbg06 (statistical landuse). The best fit, as judged by the adjusted R^2, is generated by the south.2008 model with a value of 0.584. This model also contains the most data points.

4.2.5 *Covariable ranking for regional models*

Table 10. Rank table of selected covariates, based on regional models

Rank	Name of covariabele	# times selected
1	gt06	12 (2 x forced)
2	bbg06, lgn6, kwel2	8
3	om60, geom	6
4	om10	5
5	om40, om100, stone6, ahn, nhx, slont, draf	4
6	om05, om80, pawn, stone5, stone8, laf, slaf, vds	3
7	om25, om120, stone7	2
8	lont, gronds, om05	1
9	dront	-

Since twelve models have been made, for 4 regions and 3 different years, it may be interesting to see which covariates were selected, and how often. The ranking table is based on selection by the stepwise procedure. Absolute winner is the covariate with groundwater tables gt06, with presence in all twelve models. In two of these models the parameter was first rejected automatically however. The covariables with land use properties bbg06 and lgn6 rank at a high second place.

This ranking does no justice to the significance of the parameters. It is also difficult to compare categorical variables with up to 21 levels to continuous variables. The significance of the parameters can be found in the verbose model summary, indicated by stars and dots (*** for highly significant, to (' ') or no indication for the lowest significant level).

Not once selected in the regional models was 'dront', being the drainage resistance at profound depth. Only once selected was the simplified soil map 'gronds', whereas the other soil maps vds and pawn were selected 3 times each.

4.2.6 *Linear models nation wide*

Table 11. Nationwide model summary. Common variables are shaded.

Model name (.mod)	Parameters	#data points	R^2	R^2_a
all.regions.2007	om10 + stone5 + stone6 + nhx + gt06 + ahn + bbg06 + gronds + kwel2 + pawn + lgn6 + geom + dront + laf	5383	0.472	0.462
all.regions.2008	om60 + om80 + om100 + stone5 + stone6 + stone7 + stone8 + nhx + gt06 + ahn + bbg06 + kwel2 + pawn + lgn6 + geom + slaf + slont	6243	0.554	0.547
all.regions.2009	stone5 + stone6 + stone7 + stone8 + nhx + gt06 + ahn + kwel2 + pawn + lgn6 + geom + dront + slaf + vds	5721	0.435	0.426

Common parameters in the all.regions models are stone5 and stone6, nhx, gt06, ahn, bbg06, lgn6, kwel2, pawn and geom. The covariate maps with animal manure/fertilizer N are important and except for the first year 2007, all four of these stone-maps are included. Organic matter layer maps are present in two of the three years models. A combination of two hydraulic parameter maps seem to be an integral part of the models too, but the same set never occurs twice. All.regions.2008 appears to be the best model, as judged by the R^2/adj-R^2 statistic. The other two have similar scores.

A ranking is not made, since only three models were made. Not selected covariates include om05, om25, om40, om120, lont and draf. Compared to the regional models, the three nationwide models are based more often on the stone-maps and nhx parameter. In the regional models, nhx occurs almost only in the three south models. The layer maps with organic matter are not so popular in the nationwide models as they are in the regional models, only three out of eight available depth layers were included. The digital elevation model ahn is also present in all three nationwide models, were this is true for the regional models only in 4 of the 12 cases.

4.3 Combined prediction results from regression and kriging

From the final prediction images (the sum of the regression prediction and the kriging prediction) some statistics can be generated. These can be used to check for unusual or unrealistic outcomes, but also to compare with the original features of the sample data set, as given in Table 4 of Section 2.3.

Regional models

*Table 12. Mean, median, sd, and min/max for **regionally** predicted values of NO_3.*

NORTH					
year	**mean**	**median**	**sd**	**min**	**max**
2007	45.8	27.3	51.7	0.02	987
2008	33.8	21.5	36.7	0.01	824
2009	27.4	16.9	35.3	0	1284
EAST					
year	**mean**	**median**	**sd**	**min**	**max**
2007	74.0	54.7	232.1	0.01	43061
2008	90.6	34.3	130.3	0	3499
2009	43.0	31.9	71.0	0.02	8737
CENTRE					
year	**mean**	**median**	**sd**	**min**	**max**
2007	34.2	11.1	88.8	0.11	3229
2008	27.5	19.7	24.0	0.61	185
2009	16.5	7.1	20.7	0.07	522
SOUTH					
year	**mean**	**median**	**sd**	**min**	**max**
2007	88.0	66.9	91.3	0.01	3408
2008	87.4	51.8	120.9	0	3088
2009	95.1	44.3	14777	0	22392424

Shaded: unlikely extreme values, probable prediction artefacts.

In Table 12, the properties for the predicted values are presented per year and region, each of these being the result of one of the twelve unique models. Some of the prediction results show large maximum values, for instance south.2009. This maximum value is not a very realistic outcome and must be attributed to an unstable factor, maybe a singularity in the measurement locations. The highest predictions are much higher than those present in the sample data set. This is true for all almost all separate regions, save the moderate outcomes for the centre and east region in 2008 (lower than the maximum in the sample data in 2008). The predicted mean and median range of values are generally comparable to those in the sample data set.

Nationwide models

*Table 13. Mean, median, sd, and min/max for **nationally** predicted values of NO_3. Calculations were performed per region, using nationwide predicted residuals.*

all.regions N O R T H					
year	**mean**	**median**	**sd**	**min**	**max**
2007	47.5	30.8	50.0	0.01	962
2008	35.1	22.0	38.0	0.03	811
2009	27.5	18.0	27.8	0.10	516
all.regions E A S T					
year	**mean**	**median**	**sd**	**min**	**max**
2007	67.5	53.3	58.4	0.15	1295
2008	51.7	36.1	53.6	0.05	854
2009	39.1	26.7	40.7	0.14	851
all.regions C E N T R E					
year	**mean**	**median**	**sd**	**min**	**max**
2007	29.3	9.53	46.1	0.08	872
2008	20.4	7.0	32.7	0.09	619
2009	21.4	11.5	29.5	0.13	689
all.regions S O U T H					
year	**mean**	**median**	**sd**	**min**	**max**
2007	85.2	61.3	88.7	0.06	1355
2008	83.4	51.2	103.9	0.03	2503
2009	67.9	39.0	86.7	0.17	2600
All.regions combined					
year	**mean**	**median**	**sd**	**min**	**max**
2007	61.8	40.4	69.0	0.01	1355
2008	52.6	29.1	72.6	0.03	2503
2009	42.6	23.7	59.3	0.10	2600

The results for the three nationwide models are given in Table 13. Since the regions were calculated separately, but with the same model and data for each year, results per region can be compared with the regional model results. The nationwide models yield combined results also. This is not possible for the regional model results in Table 12, since the models were different for every region.

Maximum predicted values for the nationwide models are without unrealistic high values, like in the results for the regional model predictions. They reflect the range of values in the sample data set (found in Table 4). The median value results of all.regions combined, match the trend and range of values from the sample data reasonably well. The minimum predicted values are always lower than the lowest sampled value of 1.02.

Compared to the regional predictions, predicted mean and median values are generally somewhat lower per region, with the exception of the north region. Here the all.regions combined prediction are higher than those of the regional models for this area.

4.4 Cross validation of kriging results

Cross validation was used to obtain the results of Table 14 and Table 15. In addition, for each cross-validation result the MPE, or mean prediction error, was calculated. The MPE-value should be close to zero. RMSPE (root mean square prediction error) and AKSE (average kriging standard error) are given as well. The latter two error values should approximate each other, indicating prediction stability.

*Table 14. LOO-Cross validation zscore and MPE results for the regional models**

region	year	var(zscore)	mean(zscore)	MPE	RMSPE	AKSE
North	2007	1.0440	-0.0144	-0.006095	0.4177	0.4075
	2008	0.9357	-0.0147	-0.006646	0.4007	0.4137
	2009	1.1074	**-0.0030**	-0.001479	0.4053	0.3833
East	2007	1.0186	-0.0097	-0.003652	0.4241	0.4197
	2008	0.9656	-0.0120	-0.004866	0.4508	0.4589
	2009	1.0143	-0.0210	-0.009602	**0.4251**	**0.4222**
Centre	2007	1.1456	-0.0114	-0.006578	0.3583	0.3365
	2008	1.0802	-0.0094	-0.004537	0.3963	0.3815
	2009	1.1449	-0.0122	0.004637	0.3444	0.3162
South	2007	**0.9829**	-0.0071	-0.003509	0.4161	0.4201
	2008	1.0349	-0.0041	-0.001445	0.3892	0.3801
	2009	1.0245	-0.0031	**-0.000783**	0.4318	0.4264

*best individual score in bold; RMSPE & AKSE: smallest difference

Comparing the scores for the models, based on the three CV-parameters is difficult, since all models score reasonably. It can be noted that all three centre models have a high var(zscore) and for 2007 and 2008 somewhat worse mean zscores. The mean prediction error (MPE) results for these models on the other hand, are among the lowest values found. The RMPSE and AKSE errors are lower than for the other regional models. When examining the data cloud for the models in appendix VIb, the centre region point clouds have less data and appear less clustered when compared to the other regions. The best model, taking the three parameters into account seems to be one of the south models, since they score quite well in the zscore values.

*Table 15. LOOCV-zscore and MPE results for the nationwide models**

	year	var(zscore)	mean(zscore)	MPE	RMSPE	AKSE
all.regions	2007	0.9411	**-0.0054**	**0.00102**	0.4165	0.4288
all.regions	2008	0.9692	-0.0065	-0.00282	0.4083	0.4131
all.regions	2009	**0.9796**	-0.0102	-0.00454	**0.4237**	**0.4278**

*best individual score in bold; RMSPE & AKSE: smallest difference

Since the nationwide models 'all.regions' were actually calculated in separated regions, for each of these the scores can be calculated as well, but as the input and models were the same, only the values of Table 15 are unique. The cross-validation results are very similar, but the results for 2008 seem slightly better overall, depending on which value is judged.

4.5 Year comparison of one region for 2007, 2008 and 2009

In Section 4.1, the method for calculating the predictions for the south region in 2008 was explained. In this section, the outcomes of the predictions in region south for three consecutive years, 2007-2009 are presented and compared.

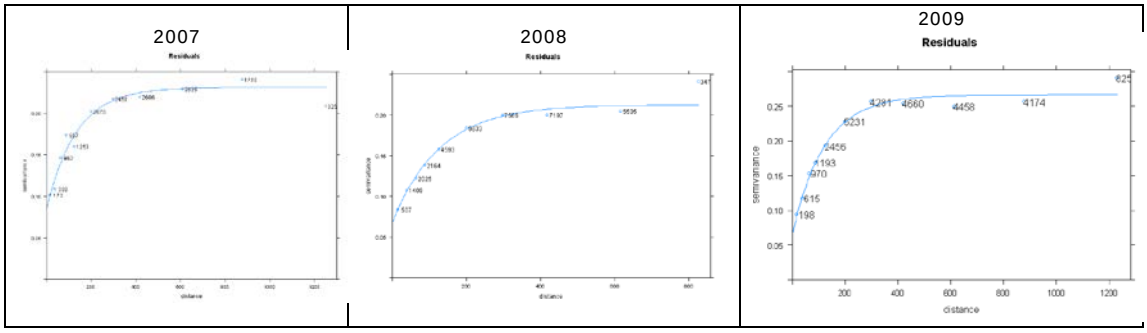
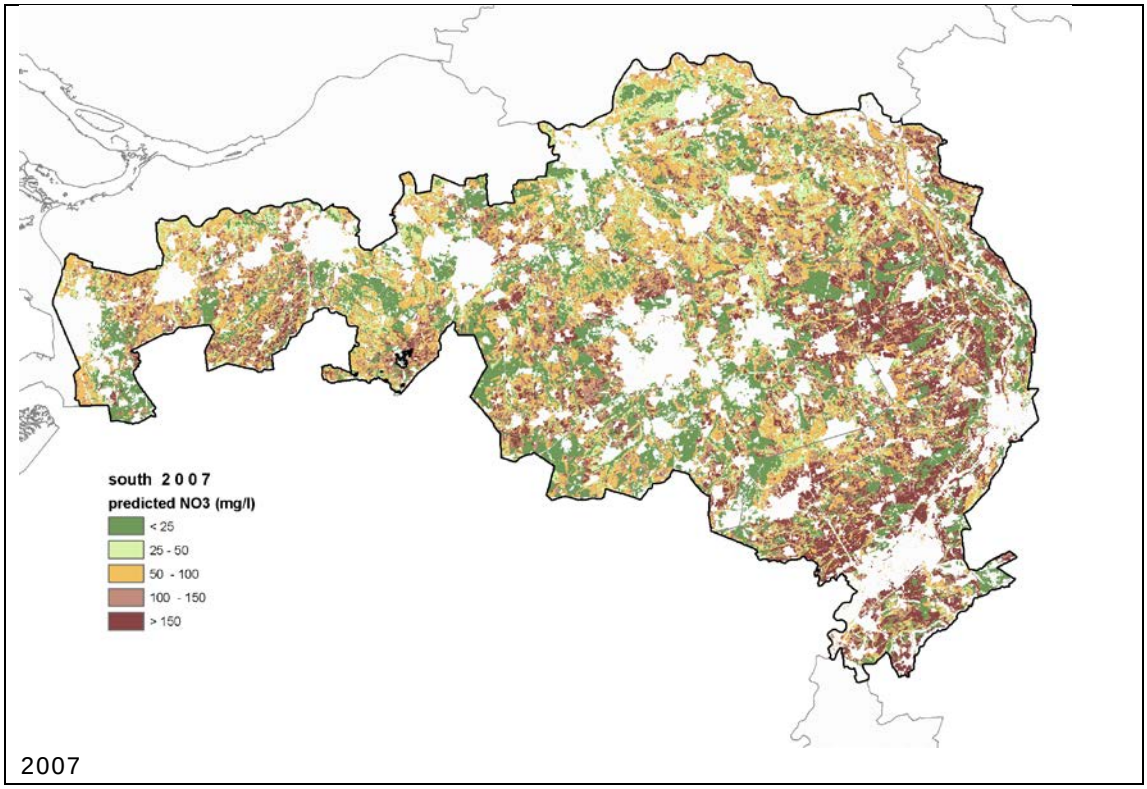


Figure 20. Variograms of three different years for region south.

The variograms are very similar, no different behaviour is to be expected from the kriging prediction. The difference is in the data values, data configuration and the model selection of covariates, as was presented already in Section 4.2.4.

In Figure 21, the regression-kriging predictions are displayed for the three different years. They reflect the small changes in the model selections: some white spots appear where they have predictions in one of the other maps, except for common no-data values like cities and infrastructure.



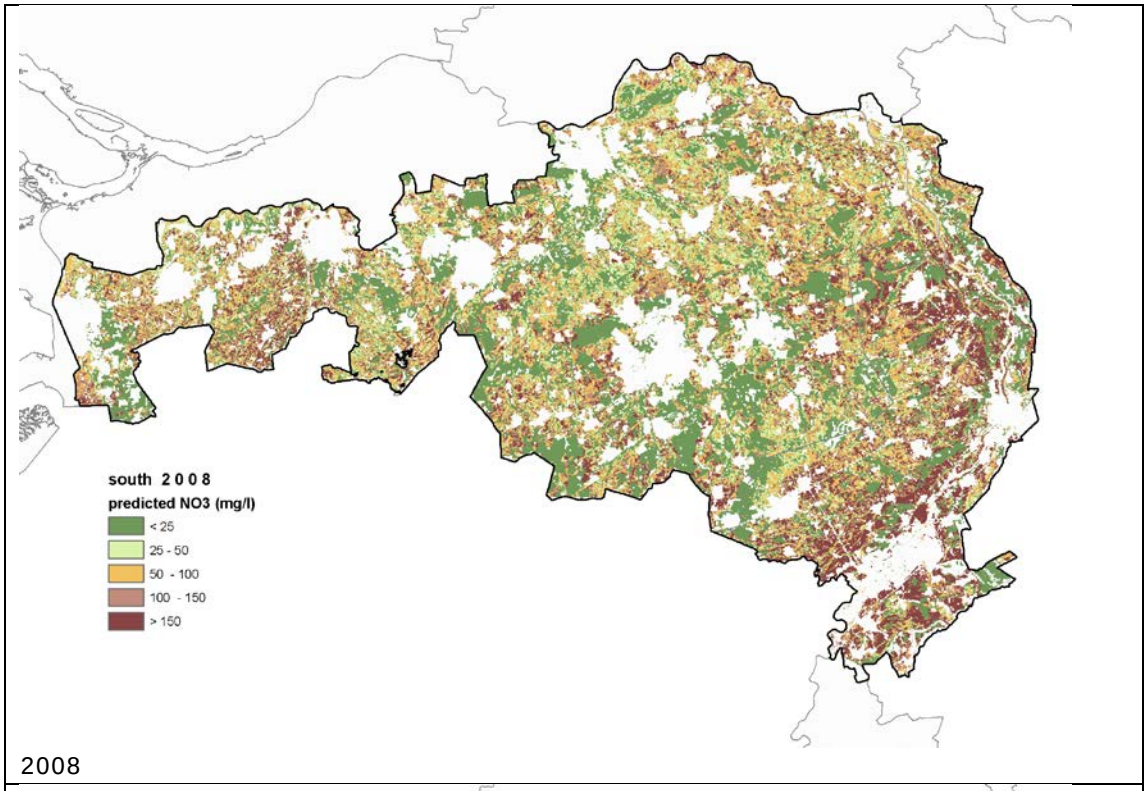
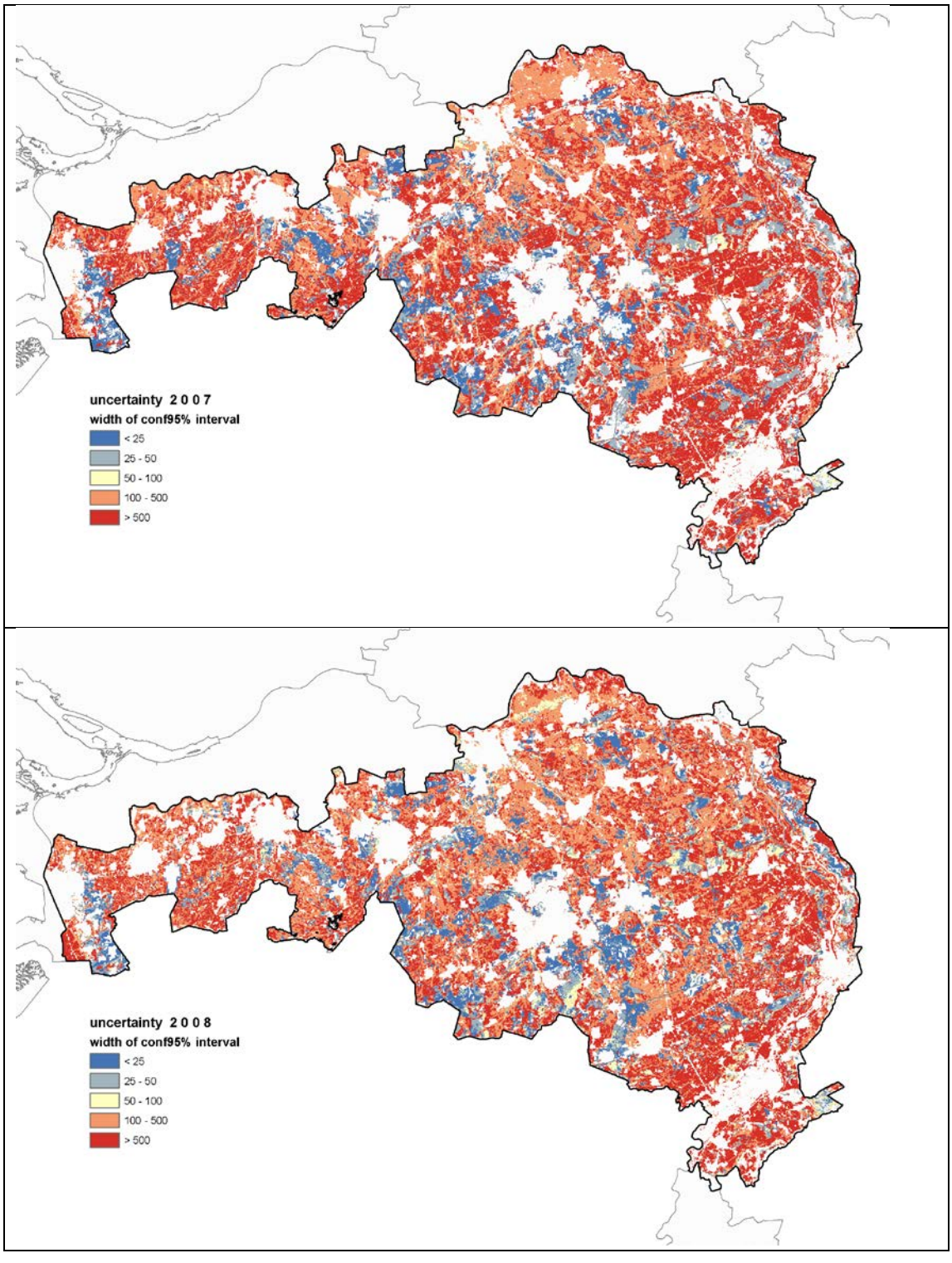


Figure 21. Prediction results for region South in 2007, 2008 and 2009.

4.6 Uncertainties of the results for 2007, 2008 and 2009

In order to judge prediction outcomes, it is necessary to define some way of uncertainty ranges. This can be done by calculating the upper and lower confidence limits of the predicted value. The distance between these two limits is then the range in which the predicted value is sure to be found (with a 95% confidence in this case). It is suggested that these maps always need to be studied when using the prediction results


uncertainty 2007
width of conf95% interval
< 25
25 - 50
50 - 100
100 - 500
> 500
uncertainty 2008
width of conf95% interval
< 25
25 - 50
50 - 100
100 - 500
> 500

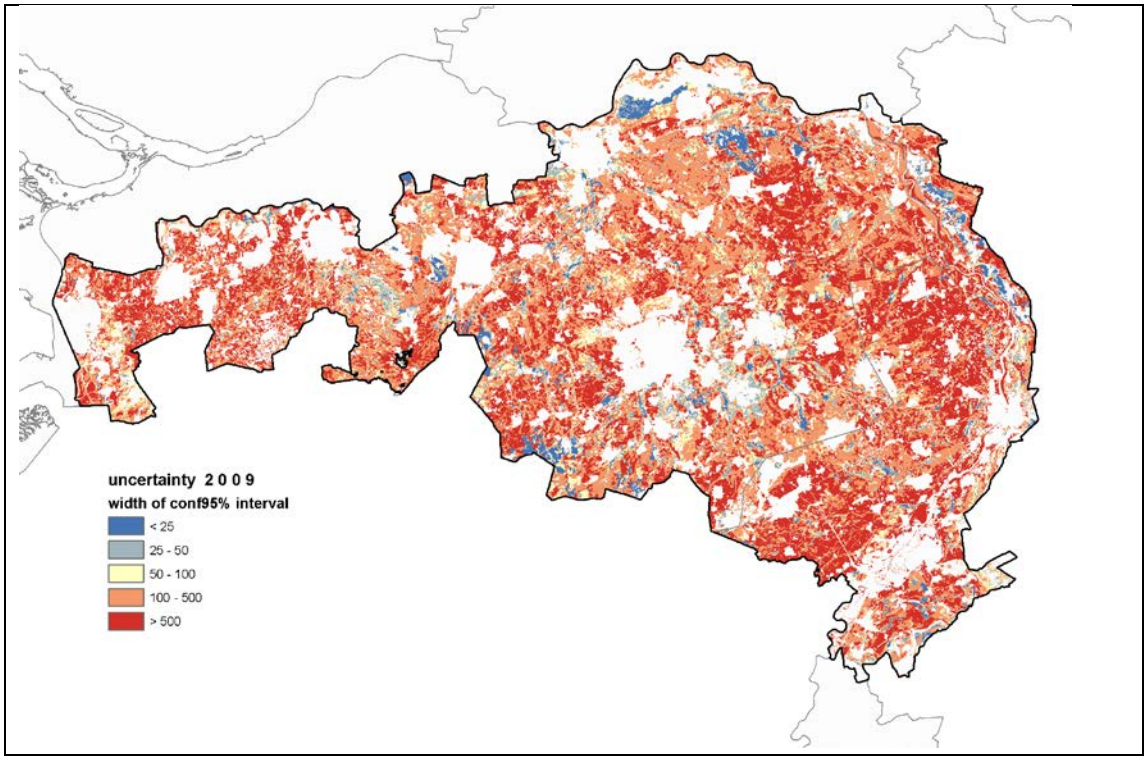


Figure 22. Uncertainty maps for the three South region models, giving the range of values in which a predicted value can be found with 95% confidence.

When looking at the predicted values for a location (for instance from Figure 21), studying the confidence class at the same location, for the corresponding year, gives an indication of the probability that the predicted value is indeed in the predicted range of results. From Figure 22, most of the surface is within the reddish colour classes. This indicates that the prediction results have a 95% certainty to range at least from 100-500, and probably more. The blue colour defines the narrowest class of confidence, where deep red means the widest class. It seems the narrow blue class value gains more weight in 2008, dropping in 2009 again. This is also the case for the extreme wide class (in dark red), achieving more presence in 2008, and with a decrease in favour of a narrower class in 2009. When comparing 2007 with 2009, the overall impression is that darker red is turning towards a more orange/yellow appearance. This would point towards a decreasing uncertainty between 2007 and 2009. The narrower dark blue class however has decreased presence in favour of yellow and light red, balancing the overall uncertainty as neutral compared to previous years.

4.7 Comparison of a regional model prediction with a nationwide model outcome

To evaluate the effects of RK-modelling for the same region with regional data only and that of combined nationwide modelling, the results of the region itself have to be compared. In this section, the region south from the model `south.2008` is compared with `all.regions.2008`, for just the southern part. The same exercise could be repeated for any of the other three regions, but are put only in the appendices.

Data from the year 2008 is used, first only regional data in a regional model (`south.2008`), then the data from all sandy regions are used to predict the same extent (south) but in a combined 'nationwide' model, `all.regions.2008`. What are the differences in model selection, and what can be said about the model accuracy? In Figure 23 the prediction outcome of both models is presented, in order to compare qualitatively. At general first view they seem equal, but small differences can be noted. For instance, the locations of the really green areas differ, some white unpredicted areas are different in both images.

Since the model composition is not the same, and the absence of a covariable (map) may lead to holes in the prediction maps, it is likely that white spots would occur in of the images, whereas they would not in the other. It seems however that both models predict generally at the same locations.

Table 16. The linear model selections for the regression prediction phase, for south.2008 and all.regions.2008.

model	included covariate
(regional) south.2008	om05 + om10 + om40 + om60 + gt06 + stone5 + stone6 + stone7 + nhx + bbg06 + kwel2 + lgn6 + geom + slaf + draf
(combined) all.regions.2008	om60 + om80 + om100 + stone5 + stone6 + stone7 + stone8 + nhx + gt06 + ahn + bbg06 + kwel2 + pawn + lgn6 + geom + slaf + slont

The linear models from the two approaches for south in 2008 are sharing 11 covariates. Most obvious difference is the preference of the regional model for the superficial organic matter layers, while in the combined nationwide model the deeper layers with organic matter are included, next to elevation (ahn) and a soil relation (pawn).

To see real differences, the RK-prediction results need to be subtracted from each other. This has been done in Figure 24, only for grid cells where in both results there are predictions. Now the differences are pronounced: blue means that the regional prediction was higher than the combined regions model prediction, while yellow to red areas indicate where this is just the opposite. The range of differences is not very large and the largest differences are also found in regions with high predictions (the dark blue and in the absolute difference map coincides with that in the prediction maps). Most of the surface is in the two lower classes surrounding zero difference. Overall, by eyeball, the blue colours are more present. This points at the regional model predicting higher values.

In the next two tables we will compare at a few calculated parameters. In Table 17, the model results are compared quantitatively also, using the cross-validation results. The scores are very close and no apparent winning approach is visible. This means both model predictions are comparably stable.

Table 17. Comparison of CV-values for regional and nationwide model

2008, south	var(zscore)	mean(zscore)	MPE
Regional	1.0349	-0.0041	-0.00145
Nationwide	0.9692	-0.0065	-0.00282

Table 18. mean, median, sd and min/max values for RK-prediction results for a national and regional model, south 2008

model	mean	median	std. dev.	min	max
Regional	87.42	51.8	120.86	0	3088
Nationwide	83.35	51.2	103.93	0.03	2503

The mean, median and maximum of the predicted target value for NO_3 are slightly higher for the regional model (Table 18). The regional model predicts somewhat higher values for all parameters. This could be different for other region-all.region comparisons though.

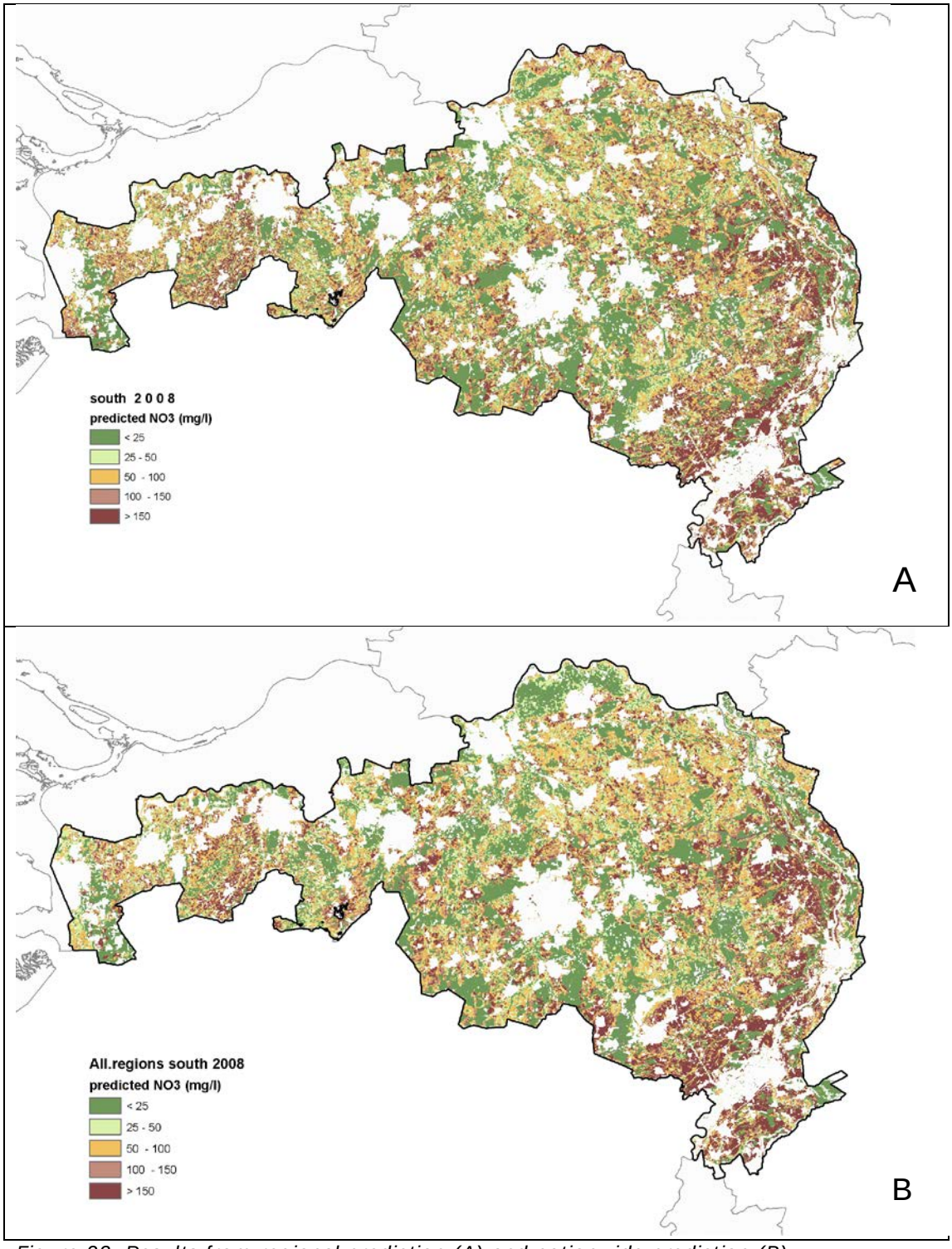


Figure 23. Results from regional prediction (A) and nationwide prediction (B)

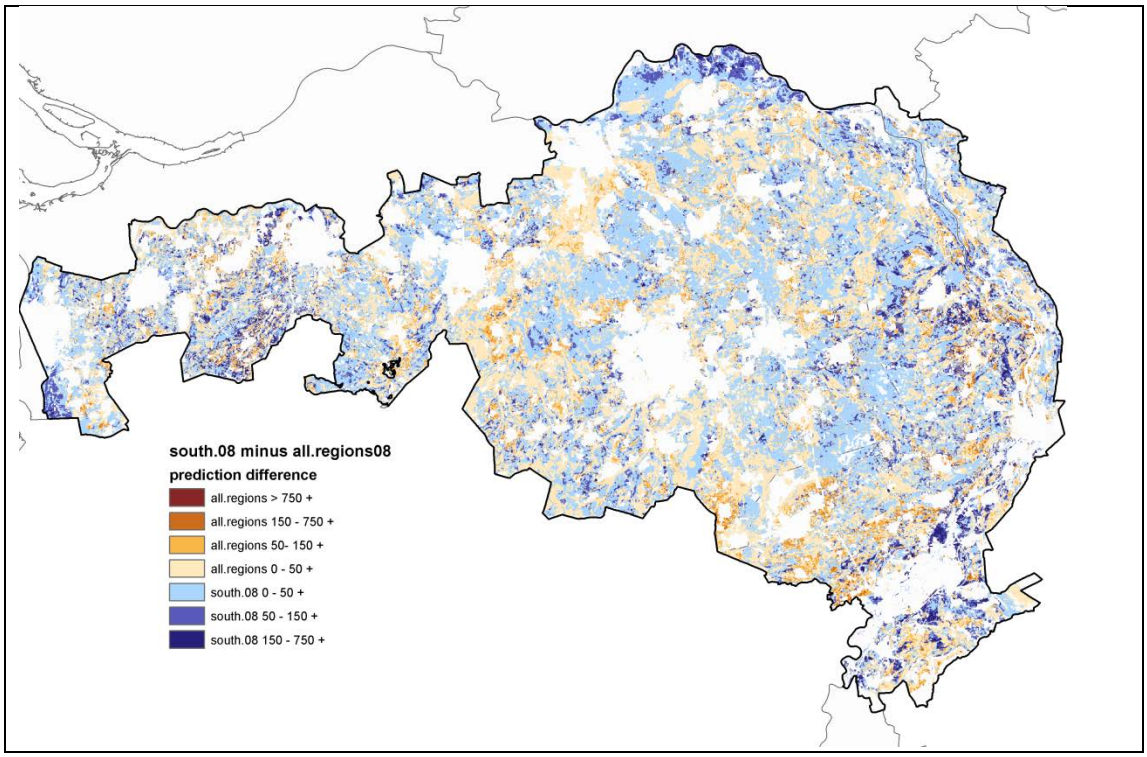


Figure 24. Difference between regional and nationwide model predictions for region south in 2008.

5 Discussion

Regression Kriging was used in this research as a method for the interpolation of measured values at point locations, in order to predict at unsampled locations. In Chapter 4 the results from this interpolation were presented. In this chapter these results are discussed.

5.1 Model results

By means of multiple linear regression, in every year-region combination, selections of covariates were modelled to represent the mean component for each RK-prediction.
From Section 4.2.6 it appears the three models for all regions include more common covariates than those of the twelve regional models. In the nationwide models, ten covariates are present in all three year-models, whereas with the data selection limited to regional models boundaries, as low as only three variables were selected (for the centre region). Some covariates, notably `dront`, were not included in regional models, but did appear in the nationwide models. Actually, when both the regional models and all.regions combined models are considered, all offered covariates were used at least once. Not used often were the organic matter at 5cm depth map (`om05`) and the soil map `gronds`.
Most successful covariate is the `gt06` map with the groundwater tables. This covariate was present in all models (albeit forced twice).

Since all models received the same treatment for selecting covariates, the stepwise deletion method, it is fair to compare them based on the R^2-a. The best models were then the south.2008 (58.4% variation explained) and centre.2007 (55.8% explained) regional models, followed by the all.regions.2008 model (54.7% explained).

The linear model that was established for all 12 regional models and 3 all-regions combined models are only half of the regression kriging product. The residuals received from the linear prediction are very important however, since they form the mean component of the regression kriging prediction (see Figure 8). At many locations, this may be the main portion of, or even the only predicted value, as the effect of the kriging prediction is sometimes very local.

Reclassification of covariate maps was not investigated. Since these maps hold many different classes one could argue that we should focus on classes that matter mainly to agriculture and nature. The risk of doing this is that classes become so general that we lose specific relationships. It may also be time consuming and would require either pre-knowledge on the importance of a class, next to testing and assessing the significance of a class for the model selection. The practical way to deal with was to simply look at the presence of classes for each map in the sampled data and let the stepwise regression selection judge the importance of these classes. As an iteration step, in some later stage the classes can be grouped together or taken separately as input maps.

5.2 Kriging results

The residuals from the linear regression prediction were the input for the ordinary kriging predictions. Computationally, the kriging part is the most time-consuming part. Predictions could last for days when no limit was set on the amount of point pairs to consider for the kriging. To have a practical measure, the limit was set at `nmax=100`, resulting in calculation time between 47 and 143 minutes on the available hardware. Higher numbers resulted in very long, impractical processing times.

Looking at the quantitative results from the cross validation, there seems to be little difference between the models. There are no obvious indications that the kriging prediction performed worse in one model or another.

The variogram shapes and ranges do not display much variety. Between years in the same region, often the same range was found. The exponential model was chosen most of the time for the best model fit. The bandwidth of separation distance just between years for the same region, are similar to those between for instance a regional approach and a nationwide approach (see results in appendix IVb for all variogrammes).

One recognizable effect can be seen in the results for the south region model predictions, when comparing the results for 2008 or (2007) with those of 2009. In the dataset with sample values, the samples from the TMV (located in nature reserves) are almost absent in 2009, while in 2007 and 2008 they are present. See Figure 3, where the peak is absent in early 2009, and compare the green line in Figure 16 with the data from those in the appendix IVb for South 2009 (page 105). In the map results after the kriging in Figure 21, the areas within class <25 NO_3 per mg/l are almost the same for the three years, but maybe a bit less pronounced in 2009. Now when we look at the uncertainty maps in 2008 en 2009 (Figure 22), and take a closer look at the blue class of < 25 (mg/l), in the map for 2009 this class is almost not present. This effect is also visible in the uncertainty maps for the nationwide predictions all.regions for 2008 and 2009 (see Appendix Vb). The absence of measurements in nature reserves in the sample data set, is translated in higher uncertainties for the map locations within nature reserves.

An effect like this can also be seen in the centre region, where large contiguous nature reserve can be found.

There are some known limitations of the regression kriging method. (Hengl et al., 2007), state a few weaknesses that can possibly explain some of the results:
1. data quality
2. under-sampling
3. reliable estimation of covariance/correlation structure
4. extrapolation outside the sampled feature space
5. Predictors with uneven relation to the target variable
6. intermediate scale modelling

Ad 1. This is true for both the measured variable and the quality of the covariates.

Ad 2. In some categories, under- sampling may have been experienced, since no minimum number of required observations was set. Missing categories were removed, but once a single observation was available in a class of a categorical variable, this variable was used. Having only a single representation for a category is not such good practice. This may also explain why some of the model coefficients are not very significant (see appendix VI – verbose model summaries).

Ad 3. The covariates may not be used in the right way, could be improved or there could be dependence on some other variable. In the same way, the variogram that is used to solve the kriging equation may be estimated poorly. There is always an amount of user-judgement when making these.

Ad 4. This has limited application in this thesis. In some regional models extreme high values have been predicted, other than sampled values. This concerns only a limited range of cells, but should be further investigated.

Ad 5. The application of animal manure and artificial fertilizer is under a lot of attention from governing organizations and limited by national and EU-legislation (Boumans & Fraters, 2011; Willems et al., 2005). The legally allowed amount of applied manure and fertilizer is differentiated by land use categories. Grassland is considered to have higher denitrification capacities than other land use, for instance arable farming, and therefore is allowed to have a higher burden of nitrogen administered to it. Nitrate levels have not been reported to increase at the rate of this application, making the relation with applied manure uneven. The use of the stone-input maps can thus be considered as an uneven related predictor, which should be modelled differently.

Ad 6. The regional modelling approach has the disadvantage that less data is available for relationship with the covariate maps. A slightly larger area includes more data, making relations stronger (and maybe some coincidental relations weaker).

5.3 Final map results

In general, the regional model predictions were predicting slightly higher values than those with the predictions that had all regions combined. This may have to do with the differences in linear models, extrapolating to higher values.

Judging the maps by appearances, there seem to be only reasonable results. Covariate map influence is clearly visible, as the location of for instance natural terrain is well recognizable, especially in the maps for the centre region, where large contiguous nature areas are found. Manure and deposition are highly correlated with agricultural practices. These maps reflect the nitrate value found in groundwater. The groundwatertable map finally, is present in every model selection, stressing the importance of this covariate in explaining nitrate levels. The soil map and pawn map are closely related to this covariate.

The choice to model regionally or combine all regions in a nationwide model can be made based on the properties of the prediction results. The nationwide results proved to be more stable, whereas the maps providing the uncertainties display a smaller bandwidth, meaning that the predictions bear a greater certainty.

The uncertainties that are calculated by the procedure with the 95% confidence limits, are expressed in maps, that accompany each regression kriging prediction result. In this way a user will have the opportunity to see whether a prediction is ‘reliable’ or not, and how broad the uncertainty class is for a certain area or even location. It appears that these uncertainties are rather large for specific areas, and that the prediction outcomes can, and should be only used to identify change over the years per region, and to pinpoint certain areas that have problematic developments. Possibly, the practical width of the calculated confidence interval can be decreased somewhat by using Block kriging. In this variant of kriging, a block size is set for the predictions, reducing the variances. This was not pursued in this thesis.

Concerning the kriging variances that were generated with the kriging predictions, there are some artefacts that were unexpected. A certain explanation cannot be given since this requires some added investigation. However, the number of points that are considered for the kriging equation at each target cell, is limited to 100, for practical reasons. In the dataset however, clusters of farms can be identified, totalling to almost these 100 points. This will make the prediction around likewise clusters a very local procedure. The mentioned artefacts are likely to be caused by this extreme clustering, and increasing the

number of points to use in the kriging process should generate variance maps without the artefacts.

The maps should not be used to assess whether at a certain location a set standard is exceeded or not, without consulting the uncertainty maps. Assuming the predicted values as certain can be deceitful, given the width of the calculated 95%-confidence interval. More specifically, there are areas in the map however, that can be identified as being quite certain to have the a predicted value below the standard, since the associated interval width at these locations is of a very small magnitude. This mostly concerns the values far below the standard.

5.4 Overall

General points

Data from two monitoring networks were used. The main reason for this was to provide data coverage for the non-agricultural soils. Even though the methods for obtaining a field sample are equal, there is a notable difference (see Figure 16). Measurement values obtained from the TMV-monitoring network generally have much smaller values, but also the threshold value seems to be lower. Another difference is the peak in seasonal timing of the two networks (see Figure 3). Then again, the data share of the nature monitoring network TMV is relatively small compared to that of the agricultural monitoring network LMM. In 2009 the data share from TMV is almost absent, when a break year in sampling for TMV was introduced. Looking at the prediction-vs-observation diagrams for 2009, the 'spread' of the data appears more clustered in around one group, compared with those of 2007 and 2008, where there seem to be two groups. The effects on the model behaviour and regression kriging-results for that year do not differ substantially from 2007 and 2008.

Factors which are not taken into account:

- seasonal fluctuations of nitrate levels (spring/summer/autumn, difference in monitoring program). Solely discrimination in the year in which the sample was taken. Therefore a sample taken on January 5 in 2008 may still be effectively a late response of all factors influencing nitrate levels in the year 2007. LMM-data are gathered in summer or autumn while TMV-data are collected on purpose in wintertime.
- effects caused by weather fluctuations (temperature, precipitation surplus, regional variations). Indirectly, the results may be present in the nitrate levels.
- discrimination of farm type (arable vs dairy farming and intensity of land use).
- soil variability within sandy soil types. For instance, (Sonneveld et al., 2010) found that glacial till within sandy soils in the North of the Netherlands correlated with a much lower nitrate level in upper groundwater than soils missing this glacial till. However, the geomorphological covariate map may explain some of these differences.

Methodological weaknesses:

- when overlaying covariate data with sample site locations, not all covariate classes are present in the same density in the sample data set. For some of these classes, no predictions are made. For the classes that are present, there was no minimal observation level (1 observation was enough). This might lead to overfitting (see point 2 and 4 of the Regression Kriging limitations below).
- overlaying the observations with the covariate grids, which were in 25 meter cell resolution sometimes forces observations in a nearby class or category, where in reality this relation is not true. This leads to a (small) decline in the linear relations we seek to establish. An example would be a sample taken in a grassland field, near the road or near a forest border, that could, wrongly, be attributed to the land use class 'infrastructure' or 'forest'. It does not seem to occur very often however (see Table *3*).

Specific points

In the modelling results, the same model-map year for stone-maps was selected in two models, specifically in `east.2008` and in `all.regions.2008`. See also the description in Section 3.1. It can be argued whether this is okay or not. As a statistical correlation, it is allowed, but as a causal relation, it is up for debate.

Data from the stone-input maps for combined nitrogen application of animal manure and fertilizer are valid for the year they are made for. A selection of a certain year in the same year-model may have a debatable amount of nitrogen still in storage in the topsoil which is not yet decomposed and still available to plant roots for crop uptake. Possibly, part of this nitrogen is not used and will infiltrate with the groundwater later in that year, or the next year. Measurements in samples in the same year in the shallow groundwater below this, will not register this. It could also happen that sampling in a certain year takes place even before the application of manure or fertilizer that year. The route of artificial fertilizer may be much more rapid and be flushed out quickly, depending on circumstances. More certainty on this can be provided by further literature study. This realisation shows that it may be better to do the following two things:

1. Only allow the stone-input maps for the model (year-1) or earlier. This means, that for instance for model year 2008 only stone 2008-1 = 2007, or earlier (2006 and 2005) can be used in models.
2. Artificial fertilizer and animal manure are currently combined in one nitrogen addition map. Offering the two variables separately might yield different results, as well as the possibility for interaction between variables in the linear models. More literature research on the valid years of application is necessary.

Number of classes in the categorical covariates

In some of the covariate maps, many classes occur. Some of these classes are not that significant, as can be judged from the verbose model listings (Appendix VI). Reclassification of these classes, or perhaps selection of only a few of the most significant classes as a single covariate map may change the model composition of the linear models. Regarding the groundwater table covariate `gt06`, for instance, three main classes could be used, instead of the 11 classes that were available in the models in this research. The land use maps lgn6 and bbg06 could also be simplified to fewer classes.

6 Conclusions and recommendations

The research objective of this study was “to predict nitrate levels at unsampled locations in upper groundwater in sandy soils in the Netherlands using Regression Kriging, and to assess the accuracy of these predictions”, as written in Section 1.3. This objective was successfully achieved. Below, each of the research questions are answered.

1. Which covariates are both relevant and available for this study?

This question has been answered by the model selection. All covariates have been selected at least once in one of the models. The most successful covariate is gt06, the map with groundwater table classes. If a ranking needs to be made:

1	gt06
2	bbg06, lgn6, kwel2
3	om60, geom
4	om10
5	om40, om100, stone6, ahn, nhx, slont, draf
6	om05, om80, pawn, stone5, stone8, laf, slaf, vds
7	om25, om120, stone7
8	lont, gronds, om05

Important covariates are land use (bbg06, lgn6), geomorphology (geom), infiltration and seepage (kwel2) and the organic matter maps in various depth layers (omxx). After this, the maps which explain where and how much nitrogen is added come in view (nhx, stone). Hydraulic property-maps are not always the first explaining covariate, but these are occasionally selected (slont, draf, laf etc.).

2. How can the covariates, determined in (1), be used in a regression model?

The answer of this research question is closely related to the first one. After a stepwise multilinear approach, a selection of covariates forms a model. This model determines how well the covariate and the combination with other selected covariates are explaining the target value. Each covariate is assigned a regression coefficients. Each model is characterized by R^2-values. This is a measure for the total variation explained by a model. The highest value found is in the regional model south.2008 with 0.594, or 59.4% of the variation explained.

3. How can the regression model be combined with kriging (point support) in the case of nitrate levels in upper groundwater and how accurate are the regression kriging results?

The results are given in Chapter 4. The regression model is used to predict residual values for the sample data set using the covariates. These residuals are then modelled by means of a variogram. With this variogram a kriging prediction is made for the target locations.
The regression model is also used to predict a regression prediction for the same target locations. The combined result of both prediction surfaces makes up the final RK-prediction. Accompanying uncertainty width maps are calculated with a 95% confidence level. These allow for regional assessment and indication of areas with large problems, not for assessing exactly whether (predicted) values exceed a legal limit.

4. What are the differences between three years when the same methodology is applied, and can these differences be explained?

Differences between years can be explained by the variation in distribution of the sample points, both geographically and value of measurements. The selection of linear model components depends strongly on the available data set. The kriging part of the regression kriging is of limited influence. It merely refines the products that are produced by the regression prediction.
When looking at the prediction and uncertainty results from the south region for 2008 and 2009, a difference in the predicted values for natural areas, as described in the discussion of Section 5.3. This can be a attributed by the difference in the sample data set for those years.

5. Will the model, when constructed at two extents (national and regional), differ in structure and accuracy, and can these differences be explained?

Yes. The covariates that are selected are different for regional and nationwide models.
Compared to the regional models, the three nationwide models are based more often on the stone-maps and nhx parameter. In the regional models, nhx occurs almost only in the three south models. The layer maps with organic matter are not so popular in the nationwide models as they are in the regional models, only three out of eight available depth layers were included. The digital elevation model ahn is also present in all three nationwide models, were this is true for the regional models only in 4 of the 12 cases.

The region that was compared in Section 4.7, had many common covariates in the linear regression model. They also differed, since the ahn (elevation) and pawn (soil/hydrology) covariate were only present in the nationwide model.

The explanation apparently lies in the greater amount of data that is available. The measurements from other regions can be used to model relationships which normally would be weak or not so pronounced, because the data collection was not covering the region, or just partly. The difference in Elevation ahn for instance, might not be so large within one region, but when all regions are combined, it suddenly may be significant in explaining nitrate levels. This can however work in two directions: important relations in one region may not be that important in another and vice versa. Therefore, some local phenomena might be ‘suppressed’ by another, more often seen relation.

Finally

When looking at the range of predicted values, the nationwide models seem to produce more stable results. Also, the uncertainty outcomes that pair with the predictions, seem to be of a smaller bandwidth at nationwide approach.

Results from this research study can be used to evaluate policy measures at national scale. Regions having more problems with high nitrate concentration in groundwater levels can be identified, by making use of the information from covariate resources. The predictions are surrounded by relatively large confidentiality estimates, making it not so suitable for the evaluation of local results for legal standards regarding nitrate concentrations.

Recommendations

Split the data set:

- Effects of mixing data from two measuring networks LMM & TMV can be seen:
  In 2009 fewer samples were taken in TMV (winter season) and in the plots of that year, effects can be seen in the diagrams of prediction-observed when comparing with the plots of previous years for the same region. Maybe splitting the dataset and repeating the method for both sets might improve results.

Use 'better' data:

- Use precipitation- and weather corrected nitrate measurements in the dataset
- An update of the boundary files for the regions exists, matching the soil type in the region on a finer scale. This should provide better matches with soil-related maps.
- Discriminate between animal manure N-load and fertilizer N-load
- Newer versions for a few covariates have been introduced: the elevation model AHN-2 is available, with increased resolution, *Bofek-2012* now replaces the predecessor pawn, which is regarded as outdated, and soil map-updates are soon to be released. Newer maps with more recent data might improve regression kriging results. Since the start of this research project for instance BBG 2008 and LGN7 were published, possibly more suitable than the maps that were used so far. The aforementioned maps reflect more or less the same timeframe as the data from the groundwater samples. The procedures should be repeated with the newly available data to check whether there is improvement.
- Investigate whether categorical maps can be reclassified into fewer classes, enabling model selection on more significant map classes

Modify data procedures:

- Try different box-cox transformations (e.g. in geoR). This concerns data transformation techniques, enabling the linear modelling perhaps a better fit of the data.
- Enable interactions between covariates in the modelling phase, this might also lead to stronger correlation in the linear regression.
- Correct or exclude the measurements with LOD-values (lower than limit of detection), and extreme high values. This will smooth the sampling data around the median values. Though this will increase the R^2, a critical look at the prediction results is needed.
- Select fewer or only most significant covariates instead of accepting the stepwise results.
- Use block kriging to decrease variances for the prediction results. This can decrease the range of uncertainty calculated with the 95% confidence-interval.
- Speed up calculations by improving the coding, for instance by using parallel computing or more efficient grid-handling techniques (eg. R-Package 'Raster')

Literature

Alterra/Stiboka (Cartographer). (2006). Bodemkaart van Nederland 1:50000.

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*: Springer.

Bonten, L. T. C., Mol-Dijkstra, J., Wieggers, H. J. J., de Vries, W., van Pul, W. A. J., & van der Hoek, K. W. (2009). Linking nitrogen deposition to nitrate concentration in groundwater below nature areas - Modelling approach and data requirements. Wageningen: Alterra.

Boumans, L. J. M., & Fraters, D. F. (2011). Nitraatconcentraties in het bovenste grondwater van de zandregio en de invloed van het mestbeleid. Visualisatie afname in de periode 1992 tot 2009. Bilthoven: RIVM.

Boumans, L. J. M., Fraters, D. F., & van Drecht, G. (2004). Nitrate leaching by atmospheric N deposition to upper groundwater in the sandy regions of The Netherlands in 1990. *Environmental Monitoring and Assessment, 93*(1-3), 1-15.

Boumans, L. J. M., Fraters, D. F., & van Drecht, G. (2008). Mapping nitrate leaching to upper groundwater in the sandy regions of The Netherlands, using conceptual knowledge. *Environmental Monitoring and Assessment, 137*(1-3), 243-249.

Buis, E., van den Ham, A., Boumans, L. J. M., Daatselaar, C. H. G., & Doornewaard, G. J. (2012). Landbouwpraktijk en waterkwaliteit op landbouwbedrijven aangemeld voor derogatie. Resultaten meetjaar 2010 in het derogatiemeetnet (pp. 107). Bilthoven: RIVM.

Burrough, P. A., & McDonnell, R. A. (1998). *Principles of Geographical Information Systems* (2nd, 2000 print ed.). New York: Oxford University Press.

CBS (Cartographer). (2008). Productbeschrijving BBG (pdf) versie 2008.1.

CCRX. (1995). Metingen in het milieu in Nederland 1993 (Monitoring the environment in the Netherlands, 1993). Bilthoven: RIVM, National Institute for Public Health and Environmental Protection.

de Goffau, A., van Leeuwen, T. C., van den Ham, A., Doornewaard, G. J., & Fraters, D. F. (2012). Minerals Policy Monitoring Programme report 2007-2010. Methods and Procedures. (pp. 97). Bilthoven: RIVM.

de Goffau, A., Wattel-Koekkoek, E. J. W., van der Hoek, K. W., & Boumans, L. J. M. (2009). Evaluatie TrendMeetnet Verzuring (pp. 78). Bilthoven: RIVM.

de Vries, F. (1999). Karakterisering van de Nederlandse gronden naar fysisch-chemische kenmerken. Wageningen: DLO Staring Centrum.

Fraters, B. F., Boumans, L. J. M., van Drecht, G., de Haan, T., & de Hoop, W. (1998). Nitrogen monitoring in groundwater in the sandy regions of the Netherlands. *Environmental Pollution, 102*(S1), 479-485.

Fraters, B. F., Kovar, K., Willems, W. J., Stockmarr, J., & Grant, R. (2005). Monitoring effectiveness of the EU Nitrates Directive Action Programmes (pp. 290). Bilthoven: RIVM.

Goovaerts, P. (1999). Geostatistics in soil science: state of the art and perspectives. *Geoderma, 89*(1-2), 45.

Gotway, C. A., & Hartford, A. H. (1996). Geostatistical Methods for Incorporating Auxiliary Information in the Prediction of Spatial Variables. *Journal of Agricultural, Biological, and Environmental Statistics, 1*(1), 24.

Hazeu, G. W., Schuiling, C., Dorland, G. J., Oldengarm, J., & Gijsbertse, H. A. (2010). Landelijk Grondgebruiksbestand Nederland versie 6 (LGN6); Vervaardiging, nauwkeurigheid en gebruik. (pp. 132). Wageningen: Alterra.

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*.

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers and Geosciences, 33*(10), 1301-1315. doi: 10.1016/j.cageo.2007.05.001

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma, 120*(1-2), 75-93. doi: 10.1016/j.geoderma.2003.08.018

Isaaks, E. H., & Srivastava, R. M. (1990). *An Introduction to Applied Geostatistics*. New York: Oxford University Press, Inc.

Kempen, B. (2011). *Updating soil information with digital soil mapping.* (PhD), Wageningen University, Wageningen.

Knotters, M., Heuvelink, G. B. M., Hoogland, T., & Walvoort, D. J. J. (2010). A disposition of interpolation techniques *WOt* (Vol. 190, pp. 90). Wageningen: Statutory Research Tasks Unit for Nature and Environment.

Koomen, A. J., & Maas, G. J. (2004). Geomorfologische Kaart Nederland (GKN); Achtergronddocument bij het landsdekkende digitale bestand (pp. 38). Wageningen: Alterra.

Kruseman, G., Luesink, H., Blokland, P. W., Hoogeveen, M. W., & De Koeijer, T. J. (2011). Mambo 2.x Design principles, model structure and data use *WOT Werkdocumenten 307*. Den Haag: LEI, Wageningen UR.

Lamsal, S., Bliss, C. M., & Graetz, D. A. (2009). Geospatial Mapping of Soil Nitrate-Nitrogen Distribution Under a Mixed-Land Use System. *Pedosphere, 19*(4), 434-445. doi: Doi: 10.1016/s1002-0160(09)60136-3

Masselink, N. J., & de Goffau, A. (2010). TrendMeetnet Verzuring - Monsternemingen in 2007/2008 (pp. 143). Bilthoven: RIVM.

Mitas, L., & Mitasova, H. (1999). Spatial Interpolation. In P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind (Eds.), *Geographical Information Systems: Principles, Techniques, Management and Applications* (2nd edition ed., pp. 404): Wiley and Sons.

Pebesma, E. J. (1996). *Mapping Groundwater Quality in the Netherlands.* (PhD), Universiteit Utrecht, Utrecht. (ISBN 90-6266-127-0)

Pebesma, E. J. (2004). Multivariate geostatistics in S: the gstat package. *Computers and Geosciences, 30*, 683-691.

Reijnders, H. F. R., van Drecht, G., Prins, H. F., Bronswijk, J. J. B., & Boumans, L. J. M. (2004). De kwaliteit van ondiep en middeldiep grondwater in het jaar 2000 en verandering daarvan in de periode 1984-2000. Bilthoven: RIVM.

Rivett, M. O., Buss, S. R., Morgan, P., Smith, J. W. N., & Bemment, C. D. (2008). Nitrate attenuation in groundwater: A review of biogeochemical controlling processes. *Water Research, 42*(16), 4215-4232. doi: http://dx.doi.org/10.1016/j.watres.2008.07.020

RIVM. (2002). Minas en Milieu - Balans en verkenning (pp. 205 pp). Bilthoven: RIVM.

Sonneveld, M. P. W., Brus, D. J., & Roelsma, J. (2010). Validation of regression models for nitrate concentrations in the upper groundwater in sandy soils. *Environmental Pollution, 158*(1), 92-97. doi: 10.1016/j.envpol.2009.07.033

Stigter, T. Y., Ribeiro, L., & Dill, A. M. M. C. (2008). Building factorial regression models to explain and predict nitrate concentrations in groundwater under agricultural land. *Journal of Hydrology, 357*(1-2), 42-56. doi: DOI: 10.1016/j.jhydrol.2008.05.009

Tiktak, A. (1999). *Modeling Non-Point Source Pollutants in Soils - Applications to the Leaching and Accumulation of Pesticides and Cadmium.* (PhD), University of Amsterdam (UvA), Amsterdam. (ISBN 90-9012365-2)

van der Gaast, J. W. J., Massop, H. T. L., Vroon, H. R. J., & Staritsky, I. G. (2006). Hydrologie op basis van karteerbare kenmerken. Wageningen: Alterra.

Vissenberg, H. A. (1994). Bepaling van een aantal kenmerken voor de nitraatbepaling in grondwater met de Nitrachek (pp. 50). Bilthoven: RIVM.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists* (Second ed.). Chichester: Wiley.

WHO. (1998). Guidelines for Drinking-Water Quality: Recommendations - addendum to Volume 1 *Guidelines for Drinking-Water quality* (2nd edition ed., Vol. 1, pp. 31). Geneva: World Health Organization.

Willems, W. J., Beusen, A. H. W., Renaud, L. V., Luesink, H., Conijn, J. G., Oosterom, H. P., . . . Schoumans, O. F. (2005). Nutriëntenbelasting van bodem en water.

Verkenning van de gevolgen van het nieuwe mestbeleid. *Evaluatie Mestbeleid* (pp. 111). Bilthoven: Milieu- en Natuurplanbureau.

Wolf, J., Beusen, A. H. W., Groenendijk, P., Kroon, T., Rötter, R., & van Zeijts, H. (2003). The integrated modeling system STONE for calculating nutrient emissions from agriculture in the Netherlands. *Environmental Modelling & Software, 18*, 20.

Woodard, D. B., Wolpert, R. L., & O'Connell, M. A. (2010). Spatial Inference of Nitrate Concentrations in Groundwater. *Journal of Agricultural, Biological, and Environmental Statistics, 15*(2), 209-227. doi: 10.1007/s13253-009-0006-x

Wösten, J. H. M., de Vries, F., Denneboom, J., & van Holst, A. F. (1998). Generalisatie en bodemfysische vertaling van de Bodemkaart van Nederland, 1:250 000 ten behoeve van de PAWN-studie. Wageningen: Stiboka.

Appendices

- I R-Code
- II Utilized software and version listing
- III Modelbuilder diagram
- IVa Model summary
- IVb Variogrammes and data depiction
- Va Data description of covariates
- Vb Data description of covariates - images
- VI Verbose model summary

Data disclaimer

In this research project, a data file containing point measurements is used. This data file is the result of collected field measurements undertaken by RIVM, by request of the Dutch ministries of Economic Affairs (EZ) and Infrastructure and Environment (I&M). It contains nitrate measurements at point level and some other attributes, next to GPS measured x-y coordinates of the location. Measurements were collected between 2006 and 2010. The samples were taken with consent at private farm locations and results are therefore strictly confidential.

Appendix I – R code scripts

- Base script for regional models (adapted to south.2008)

```
# GIMA thesis script Cor de Jong
# RK script south 2008

rm(list = objects())

# assign working directory
# located at network drive at RIVM!
setwd(dir = "N:/GIMA/Afstuderen/VelddataLMM")

# load libraries
library(sp)
library(maptools)
library(gstat)
library(rgdal)
library(spatstat)
library(foreign)
library(lattice)
library(shapefiles)

# read data (field file having nitrachek measurements 2006-2010)
no3xy <- read.table(file = "xyavgno3_3.txt", header = T, sep = "\t")

# Make subset with essential data
basicno3 <- subset (no3xy, select = c(x,y,jaar,avgno3))
basicno3 <- subset (basicno3, jaar ==2007 | jaar==2008 | jaar==2009)

# Check and find duplicate locations
dupch <- data.frame(X=basicno3$x, Y=basicno3$y)
dupch2 <- duplicated(dupch)
basicno3$duploc <- dupch2
rm(dupch, dupch2)
basicno3 <- subset(basicno3, duploc==FALSE)
basicno3 <- subset (basicno3, select = c(x,y,jaar,avgno3))
# NB: this operation removes (18960-18949=) 11 duplicate locations

# transform avgno3 and add log10no3 as a field
basicno3$log10no3 <- log10(basicno3$avgno3)

# convert data frame to SpatialPointsDataFrame and assign 'dutch' projection
coordinates(basicno3) <- ~ x+y
proj4string(basicno3) <- CRS("+init=epsg:28992")
rm(no3xy)
gc() # clean up memory

# Read auxiliary data and make an archive ON (“Object Nest”) with it
# read first continuous regional map data for south
# om = organic matter, 8 maps at different depths
om05 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os5")
ON <- om05
ON$om05 = ON$band1
ON$band1 = NULL
rm(om05)

ON$om10 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os10")$band1
ON$om25 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os25")$band1
ON$om40 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os40")$band1
ON$om60 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os60")$band1
ON$om80 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os80")$band1
ON$om100 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os100")$band1
ON$om120 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_os120")$band1
ON$stone5 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_stone05n")$band1
ON$stone6 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_stone06n")$band1
ON$stone7 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_stone07n")$band1
ON$stone8 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_stone08n")$band1
ON$nhx <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_nh3_10")$band1
ON$ahn <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_ahn")$band1
ON$kwel2 <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_kwel2")$band1
gc()
```

```
gc()
# remark: in the maps with organic matter, om = -9 occurs!
# replace this with om = NA (concerns built up area)
ON$om05[ON$om05<0] <- NA
ON$om10[ON$om10<0] <- NA
ON$om25[ON$om25<0] <- NA
ON$om40[ON$om40<0] <- NA
ON$om60[ON$om60<0] <- NA
ON$om80[ON$om80<0] <- NA
ON$om100[ON$om100<0] <- NA
ON$om120[ON$om120<0] <- NA
gc()
gc()

# groundwater tables; as factor
ON$gt06 <- as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_gt06")$band1)
ON$bbg06 <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_bbg06")$band1)
ON$gronds <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_gronds")$band1)
ON$pawn <-as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_pawn")$band1)
ON$lgn6 <- as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_lgn6")$band1)
ON$geom <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_aggeom")$band1)
ON$draf <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_drwst_afw")$band1)
ON$dront<-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_drwst_ontw")$band1)
ON$laf <- as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_l_afw")$band1)
ON$lont <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_l_ontw")$band1)
ON$slaf <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_slafst_afw")$band1)
ON$slont <-
+ as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_slafst_ontw")$band1)
ON$vds <- as.factor(readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/Afstuderen/Data/south/s_vds")$band1)
gc()
gc()
gc()
proj4string(ON) <- CRS("+init=epsg:28992") # assign dutch projection

# The GRID "ON" now contains all auxiliary data, these values must be added to
# the point selection of basicno3
# overlay with measurement data to pointfile (from package "sp": command 'over')
ovlall <- over(basicno3, ON)
basicno3$om05 <- ovlall$om05
basicno3$om10 <- ovlall$om10
basicno3$om25 <- ovlall$om25
basicno3$om40 <- ovlall$om40
basicno3$om60 <- ovlall$om60
basicno3$om80 <- ovlall$om80
basicno3$om100 <- ovlall$om100
basicno3$om120 <- ovlall$om120
basicno3$stone5 <- ovlall$stone5
basicno3$stone6 <- ovlall$stone6
basicno3$stone7 <- ovlall$stone7
basicno3$stone8 <- ovlall$stone8
basicno3$nhx <- ovlall$nhx
basicno3$gt06 <-ovlall$gt06
basicno3$ahn <- ovlall$ahn
basicno3$bbg06 <- ovlall$bbg06
basicno3$gronds <- ovlall$gronds
basicno3$kwel2 <- ovlall$kwel2
basicno3$pawn <-ovlall$pawn
basicno3$lgn6 <-ovlall$lgn6
basicno3$geom <-ovlall$geom
basicno3$draf <-ovlall$draf
basicno3$dront <-ovlall$dront
basicno3$laf <-ovlall$laf
basicno3$lont <-ovlall$lont
basicno3$slaf <-ovlall$slaf
basicno3$slont <-ovlall$slont
basicno3$vds <-ovlall$vds
rm(ovlall)

# data clean-up, to assure every point in the measuring set has data in all covariate
```

```
# layers, also prevent problems with unequal nr of residuals later:
# select only those pts within the region "south"; these have a value, others don't
# check with summary(basicno3south) where NA's occur.
# clean missing data: some pts do not have a value for map ahn,kwel2 and om10
# use only those points that have coverage in the organic matter map
basicno3south <- subset(basicno3, basicno3$om10 > 0)
basicno3south <- subset(basicno3south, basicno3$gt06 > 0)
# Next, dispose of points with 'NA' data in ahn-map
basicno3south <- subset(basicno3south, (basicno3south$ahn >= min(basicno3south$ahn, na.rm=T)))
# dispose of points with 'NA' data in kwel2-map
basicno3south <- subset(basicno3south, (basicno3south$kwel2 >= min(basicno3south$ahn, na.rm=T)))
# dispose of points with 'NA' data in stone-map
basicno3south <- subset(basicno3south, basicno3south$stone5 != "")
# note: 125 points, 2% of total (mostly 2008 (#64))
# examine these missing points in: stonmiss <- subset(basicno3south, is.na
# (basicno3south$stone5))
# dispose of points with 'NA' data in draf-map
basicno3south <- subset(basicno3south, basicno3south$draf != "")
# dispose of points with 'NA' data in dront-map
basicno3south <- subset(basicno3south, basicno3south$dront != "")
# dispose of points with 'NA' in geom-map
basicno3south <- subset(basicno3south, basicno3south$geom != "")
#clean up memory a bit
rm(basicno3)
gc()
gc()
gc()
gc()

# work with 2008 data
south.2008 <- subset(basicno3south,jaar==2008)
rm(basicno3south)
summary(south.2008)

# any missing data or empty levels? --> use drop.levels
table(south.2008$gt06)
south.2008$gt06 <- droplevels(south.2008$gt06)
table(south.2008$bbg06)
south.2008$bbg06 <- droplevels(south.2008$bbg06)
table(south.2008$pawn)
south.2008$pawn <- droplevels(south.2008$pawn)
table(south.2008$gronds)
south.2008$gronds <- droplevels(south.2008$gronds)
table(south.2008$lgn6)
south.2008$lgn6 <- droplevels(south.2008$lgn6)
table(south.2008$geom)
south.2008$geom <- droplevels(south.2008$geom)
table(south.2008$draf)
south.2008$draf <- droplevels(south.2008$draf)
table(south.2008$dront)
south.2008$dront <- droplevels(south.2008$dront)
table(south.2008$laf)
south.2008$laf <- droplevels(south.2008$laf)
table(south.2008$lont)
south.2008$lont <- droplevels(south.2008$lont)
table(south.2008$slaf)
south.2008$slaf <- droplevels(south.2008$slaf)
table(south.2008$slont)
south.2008$slont <- droplevels(south.2008$slont)
table(south.2008$vds)
south.2008$vds <- droplevels(south.2008$vds)

## commence stepwise regression

south.2008.mod <- lm(log10no3 ~ . -x -y -jaar -avgno3, data=south.2008)
summary(south.2008.mod) # basic model; includes all possible variables
# now optimize with stepwise deletion (automatic determination of most significant parameters
# yields best lm; k-factor=penalty factor)
south.2008.mod2 <- step(south.2008.mod, k=4) # optimized model
summary(south.2008.mod2)

## plot some diagnostics
# draw QQ-plot
qqnorm(residuals(south.2008.mod2), main="residuals south.2008.mod2")
qqline(residuals(south.2008.mod2))
# draw plot of residuals with predictions
```

```
south.2008$pred <- predict(south.2008.mod2)
plot(south.2008$log10no3,south.2008$pred)
plot(south.2008$pred,residuals(south.2008.mod2))

# check influence of x-y coordinates given last two plots: results in worse fit
# maybe spatial dependency given the decrease in residuals with increasing no3 values

if(T){
+  south.2008.mod2xy <- lm(formula = log10no3 ~ x + y + om05 + om10 + om40 + om60 + gt06 + stone5+
+  stone6 + stone7 + nhx  + bbg06 + kwel2 + lgn6 + geom + slaf + draf, data = south.2008))
  summary(south.2008.mod2xy)
  south.2008$pred.xy <- predict(south.2008.mod2xy)
  plot(south.2008$log10no3,south.2008$pred.xy,xlim=c(0,3),ylim=c(0,3))
  plot(south.2008$pred.xy,residuals(south.2008.mod2xy))
}

# conclusion: a small influence of y-coordinates exists : avgno3 decreases with y-
# coordinate increase; this means that the more south the higher nitrate will be

###################################
# data clean up, to save memory, first save large file and reload
save(ON, file="ON")
rm(ON)
gc()
gc()
load("ON")
names(ON)
summary(ON)
# clean up unused data to liberate memory (beware: specify anew for each model):
ON$om25 <- NULL
ON$om80 <- NULL
ON$om100 <- NULL
ON$om120 <- NULL
ON$stone8 <- NULL
ON$gronds <- NULL
ON$pawn <- NULL
ON$dront <- NULL
ON$lont <- NULL
ON$laf <- NULL
ON$slont <- NULL
ON$vds <- NULL

gc()
gc()
gc()
gc()

# construct dataframe df.ON and move all NA data to df.ON.NA, on which no predictions take place
# x,y from bbox(ON)
df.ON <- expand.grid(x=seq(73876, by=25,length = 5582),y=seq(337773, by=25,length = 3566))

# merge covariate data from ON to df.ON
df.ON <- cbind(df.ON, ON@data)

# check "ON" for categorical variables with missing observations in south.2008
table(south.2008$gt06)
#  0   2   4   5   6   7   8   9  10  11
# 12  48 265 123  56 411 290 670 590  58
table(ON$gt06)
#       0       1       2       3       4       5       6       7       8       9      10      11
# 1005884   17456  170603    8343  994555  297245  221240 1280750  554267 2135791 1717174  694479
# in ON classes exist for which there are no observations in the sample data (levels 1 & 3)

Likewise for the other categorical variables:
table(south.2008$bbg06)
# 11   51   60   61   62
# 12 2272  190   38   11

table(south.2008$lgn6)
#   1   2   3   4   5   6  10  11  12  25  26  35  36  37  38  39  40  45  61
# 747 534 126  79 184 545  41 105  91   1   1   1  15   6   7  19   2  12   7

table(south.2008$geom)
# 6   7   8  10  12  13  14  15  16  22
# 6  81 225  58 310 898 583  31 324   7
table(south.2008$draf)
```

```
# 3    4    5    6    7
#42  398 1321  732   30
table(south.2008$slaf)
# 2   3   4   5   6   7   8   9
# 1   8  14  65 297 950 780 408

# now remove all NA's for gt06 from df.ON
df.ON.NA <- df.ON[is.na(df.ON$gt06),]
df.ON <- df.ON[!is.na(df.ON$gt06),]
gc()
gc()

# factor variables in south.2008.mod2 are: gt06, bbg06, lgn6, geom, draf, slaf
# levels in gt06 for 2008 are:
table(south.2008$gt06)
#  0   2   4   5   6   7   8   9  10  11
# 12  48 265 123  56 411 290 670 590  58

# commence removal until all levels are processed
gt06.levels <- levels(south.2008$gt06)
ON.levels <- levels(df.ON$gt06)
dim(df.ON)
for (ii in 1:length(ON.levels)){
  if (sum(ON.levels[ii]==gt06.levels)==0)
    df.ON$gt06[df.ON$gt06==ON.levels[ii]] <- NA
}
if(sum(is.na(df.ON$gt06)>0)) {
  df.ON.NA <- rbind(df.ON.NA,df.ON[is.na(df.ON$gt06),])
  df.ON <- df.ON[!is.na(df.ON$gt06),]
}
gc()
gc()

#repeat for other factor variables:
lgn6.levels <- levels(south.2008$lgn6)
ON.levels <- levels(df.ON$lgn6)
dim(df.ON)
for (ii in 1:length(ON.levels)){
 if (sum(ON.levels[ii]==lgn6.levels)==0)
   df.ON$lgn6[df.ON$lgn6==ON.levels[ii]] <- NA
}
if(sum(is.na(df.ON$lgn6)>0)) {
 print(sum(is.na(df.ON$lgn6)>0))
 df.ON.NA <- rbind(df.ON.NA,df.ON[is.na(df.ON$lgn6),])
 df.ON <- df.ON[!is.na(df.ON$lgn6),]
}
gc()
gc()

bbg06.levels <- levels(south.2008$bbg06)
ON.levels <- levels(df.ON$bbg06)
dim(df.ON)
for (ii in 1:length(ON.levels)){
  if (sum(ON.levels[ii]==bbg06.levels)==0)
    df.ON$bbg06[df.ON$bbg06==ON.levels[ii]] <- NA
}
if(sum(is.na(df.ON$bbg06)>0)) {
  print(sum(is.na(df.ON$bbg06)>0))
  df.ON.NA <- rbind(df.ON.NA,df.ON[is.na(df.ON$bbg06),])
  df.ON <- df.ON[!is.na(df.ON$bbg06),]
}
gc()
gc()

geom.levels <- levels(south.2008$geom)
ON.levels <- levels(df.ON$geom)
dim(df.ON)
for (ii in 1:length(ON.levels)){
  if (sum(ON.levels[ii]==geom.levels)==0)
    df.ON$geom[df.ON$geom==ON.levels[ii]] <- NA
}
if(sum(is.na(df.ON$geom)>0)) {
  print(sum(is.na(df.ON$geom)>0))
  df.ON.NA <- rbind(df.ON.NA,df.ON[is.na(df.ON$geom),])
  df.ON <- df.ON[!is.na(df.ON$geom),]
}
```

```
gc()
gc()

draf.levels <- levels(south.2008$draf)
ON.levels <- levels(df.ON$draf)
dim(df.ON)
for (ii in 1:length(ON.levels)){
  if (sum(ON.levels[ii]==draf.levels)==0)
    df.ON$draf[df.ON$draf==ON.levels[ii]] <- NA
}
if(sum(is.na(df.ON$draf)>0)) {
  print(sum(is.na(df.ON$draf)>0))
  df.ON.NA <- rbind(df.ON.NA,df.ON[is.na(df.ON$draf),])
  df.ON <- df.ON[!is.na(df.ON$draf),]
}

gc()
gc()
gc()
gc()

slaf.levels <- levels(south.2008$slaf)
ON.levels <- levels(df.ON$slaf)
dim(df.ON)
for (ii in 1:length(ON.levels)){
  if (sum(ON.levels[ii]==slaf.levels)==0)
    df.ON$slaf[df.ON$slaf==ON.levels[ii]] <- NA
}
if(sum(is.na(df.ON$slaf)>0)) {
  print(sum(is.na(df.ON$slaf)>0))
  df.ON.NA <- rbind(df.ON.NA,df.ON[is.na(df.ON$slaf),])
  df.ON <- df.ON[!is.na(df.ON$slaf),]
}
gc()
gc()
gc()
gc()

# add empty field 'pred' to df.ON.NA
df.ON.NA$pred <- NA

# predict values for all valid locations in df.ON
df.ON$pred <- predict(south.2008.mod2,df.ON)
gc()
gc()
gc()
gc()

# combine the two data frames
df.ON.pred <- rbind(df.ON,df.ON.NA)
gc()
gc()

rm(df.ON, df.ON.NA) # can both be removed

# re-order x and y in combined data frame
df.ON.pred <- df.ON.pred[order(df.ON.pred$y, df.ON.pred$x),]
gc()
gc()

# add (df.ON.pred["pred"]) to ON; this will be the prediction grid mask
ON$pred <- df.ON.pred$pred
names(ON)
gc()
gc()

# temporally save dataframes and remove from memory to enable the kriging calculation
save(ON, file="ON")
save(df.ON.pred, file="df.ON.pred")

gc()
gc()
gc()
gc()
gc()
```

```
# export to Asciigrid for display in ArcGIS
ON$btpred <- 10^(ON$pred) # !transform, only valid for quick scan, grid is also used as mask for
kriging!
writeGDAL(ON["btpred"],       "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_mask08.gtif",
drivername = "GTiff", type= "Float32", mvFlag=-9999)
# free memory
rm(ON, df.ON.pred)
gc()
gc()
gc()
gc()

### Kriging

# add column with residuals to south.2008
south.2008$residuals <- residuals(south.2008.mod2)

# calculate variogram of residuals 2008
g.res <- gstat(formula = residuals ~1, data = south.2008)
vg.res <- variogram(g.res, boundaries = c(25,50,75,100,150,250,350,500,750,1000,1500,4000))
# check range; zoom in when needed
plot(vg.res, plot.nu = T)
vgm.res <- vgm(nugget=0.10, psill=0.20, range=2000, model="Exp")

# adapt to fit with correct variogram parameter
vgm.res <- fit.variogram(vg.res,vgm.res)
plot(vg.res, vgm.res, plot.nu = T, main = "Residuals")

#load mask for prediction locations
predgrid <- readGDAL("R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_mask08.gtif")
proj4string(predgrid) <- CRS("+init=epsg:28992")

# point kriging of residuals
system.time(df.ON.kr <- krige(residuals ~1, south.2008, newdata = predgrid, vgm.res, nmax=100,
+ debug.level=-1))
names(df.ON.kr)[1] = "res.pred"
names(df.ON.kr)[2] = "res.var"
names(df.ON.kr)

# kriging above takes about 77 minutes of calculation time (jaar=2008, nmax=100)
gc()
gc()
gc()

# now add regression prediction df.ON.pred to residual kriging prediction (df.ON.kr)
load("df.ON.pred")

df.ON.kr$pred <- df.ON.pred$pred + df.ON.kr$res.pred
df.ON.kr$var <- df.ON.kr$res.var
df.ON.kr$sd <- sqrt(df.ON.kr$var)

# Backtransform predicted logno3; make new variable btfno3
df.ON.kr$btfno3 <- 10^(df.ON.kr$pred+0.5*df.ON.kr$var) # for log10
gc()
gc()
gc()

# export grids for further use in GIS

writeGDAL(df.ON.kr["btfno3"],
+ "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_btno3_08.tif", drivername = "GTiff",
+ type= "Float32", mvFlag=-9999)
writeGDAL(df.ON.kr["var"], "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_var_08.tif",
+ drivername = "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.ON.kr["sd"], "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_sd_08.tif",
+ drivername = "GTiff", type= "Float32", mvFlag=-9999)

#save final products regression kriging south 2008

save(df.ON.kr, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/df.ON.kr.s08")
save(south.2008, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/south.2008")
save(south.2008.mod2, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/south.2008.mod2")
save(vgm.res, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/south.2008.vgm.res")
save(vg.res, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/south.2008.vg.res")
save(g.res, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/south.2008.g.res")
```

```
## Uncertainties

# check if still loaded in memory, otherwise:
load("R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/df.ON.kr.s08")
# calculate 2.5% lower boundary
df.ON.kr$lb <- df.ON.kr$pred - 1.96 * df.ON.kr$sd
# calculate 97.5% upper boundary
df.ON.kr$ub <- df.ON.kr$pred + 1.96 * df.ON.kr$sd

# Backtransform LB predicted logno3;
df.ON.kr$lbtr <- 10^(df.ON.kr$lb) # for log10
df.ON.kr$ubtr <- 10^(df.ON.kr$ub) #  for log10
# Determine conf95% range:
df.ON.kr$conf95 <- df.ON.kr$ubtr - df.ON.kr$lbtr


# export grids for use in GIS
writeGDAL(df.ON.kr["lbtr"], "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_lb_08.tif",
+ drivername = "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.ON.kr["ubtr"], "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_ub_08.tif",
+ drivername = "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.ON.kr["conf95"],
+ "R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/s_wiconf_08.tif", drivername = "GTiff",
+ type= "Float32", mvFlag=-9999)

### Crossvalidation

# Leave one out crossvalidation (LOOCV); takes a while...
south.2008.cv <- krige.cv(residuals~1, south.2008, model = vgm.res, nmax=100, nfold =
+ nrow(south.2008))

summary(south.2008.cv)
var(south.2008.cv$zscore)
mean(south.2008.cv$zscore)
hist(south.2008.cv$zscore)
bubble(south.2008.cv, z="residual")

# save LooCV results
save(south.2008.cv, file="R:/MIL/Werkmappen/jongdc/GIMA/modeloutput/south/2008/south.2008.cv")
```

- Script for nationwide model (requires 4 regional runs)

```
# Nationwide model
# RK script, GIMA-thesis Cor de Jong
# directories adapted to windows environment "S:/R/jongdc/GIMA/Data/"
# Linux-Rstudio-server uses different directory paths ()
#
### up to line 560 script is generic! Calculation per region after then!
# When model is still in memory (only valid for equal years), repeat with same variogramme

## STEP 1: SAMPLE DATA

rm(list = objects())

# assign working directory
# located at network drive at RIVM
setwd("S:/R/jongdc/GIMA/Data/VelddataLMM")
# linux: setwd(dir = "/s-schijf/jongdc/GIMA/Data/VelddataLMM")

# load libraries
library(sp)
library(gstat)
library(rgdal)
library(foreign)
library(lattice)
library(raster)
library(maptools)

# read data (field file having nitrachek measurements 2006-2010)
no3xy <- read.table(file = "xyavgno3_3.txt", header = T, sep = "\t")
no3xy$avgno3 <- no3xy$avgNO3
no3xy$jaar <- no3xy$Jaar
# Make subset with essential data
basicno3 <- subset (no3xy, select = c(x, y, jaar, avgno3))
basicno3 <- subset (basicno3, jaar == 2007 | jaar == 2008 | jaar == 2009)

# Check for and find duplicate locations
dupch <- data.frame(X=basicno3$x, Y=basicno3$y)
dupch2 <- duplicated(dupch)
basicno3$duploc <- dupch2
rm(dupch, dupch2)
basicno3 <- subset(basicno3, duploc == FALSE)
basicno3 <- subset (basicno3, select = c(x, y, jaar, avgno3))
# NB: this operation has removed (18960-18949 = ) 11 duplicate locations

# add log10no3 as a field
basicno3$log10no3 <- log10(basicno3$avgno3)

# convert table to pointmap and assign 'dutch' projection
coordinates(basicno3) <- ~ x+y
proj4string(basicno3) <- CRS("+init=epsg:28992")

rm(no3xy)
gc()

##  STEP 2: OVERLAYS

# The GRIDS "NON", "EON", "CON" and "SON" now contain all auxiliary data, these values must be
# added to the point selection of basicno3
# overlay with measurement data to pointfile (from package "sp": command 'over')

# EAST

load("S:/R/jongdc/GIMA/Data/modeloutput/all4/NON")
load("S:/R/jongdc/GIMA/Data/modeloutput/all4/EON")
load("S:/R/jongdc/GIMA/Data/modeloutput/all4/CON2")
load("S:/R/jongdc/GIMA/Data/modeloutput/all4/SON")

# overlay spatialpointsdataframe "basicno3" with spatialgriddataframes "NON, EON, CON, # SON" to
# obtain pointvalues
# check for same CRS (with identicalCRS(x,y)), if FALSE then assign again with proj4string(object)
# <- CRS("+init=epsg:28992")

proj4string(NON) <- CRS("+init=epsg:28992")
north <- over(basicno3, NON)
```

```
proj4string(EON) <- CRS("+init=epsg:28992")
east <- over(basicno3, EON)
proj4string(CON) <- CRS("+init=epsg:28992")
centre <- over(basicno3, CON)
proj4string(SON) <- CRS("+init=epsg:28992")
south <- over(basicno3, SON)

# make regional files from basicno3
northpoints <- basicno3
eastpoints <- basicno3
centrepoints <- basicno3
southpoints <- basicno3

# 1 of 4 add overlay data to regional files; NORTH

northpoints$om05 <- north$om05
northpoints$om10 <- north$om10
northpoints$om25 <- north$om25
northpoints$om40 <- north$om40
northpoints$om60 <- north$om60
northpoints$om80 <- north$om80
northpoints$om100 <- north$om100
northpoints$om120 <- north$om120
northpoints$stone5 <- north$stone5
northpoints$stone6 <- north$stone6
northpoints$stone7 <- north$stone7
northpoints$stone8 <- north$stone8
northpoints$nhx <- north$nhx
northpoints$gt06 <-north$gt06
northpoints$ahn <- north$ahn
northpoints$bbg06 <- north$bbg06
northpoints$gronds <- north$gronds
northpoints$kwel2 <- north$kwel2
northpoints$pawn <-north$pawn
northpoints$lgn6 <-north$lgn6
northpoints$geom <-north$geom
northpoints$draf <-north$draf
northpoints$dront <-north$dront
northpoints$laf <-north$laf
northpoints$lont <-north$lont
northpoints$slaf <-north$slaf
northpoints$slont <-north$slont
northpoints$vds <-north$vds
rm(north)

# data clean-up, to assure every point in the measuring set has data in all covariate # layers,
# also prevent problems with unequal # of residuals later
# check with summary(northpoints) where NA's occur

northpoints <- subset(northpoints,northpoints$om10 > 0)
northpoints <- subset(northpoints,(northpoints$ahn >= min(northpoints$ahn, na.rm=T)))
northpoints <- subset(northpoints,(northpoints$kwel2 >= min(northpoints$kwel2,  na.rm=T)))
northpoints <- subset(northpoints, northpoints$stone5 != "")
northpoints <- subset(northpoints, northpoints$draf != "")
northpoints <- subset(northpoints, northpoints$dront != "")
northpoints <- subset(northpoints, northpoints$geom != "")
#clean up memory
rm(basicno3)
gc()
gc()
gc()
gc()

# 2 of 4 add overlay data to regional files; EAST

eastpoints$om05 <- east$om05
eastpoints$om10 <- east$om10
eastpoints$om25 <- east$om25
eastpoints$om40 <- east$om40
eastpoints$om60 <- east$om60
eastpoints$om80 <- east$om80
eastpoints$om100 <- east$om100
eastpoints$om120 <- east$om120
eastpoints$stone5 <- east$stone5
eastpoints$stone6 <- east$stone6
eastpoints$stone7 <- east$stone7
```

```
eastpoints$stone8 <- east$stone8
eastpoints$nhx <- east$nhx
eastpoints$gt06 <-east$gt06
eastpoints$ahn <- east$ahn
eastpoints$bbg06 <- east$bbg06
eastpoints$gronds <- east$gronds
eastpoints$kwel2 <- east$kwel2
eastpoints$pawn <-east$pawn
eastpoints$lgn6 <-east$lgn6
eastpoints$geom <-east$geom
eastpoints$draf <-east$draf
eastpoints$dront <-east$dront
eastpoints$laf <-east$laf
eastpoints$lont <-east$lont
eastpoints$slaf <-east$slaf
eastpoints$slont <-east$slont
eastpoints$vds <-east$vds
rm(east)

# data clean-up
eastpoints <- subset(eastpoints, eastpoints$om10 > 0
eastpoints <- subset(eastpoints, eastpoints$$ahn >= min(eastpoints$ahn, na.rm=T)))
eastpoints <- subset(eastpoints, eastpoints$kwel2 >= min(eastpoints$kwel2, na.rm=T)))
eastpoints <- subset(eastpoints, eastpoints$stone5 != "")
eastpoints <- subset(eastpoints, eastpoints$draf != "")
eastpoints <- subset(eastpoints, eastpoints$dront != "")
eastpoints <- subset(eastpoints, eastpoints$geom != "")

# 3 of 4 add overlay data to regional files; CENTRE

centrepoints$om05 <- centre$om05
centrepoints$om10 <- centre$om10
centrepoints$om25 <- centre$om25
centrepoints$om40 <- centre$om40
centrepoints$om60 <- centre$om60
centrepoints$om80 <- centre$om80
centrepoints$om100 <- centre$om100
centrepoints$om120 <- centre$om120
centrepoints$stone5 <- centre$stone5
centrepoints$stone6 <- centre$stone6
centrepoints$stone7 <- centre$stone7
centrepoints$stone8 <- centre$stone8
centrepoints$nhx <- centre$nhx
centrepoints$gt06 <-centre$gt06
centrepoints$ahn <- centre$ahn
centrepoints$bbg06 <- centre$bbg06
centrepoints$gronds <- centre$gronds
centrepoints$kwel2 <- centre$kwel2
centrepoints$pawn <-centre$pawn
centrepoints$lgn6 <-centre$lgn6
centrepoints$geom <-centre$geom
centrepoints$draf <-centre$draf
centrepoints$dront <-centre$dront
centrepoints$laf <-centre$laf
centrepoints$lont <-centre$lont
centrepoints$slaf <-centre$slaf
centrepoints$slont <-centre$slont
centrepoints$vds <-centre$vds
rm(centre)

# data clean-up
centrepoints <- subset(centrepoints, centrepoints$om10 > 0)
centrepoints <- subset(centrepoints, centrepoints$ahn >= min(centrepoints$ahn, na.rm=T)))
centrepoints <- subset(centrepoints, centrepoints$kwel2 >= min(centrepoints$kwel2, na.rm=T)))
centrepoints <- subset(centrepoints, centrepoints$stone5 != "")
centrepoints <- subset(centrepoints, centrepoints$draf != "")
centrepoints <- subset(centrepoints, centrepoints$dront != "")
centrepoints <- subset(centrepoints, centrepoints$geom != "")

# 4 of 4 add overlay data to regional files; SOUTH

southpoints$om05 <- south$om05
southpoints$om10 <- south$om10
southpoints$om25 <- south$om25
southpoints$om40 <- south$om40
southpoints$om60 <- south$om60
```

```
southpoints$om80 <- south$om80
southpoints$om100 <- south$om100
southpoints$om120 <- south$om120
southpoints$stone5 <- south$stone5
southpoints$stone6 <- south$stone6
southpoints$stone7 <- south$stone7
southpoints$stone8 <- south$stone8
southpoints$nhx <- south$nhx
southpoints$gt06 <-south$gt06
southpoints$ahn <- south$ahn
southpoints$bbg06 <- south$bbg06
southpoints$gronds <- south$gronds
southpoints$kwel2 <- south$kwel2
southpoints$pawn <-south$pawn
southpoints$lgn6 <-south$lgn6
southpoints$geom <-south$geom
southpoints$draf <-south$draf
southpoints$dront <-south$dront
southpoints$laf <-south$laf
southpoints$lont <-south$lont
southpoints$slaf <-south$slaf
southpoints$slont <-south$slont
southpoints$vds <-south$vds
rm(south)

# data clean-up

southpoints <- subset(southpoints, southpoints$om10 > 0)
southpoints <- subset(southpoints, southpoints$ahn >= min(southpoints$ahn, na.rm=T)))
southpoints <- subset(southpoints, southpoints$kwel2 >= min(southpoints$kwel2, na.rm=T)))
southpoints <- subset(southpoints, southpoints$stone5 != "")
southpoints <- subset(southpoints, southpoints$draf != "")
southpoints <- subset(southpoints, southpoints$dront != "")
southpoints <- subset(southpoints, southpoints$geom != "")


# STEP 3 Combine regional pointfiles to one file
###
all.regions <- rbind.SpatialPointsDataFrame(northpoints, eastpoints, centrepoints, southpoints)
###
# start modelling LM for one year: select 2007, 2008 or 2009
###

all.regions.2009 <- subset(all.regions, all.regions$jaar == 2009)

summary(all.regions.2009)

### prepare for LM

table(all.regions.2009$gt06)
all.regions.2009$gt06 <- droplevels(all.regions.2009$gt06)
levels(all.regions.2009$gt06)
table(all.regions.2009$pawn)
# zorgt later voor NA, verwijderen 1 los record
all.regions.2009 <- subset(all.regions.2009, all.regions.2009$pawn !=21)
all.regions.2009$pawn <- droplevels(all.regions.2009$pawn)
table(all.regions.2009$gronds)
all.regions.2009$gronds <- droplevels(all.regions.2009$gronds)
table(all.regions.2009$lgn6)  # missing levels?
all.regions.2009 <- subset(all.regions.2009, all.regions.2009$lgn6 !=28)
all.regions.2009 <- subset(all.regions.2009, all.regions.2009$lgn6 !=18)
all.regions.2009$lgn6 <- droplevels(all.regions.2009$lgn6)
table(all.regions.2009$lgn6)
table(all.regions.2009$geom)
all.regions.2009$geom <- droplevels(all.regions.2009$geom)
table(all.regions.2009$draf)
all.regions.2009$draf <- droplevels(all.regions.2009$draf)
table(all.regions.2009$dront)
all.regions.2009$dront <- droplevels(all.regions.2009$dront)
table(all.regions.2009$laf)
all.regions.2009$laf <- droplevels(all.regions.2009$laf)
table(all.regions.2009$lont)
all.regions.2009$lont <- droplevels(all.regions.2009$lont)
table(all.regions.2009$slaf)
all.regions.2009$slaf <- droplevels(all.regions.2009$slaf)
table(all.regions.2009$slont)
```

```
all.regions.2009$slont <- droplevels(all.regions.2009$slont)
table(all.regions.2009$vds)
all.regions.2009$vds <- droplevels(all.regions.2009$vds)
all.regions.2009 <- subset(all.regions.2009, all.regions.2009$vds !=12)
table(all.regions.2009$bbg06)
all.regions.2009$bbg06 <- droplevels(all.regions.2009$bbg06)
levels(all.regions.2009$bbg06)
summary(all.regions.2009$bbg06)

# predict for 2009

# commence stepwise regression
all.regions.2009.mod <- lm(log10no3 ~ . -x -y -jaar -avgno3, data=all.regions.2009)
summary(all.regions.2009.mod)
# stepwise (improve lm; k-factor=penalty factor)
all.regions.2009.mod2 <- step(all.regions.2009.mod, k=4)

################################################################################
## STEP 4

# Treatment in 4 blocks; predict by region and krige with nationwide model 2009
# run STEP 4 for each region, model and variogram do not change

rm(SON,CON,EON,NON)
# rm(*.levels); all unnecessary files)

## first run: south; SON
load("S:/R/jongdc/GIMA/Data/modeloutput/all4/SON")
gc()
gc()
gc()


# clean-up of unused covariates:
# used are: stone5 + stone6 + stone7 + stone8 + nhx + gt06 + ahn + kwel2 + pawn + lgn6 + geom +
# slaf + dront + vds
# then unused are: om05,om10,om25,om40, om60, om80, om100, om120, gronds, bbg06, slont, laf, lont,
draf
SON$om05 <- NULL
SON$om10 <- NULL
SON$om25 <- NULL
SON$om40 <- NULL
SON$om60 <- NULL
SON$om80 <- NULL
SON$om100 <- NULL
SON$om120 <- NULL
SON$gronds <- NULL
SON$bbg06 <- NULL
SON$slont <- NULL
SON$draf <- NULL
SON$lont <- NULL
SON$laf <- NULL
gc()
gc()
gc()

# make regional dataframe df.SON and split data into df.SON.NA (no predictions) and df.SON
# (predictions)
df.SON <- expand.grid(x=seq(73875.89,by=25,length=5582), y=seq(337772.81,by=25,length=3566))
# x,y taken from extent(SON)
# combine with covariate data from SON
df.SON <- cbind(df.SON, SON@data)

# remove all NA's in gt06
df.SON.NA <- df.SON[is.na(df.SON$gt06),]
df.SON <- df.SON[!is.na(df.SON$gt06),]
gc()
gc()

# factor variables are gt06, lgn6, pawn, geom, slaf, dront, vds

# levels present in gt06 for 2009:
table(all.regions.2009$gt06)
# 0    1    2    3    4    5    6    7    8    9   10   11
#78   10  294   23  749  589  302  844  639 1351  730  112
gt06.levels <- levels(all.regions.2009$gt06)
```

```
SON.levels <- levels(df.SON$gt06)
dim(df.SON)
for (ii in 1:length(SON.levels)){
  if (sum(SON.levels[ii]==gt06.levels)==0)
    df.SON$gt06[df.SON$gt06==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$gt06)>0)) {
  df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$gt06),])
  df.SON <- df.SON[!is.na(df.SON$gt06),]
}
gc()
gc()

vds.levels <- levels(all.regions.2009$vds)
SON.levels <- levels(df.SON$vds)
dim(df.SON)
for (ii in 1:length(SON.levels)){
  if (sum(SON.levels[ii]==vds.levels)==0)
    df.SON$vds[df.SON$vds==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$vds)>0)) {
  print(sum(is.na(df.SON$vds)>0))
  df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$vds),])
  df.SON <- df.SON[!is.na(df.SON$vds),]
}
gc()
gc()

lgn6.levels <- levels(all.regions.2009$lgn6)
SON.levels <- levels(df.SON$lgn6)
dim(df.SON)
for (ii in 1:length(SON.levels)){
 if (sum(SON.levels[ii]==lgn6.levels)==0)
   df.SON$lgn6[df.SON$lgn6==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$lgn6)>0)) {
 print(sum(is.na(df.SON$lgn6)>0))
 df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$lgn6),])
 df.SON <- df.SON[!is.na(df.SON$lgn6),]
}

gc()
gc()
pawn.levels <- levels(all.regions.2009$pawn)
SON.levels <- levels(df.SON$pawn)
dim(df.SON)
for (ii in 1:length(SON.levels)){
  if (sum(SON.levels[ii]==pawn.levels)==0)
    df.SON$pawn[df.SON$pawn==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$pawn)>0)) {
  print(sum(is.na(df.SON$pawn)>0))
  df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$pawn),])
  df.SON <- df.SON[!is.na(df.SON$pawn),]
}

gc()
gc()
geom.levels <- levels(all.regions.2009$geom)
SON.levels <- levels(df.SON$geom)
dim(df.SON)
for (ii in 1:length(SON.levels)){
  if (sum(SON.levels[ii]==geom.levels)==0)
    df.SON$geom[df.SON$geom==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$geom)>0)) {
  print(sum(is.na(df.SON$geom)>0))
  df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$geom),])
  df.SON <- df.SON[!is.na(df.SON$geom),]
}

gc()
gc()
slaf.levels <- levels(all.regions.2009$slaf)
SON.levels <- levels(df.SON$slaf)
dim(df.SON)
```

```
for (ii in 1:length(SON.levels)){
  if (sum(SON.levels[ii]==slaf.levels)==0)
    df.SON$slaf[df.SON$slaf==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$slaf)>0)) {
  print(sum(is.na(df.SON$slaf)>0))
  df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$slaf),])
  df.SON <- df.SON[!is.na(df.SON$slaf),]
}

gc()
gc()
dront.levels <- levels(all.regions.2009$dront)
SON.levels <- levels(df.SON$dront)
dim(df.SON)
for (ii in 1:length(SON.levels)){
  if (sum(SON.levels[ii]==dront.levels)==0)
    df.SON$dront[df.SON$dront==SON.levels[ii]] <- NA
}
if(sum(is.na(df.SON$dront)>0)) {
  print(sum(is.na(df.SON$dront)>0))
  df.SON.NA <- rbind(df.SON.NA,df.SON[is.na(df.SON$dront),])
  df.SON <- df.SON[!is.na(df.SON$dront),]
}

gc()
gc()


# no removal of missing values from continuous variables (should not be any)
df.SON.NA$pred <- NA
df.SON$pred <- predict(all.regions.2009.mod2, df.SON)

gc()
gc()

# combine both dataframes
df.SON.pred <- rbind(df.SON, df.SON.NA)
gc()
gc()

# reorder x and y in combined file
df.SON.pred <- df.SON.pred[order(df.SON.pred$y, df.SON.pred$x),]
gc()
gc()
gc()

# add (df.SON.pred["pred"]) to SON
SON$pred <- df.SON.pred$pred
names(SON)
gc()
gc()
gc()
gc()
gc()

# temporalily save dataframes and remove from memory to free space for kriging calc

save(df.SON.NA, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/df.SON.NA")
save(df.SON.pred, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/df.SON.pred")
rm(df.SON.NA, df.SON.pred)
gc()
gc()

# save(SON, file="SON")
# export to Asciigrid to examine in ArcGIS
SON$btpred <- 10^(SON$pred) # ! just for quick look, image also in use for prediction locations
writeGDAL(SON["btpred"],     "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.mask09s.tif",
drivername = "GTiff", type= "Float32", mvFlag=-9999)
# liberate memory
rm(SON)
rm(lgn6.levels,  pawn.levels,  slaf.levels,  dront.levels,  vds.levels,  geom.levels,  gt06.levels,
SON.levels, ii)

gc()
gc()
```

```
gc()
gc()

### STEP 5  Kriging

# Add field with recently calculated residuals from lm to all.regions.2009
# check when ok, otherwise ignore

all.regions.2009$residuals <- residuals(all.regions.2009.mod2)

# kriging preparation ------------------------------------------------


# calculate variogram of residuals 2009; check if variogram is OK.
# Should be equal for # every region in all.regions.2009!
g.res <- gstat(formula = residuals ~1, data = all.regions.2009)
vg.res <- variogram(g.res, boundaries = c(50,100,200,400,600,1000,1600,2400,3000)/5)

# kriging range indicates kriging exerts especially a local effect

plot(vg.res, plot.nu = T)
vgm.res <- vgm(nugget=0.1, psill=0.20, range=600, model="Sph") #adjust to fit
vgm.res <- fit.variogram(vg.res,vgm.res)

win.graph(7, 5, 12)
plot(vg.res, vgm.res, main = "Residuals", plot.nu = T)
savePlot(filename="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/variogram all.regions north 2009",
+ type="png")
dev.off()

#load prediction mask
predgrid <-readGDAL("S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.mask09s.tif")
proj4string(predgrid) <- CRS("+init=epsg:28992")

# point kriging residual
# sometimes an error occurs with singularity issues "Error in predict.gstat(g, newdata = predgrid,
# block = block, nsim = nsim, : LDLfactor".
# this applies to a restricted number of cells. To prevent losing all output, add the # following
# command option to krige: "set=list(cn_max=1e10),"
# commence kriging and keep time
system.time(df.SON.kr  <-  krige(residuals  ~1,  all.regions.2009,  newdata  =  predgrid,  vgm.res,
nmax=100, set=list(cn_max=1e10), debug.level=-1))
names(df.SON.kr)[1] = "res.pred"
names(df.SON.kr)[2] = "res.var"
names(df.SON.kr)

gc()
gc()
gc()

# add regression prediction(df.SON.pred)to residual kriging part (df.SON.kr)
load("S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/df.SON.pred")

df.SON.kr$pred <- df.SON.pred$pred + df.SON.kr$res.pred
df.SON.kr$var <- df.SON.kr$res.var
df.SON.kr$sd <- sqrt(df.SON.kr$var)

# Backtransform predicted logno3; make new variable btfno3
df.SON.kr$btfno3 <- 10^(df.SON.kr$pred+0.5*df.SON.kr$var) # for log10

gc()
gc()
gc()

# export grids for use in GIS
writeGDAL(df.SON.kr["btfno3"],     "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all4btno3_09s.tif",
drivername = "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.SON.kr["var"],        "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all4n_var_09s.tif",
drivername = "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.SON.kr["sd"],          "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all4n_sd_09s.tif",
drivername = "GTiff", type= "Float32", mvFlag=-9999)

#save final results regression kriging all.regions 2009, south part
save(df.SON.kr, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/df.SON.kr.09")
save(all.regions.2009, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.2009s")
```

```
save(all.regions.2009.mod2,
file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.2009s.mod2")
save(vgm.res, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.2009s.vgm.res")
save(vg.res, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.2009s.vg.res")
save(g.res, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.2009s.g.res")

rm(df.SON, df.SON.pred, predgrid)
gc()

# Uncertainties

# check if still loaded in memory, otherwise:
load("S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/df.SON.kr.09")
# calculate 2.5% lower boundary
df.SON.kr$lb <- df.SON.kr$pred - 1.96 * df.SON.kr$sd
# calculate 97.5% upper boundary
df.SON.kr$ub <- df.SON.kr$pred + 1.96 * df.SON.kr$sd

# Backtransform LB predicted logno3;
df.SON.kr$lbtr <- 10^(df.SON.kr$lb) # for log10
df.SON.kr$ubtr <- 10^(df.SON.kr$ub) #  for log10
df.SON.kr$conf95 <- df.SON.kr$ubtr - df.SON.kr$lbtr


# export grids for use in GIS
writeGDAL(df.SON.kr["lbtr"],
+ "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.s09_lb.tif", drivername =
+ "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.SON.kr["ubtr"],
+ "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.s09_ub.tif", drivername =
+ "GTiff", type= "Float32", mvFlag=-9999)
writeGDAL(df.SON.kr["conf95"],
+ "S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions.s09_conf.tif", drivername =
+ "GTiff", type= "Float32", mvFlag=-9999)

# Repeat STEPS 4 and 5 for the other three regions: CON, EON and NON
# in order to make nationwide map

### STEP 6 Crossvalidation
# identical for each region; no need to perform for other regions
# since data = all.regions.2009 and model and variogram remain unchanged

# Leave one out crossvalidation (LOOCV)
all.regions.2009.cv <- krige.cv(residuals~1, all.regions.2009, model = vgm.res,
+ nmax=100, nfold = nrow(all.regions.2009))

summary(all.regions.2009.cv)
var(all.regions.2009.cv$zscore)
mean(all.regions.2009.cv$zscore)
hist(all.regions.2009.cv$zscore)
bubble(all.regions.2009.cv, z="residual")

# save LooCV results
save(all.regions.2009.cv, file="S:/R/jongdc/GIMA/Data/modeloutput/all4/2009/all.regions_s09.cv")
```

Appendix II – Utilized software and versions

Extended R-workstation based on multi-core 64 bit Windows7 and 16Gb of memory

R-version 3.0.3 (2013-09-25 “Warm Puppy”) with RStudio 0.98.1028

Packages:

sp	1.0-15
gstat	1.0-19
geospt	1.0-0
rgdal	0.8-16
lattice	0.20-29
zoo	1.7-11
spacetime	1.1-0
xts	0.9-7
foreign	0.8-61
shapefiles	0.7
intervals	0.14.0

ArcGIS 9.3.1 with extensions (Spatial Analyst)
ArcGIS10.1 with extensions (Esri site license)

Besides the Windows-machine, a RedHat linux server was available with 32 Gb memory and multi-core processing, sporting RStudio Server. R-code is generally compatible between the platforms, save the occasional specific screen output command and directory structure.

The R-environment can be downloaded freely from http://cran.r-project.org/ and is available for Windows, Linux and MAC-OSx platforms.

R-studio is available for download at http://www.rstudio.com/.

ArcGIS is developed and maintained by Esri, Redlands USA and works on the Windows platform.

Appendix III - Modelbuilder™ diagram

Modelbuilder is a schematic scripting tool in ArcGIS, enabling automating processes using a graphic representation of commands. The script for tailoring covariate maps to equally aligned cell-sizes (25m resolution) and dividing of the four regions is listed in Figure 26.

Each of the auxiliary maps needs to be available in the same grid format with a cell size of 25 m and aligned to the same origin. Next, each of the maps is cut, matching the region boundaries to produce an auxiliary map dataset for each of the four regions. This is explained in Figure 25, using ESRI's ArcGIS Modelbuilder functions. A blue oval shape depicts an input map, yellow squares are processes and green ovals are intermediate or end products.

Example for variable *gt06* (groundwater tables)

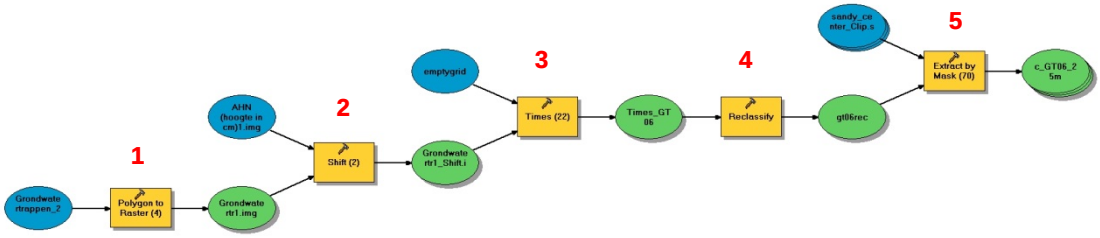


Figure 25. Modelbuilder diagram for variable gt06

First, the map is transformed from vector to raster in 25 m grid cells (1). The output is then aligned with the map AHN (2) and the cell size is adjusted (3), after which the same origin and cell size are equal for the whole research project. In this particular case, a reclassification of the categories is needed (4) to reduce the possible number of values from 23 to 12. The last step breaks down the map into four regions (5) by using a regional mask. This procedure is similar for the other auxiliary parameter maps.

The combination of all map operations yields the following model (Figure 26), making it possible to generate all necessary data sets from all map variables. Unfortunately the scale is insufficient to show details, though the four regional output groupings can be recognized where many lines converge. The AHN-image is visible at the left, from which all map extents are derived.

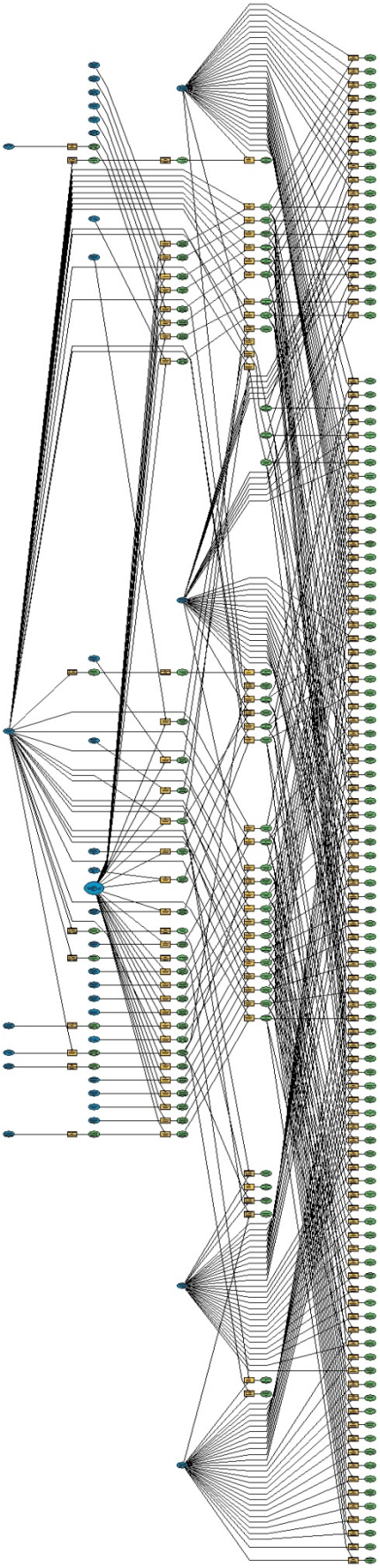


Figure 26. Modelbuilder diagram for all variables

Appendix IVa – Model summary

YEAR-REGIONAL MODELS

Model name (.mod)	**Parameters**	**#data points**	**R^2**	**R2a**
north.2007	om60 + om100 + om120 + gt06 + bbg06 + gronds + kwel2 + pawn + lgn6 + laf + slont	1983	0.493	0.473
north.2008	om60 + om100 + **gt06*** + ahn + bbg06 + kwel2 + pawn + lgn6 + lont + slaf	2167	0.504	0.488
north.2009	stone7 + stone8 + **gt06*** + ahn + bbg06 + kwel2 + lgn6 + laf + slont + vds	1979	0.390	0.372
east.2007	om10 + om40 + om80 + om120 + gt06 + bbg06 + kwel2 + geom + draf + slont	1364	0.331	0.310
east.2008	om10 + om60 + om80 + om100 + gt06 + stone6 + stone8 + lgn6 + draf	1180	0.307	0.288
east.2009	om25 + om80 + om100 + gt06 + ahn + kwel2 + lgn6 + draf	1249	0.266	0.248
centre.2007	gt06 + stone5 + stone6 + bbg06 + geom + vds	452	0.584	0.558
centre.2008	nhx + gt06 + geom	379	0.452	0.426
centre.2009	om10 + gt06 + stone5 + stone6 + stone8 + bbg06 + laf + vds	348	0.529	0.489
south.2007	om05 + om40 + om60 + gt06 + nhx + ahn + bbg06 + kwel2 + lgn6 + geom	1603	0.493	0.481
south.2008	om05 + om10 + om40 + om60 + gt06 + stone5 + stone6 + stone7 + nhx + bbg06 + kwel2 + lgn6 + geom + slaf + draf	2523	0.594	0.584
south.2009	om05 + om10 + om25 + om40 + om60 + gt06 + nhx + kwel2 + pawn + lgn6 + geom + slaf + slont	2137	0.417	0.398

***** = added manually, parameter originally rejected by automated model selection (see 4.1)

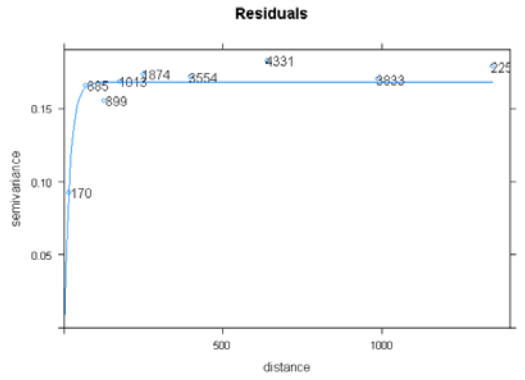
NATIONWIDE MODELS

Model name (.mod)	**Parameters**	**#data points**	**R^2**	**R2a**
all.regions.2007	om10 + stone5 + stone6 + nhx + gt06 + ahn + bbg06 + gronds + kwel2 + pawn + lgn6 + geom + dront + laf	5383	0.472	0.462
all.regions.2008	om60 + om80 + om100 + stone5 + stone6 + stone7 + stone8 + nhx + gt06 + ahn + bbg06 + kwel2 + pawn + lgn6 + geom + slaf + slont	6243	0.554	0.547
all.regions.2009	stone5 + stone6 + stone7 + stone8 + nhx + gt06 + ahn + kwel2 + pawn + lgn6 + geom + dront + slaf + vds	5721	0.435	0.426

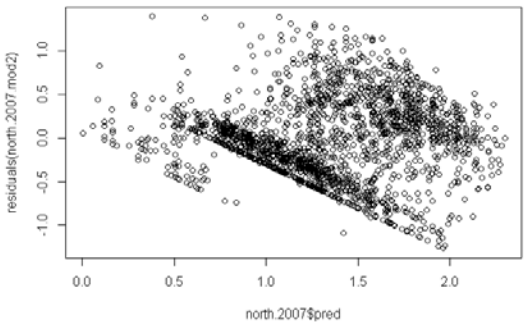
Appendix IVb – Variogrammes and data depiction

Data north 2007

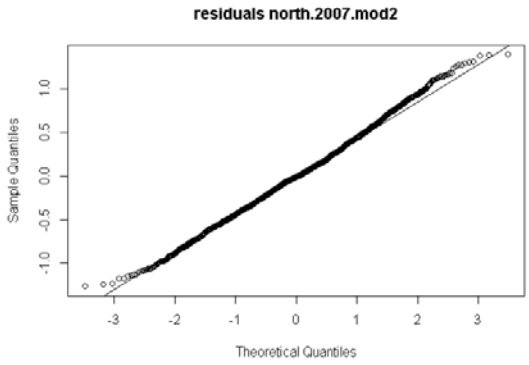
Semivariogram



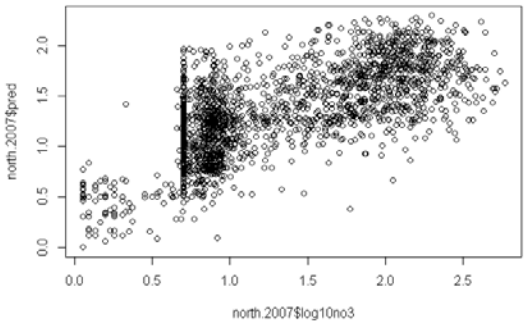
Residuals vs predictions

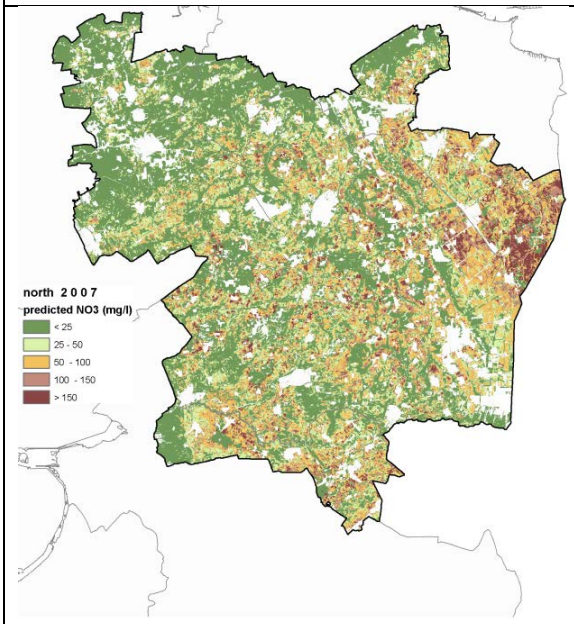


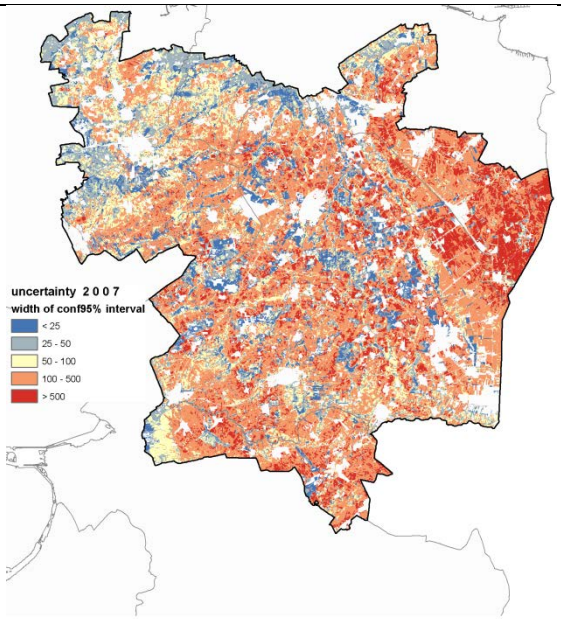
QQ-plot



Predictions vs observations

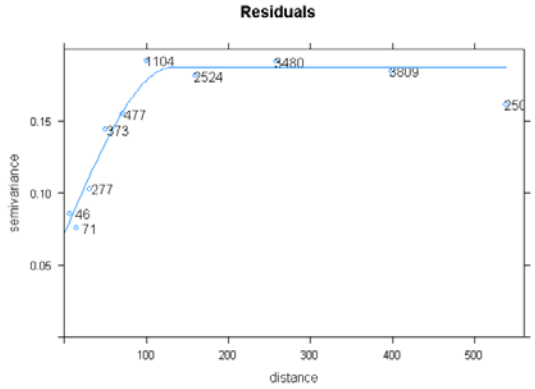




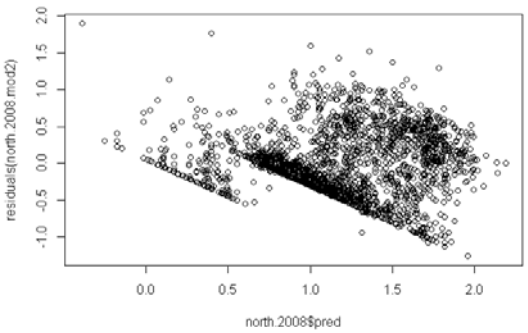


Data north 2008

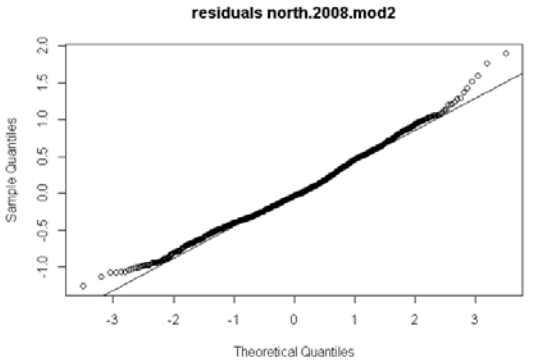
Semivariogram



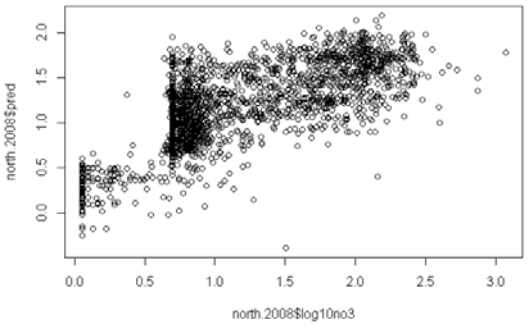
Residuals vs predictions

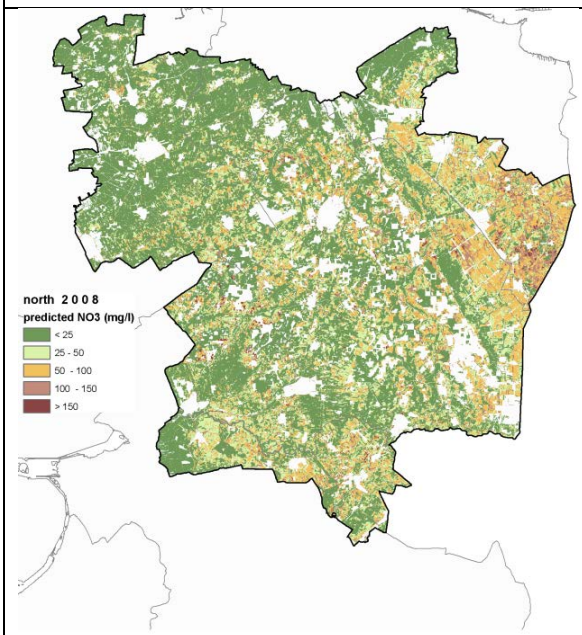


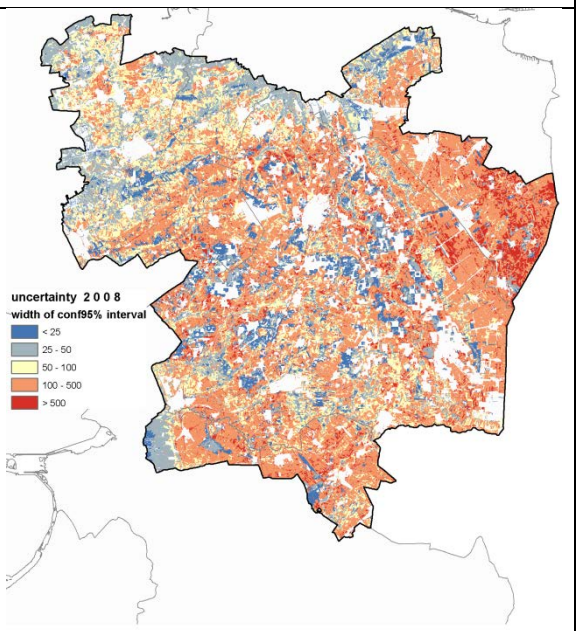
QQ-plot



Predictions vs observations

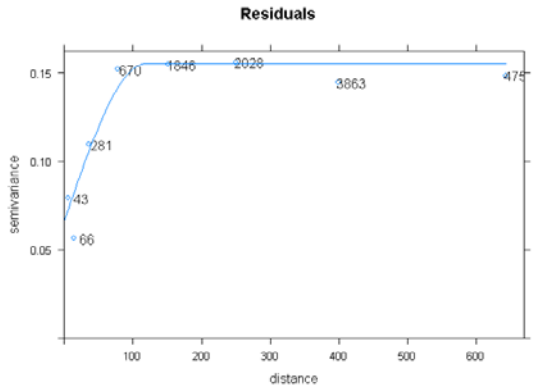




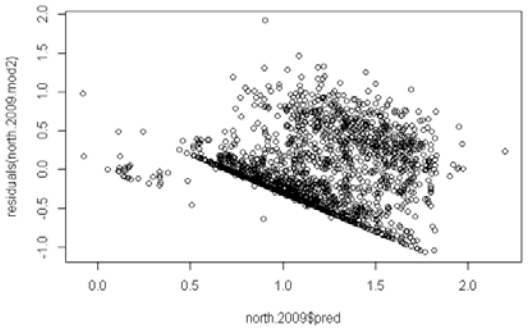


Data north 2009

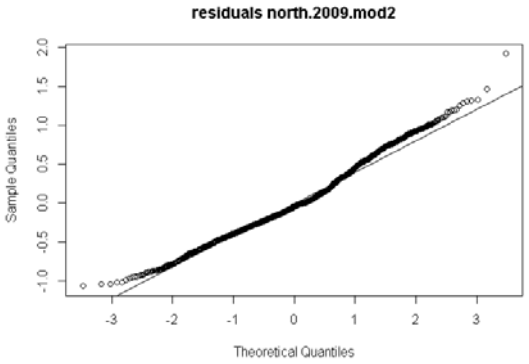
Semivariogram



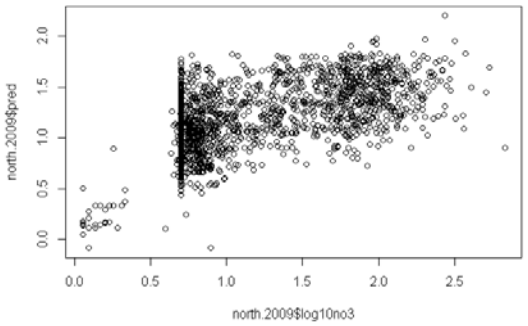
Residuals vs predictions

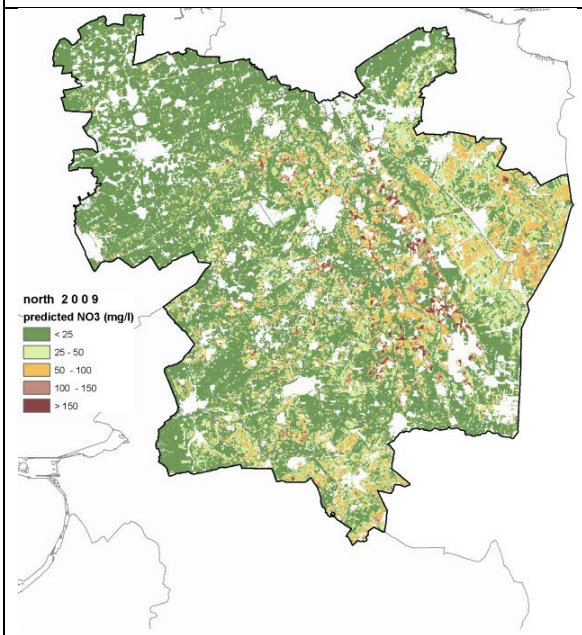


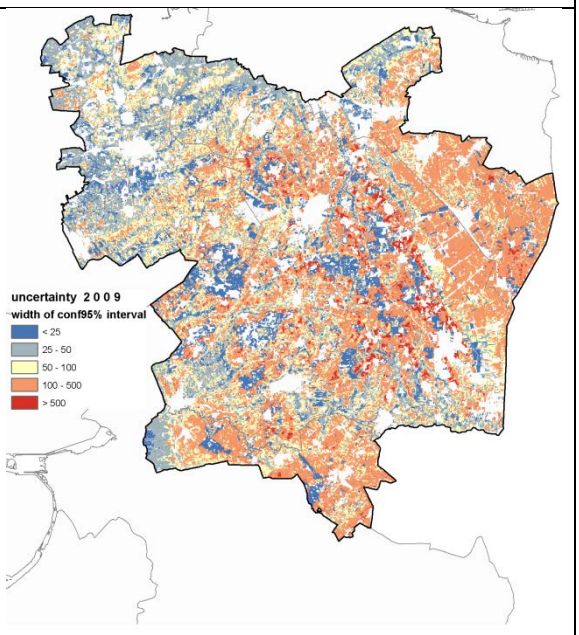
QQ-plot



Predictions vs observations

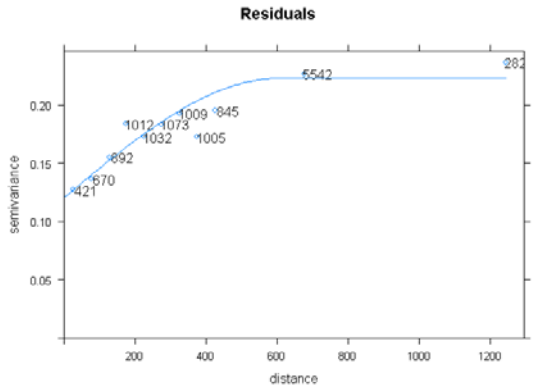




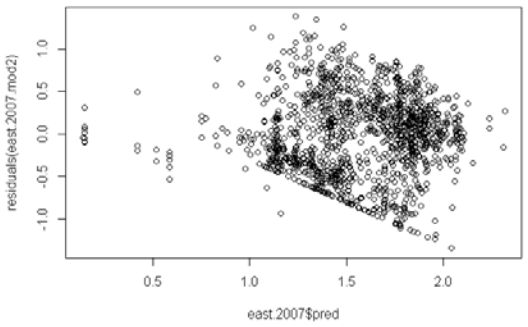


Data east 2007

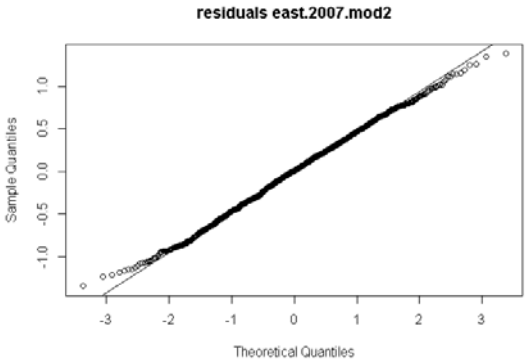
Semivariogram



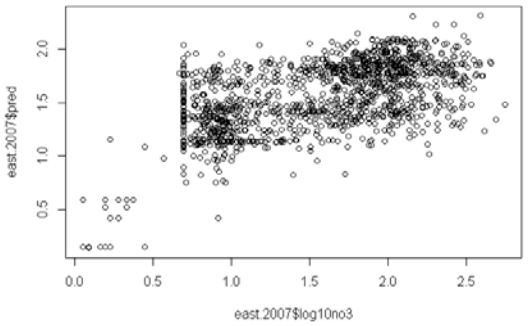
Residuals vs predictions

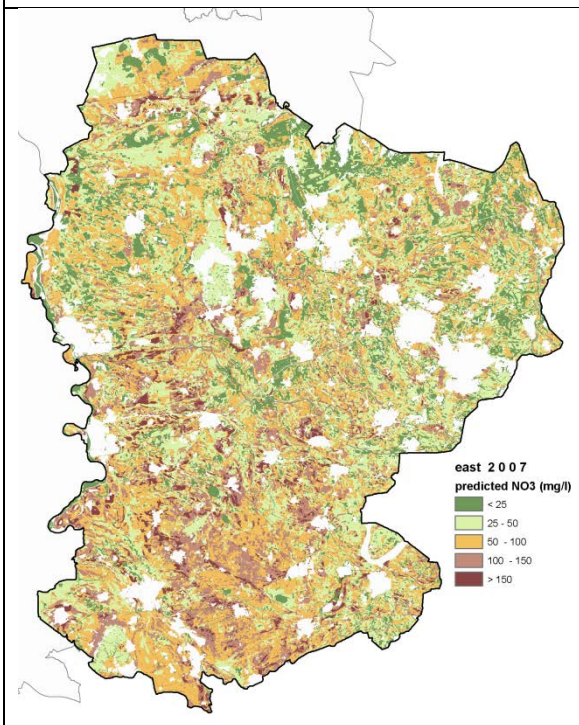


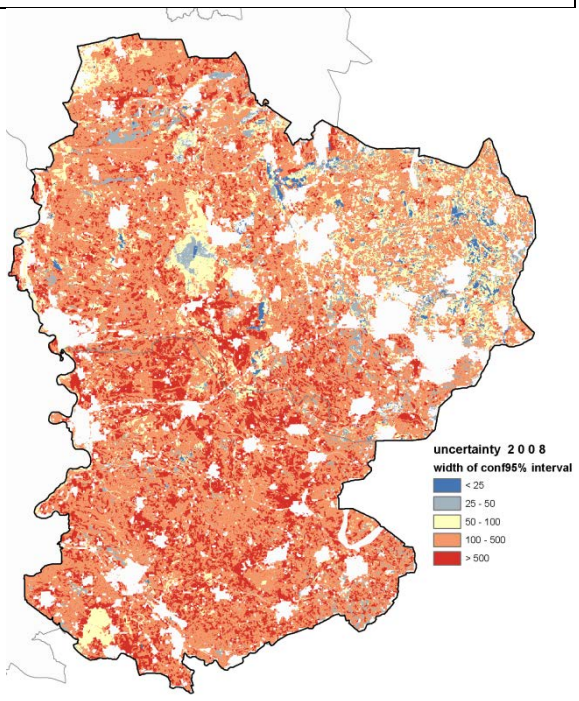
QQ-plot



Predictions vs observations

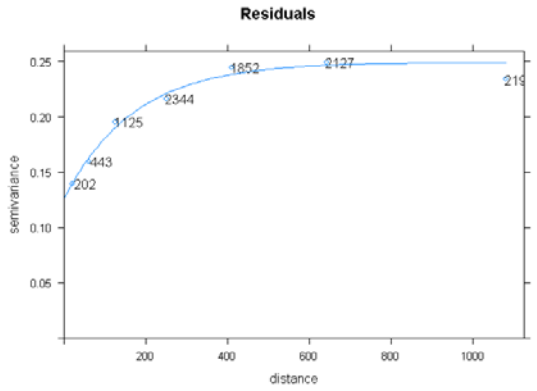




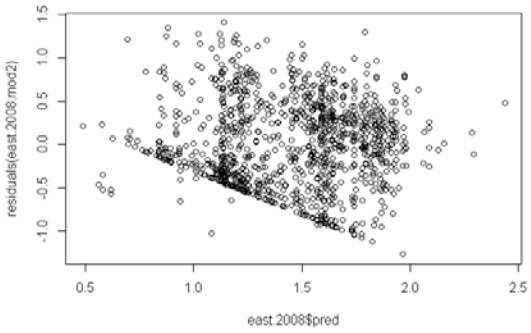


Data east 2008

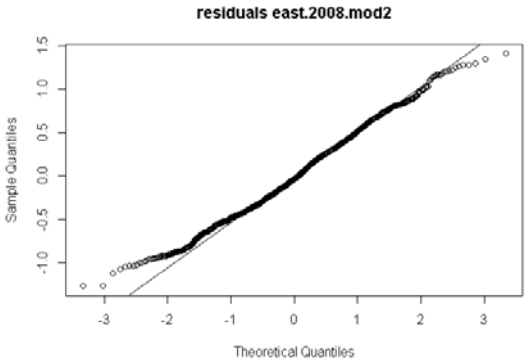
Semivariogram



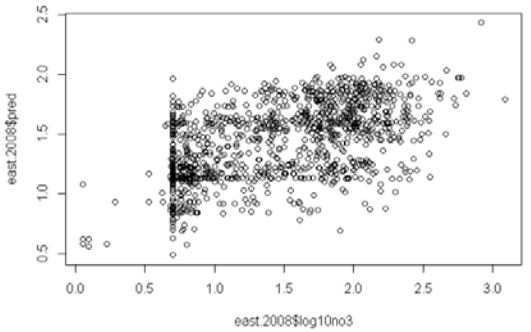
Residuals vs predictions

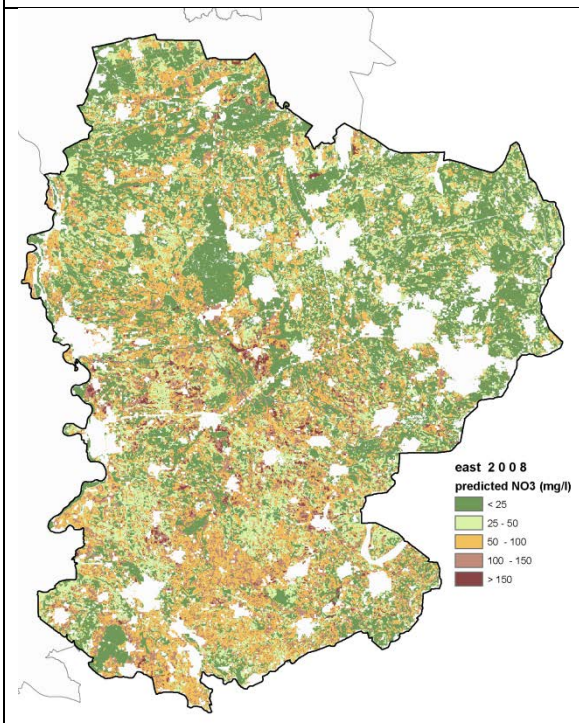


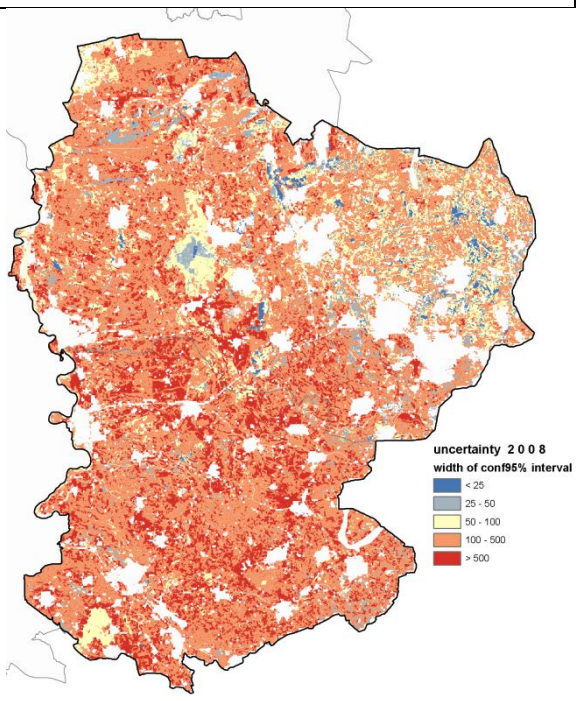
QQ-plot



Predictions vs observations

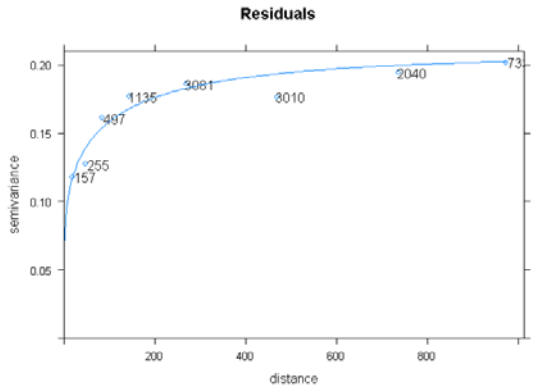




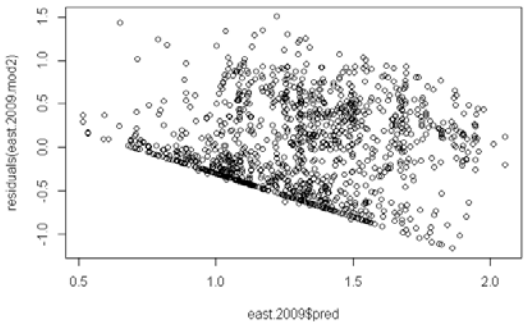


Data east 2009

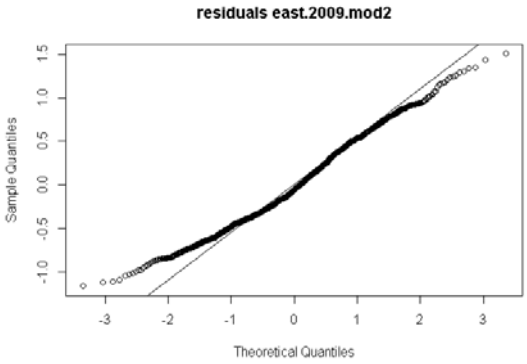
Semivariogram



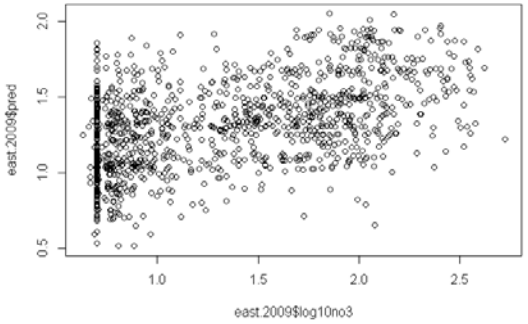
Residuals vs predictions

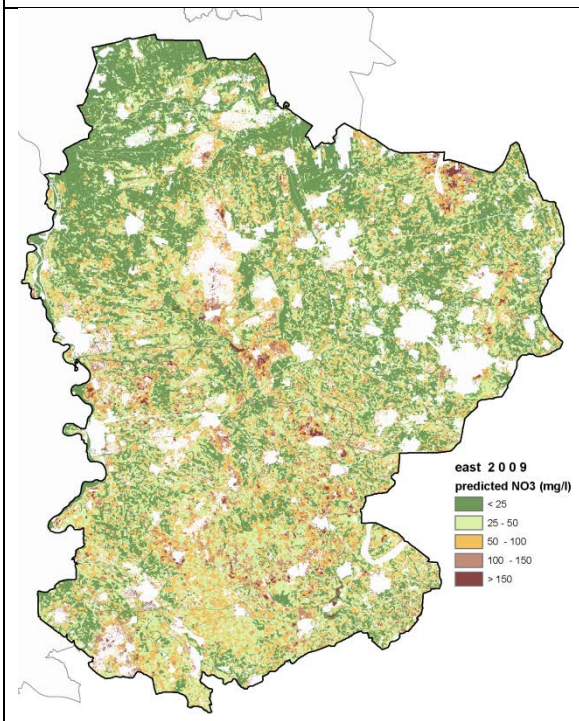


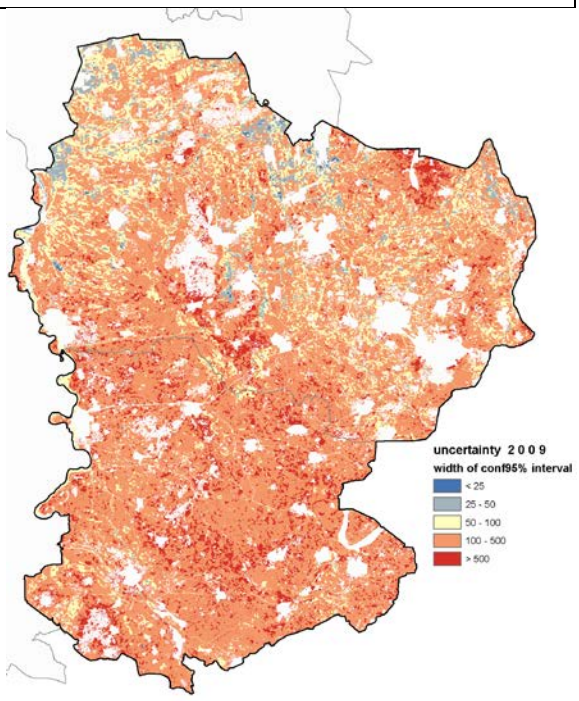
QQ-plot



Predictions vs observations

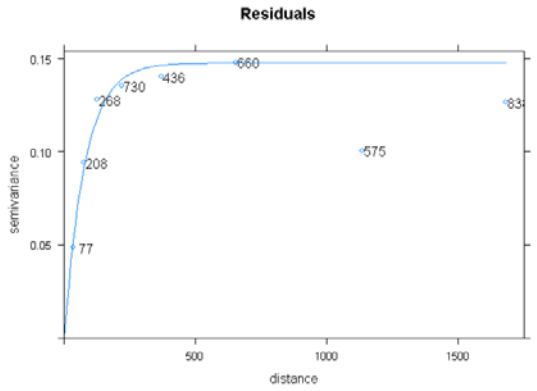




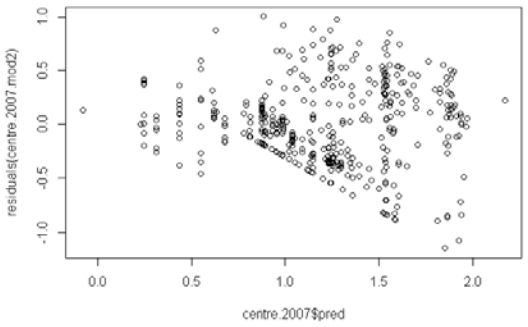


Data centre 2007

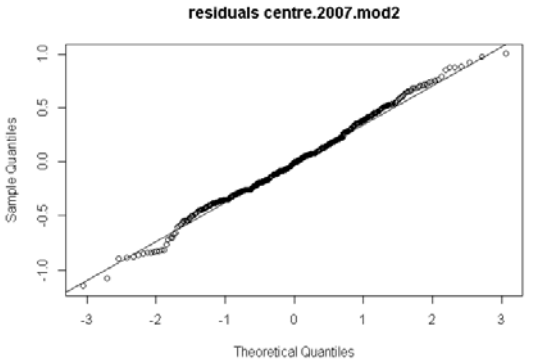
Semivariogram



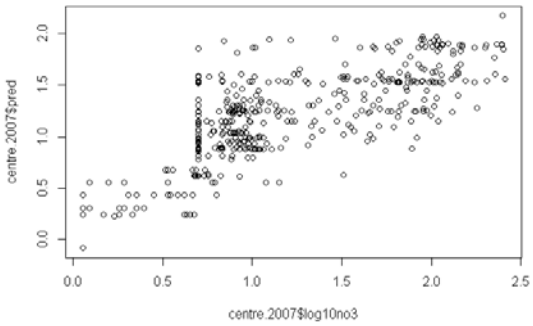
Residuals vs predictions

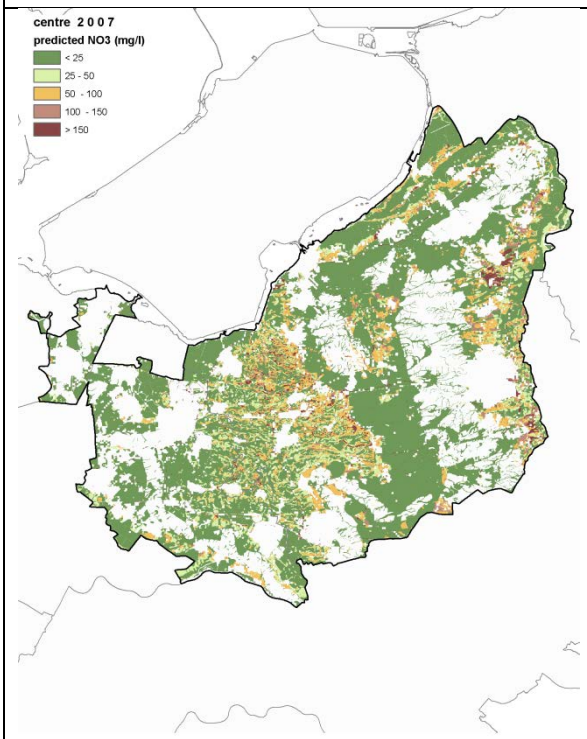


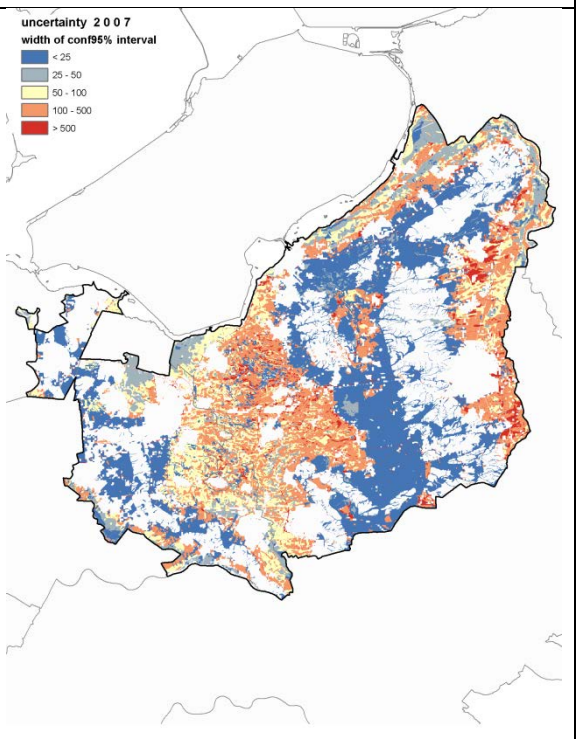
QQ-plot



Predictions vs observations

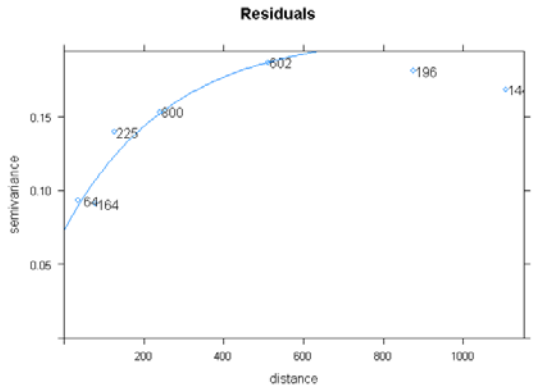




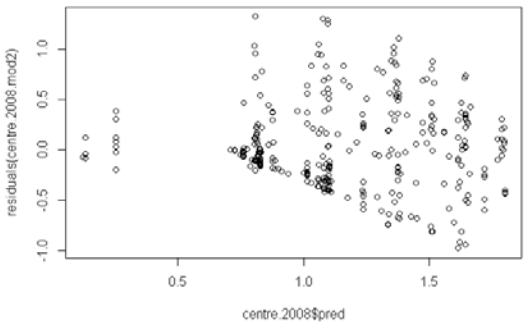


Data centre 2008

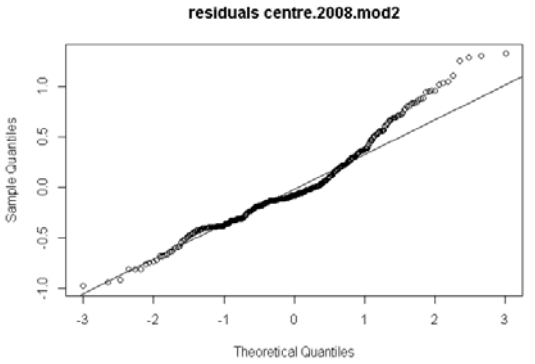
Semivariogram



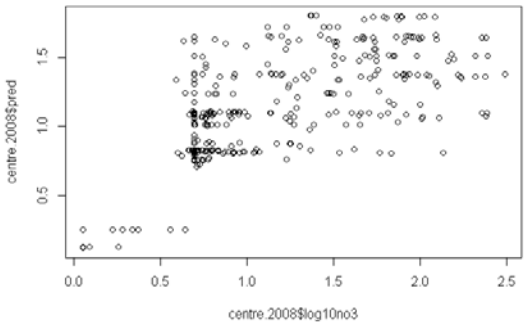
Residuals vs predictions

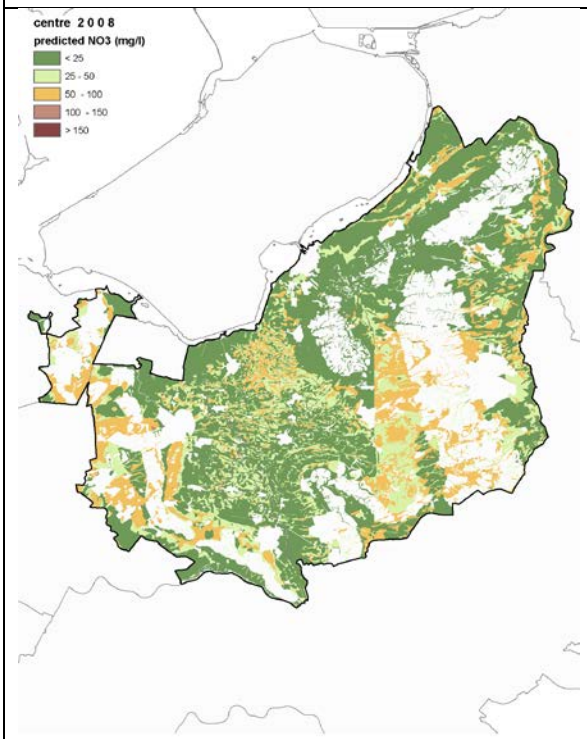


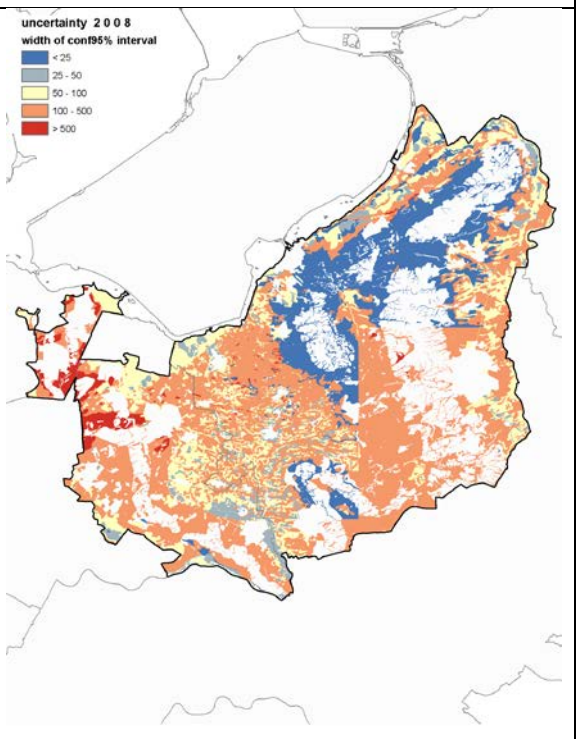
QQ-plot



Predictions vs observations

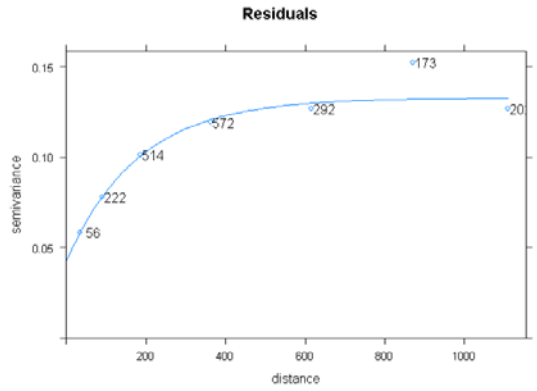




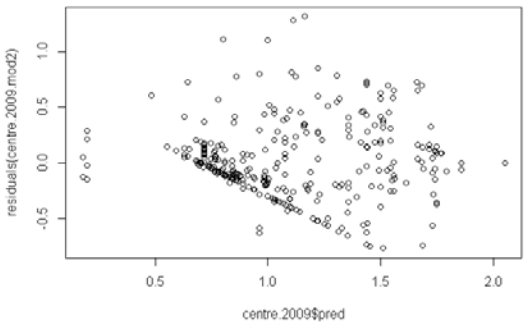


Data centre 2009

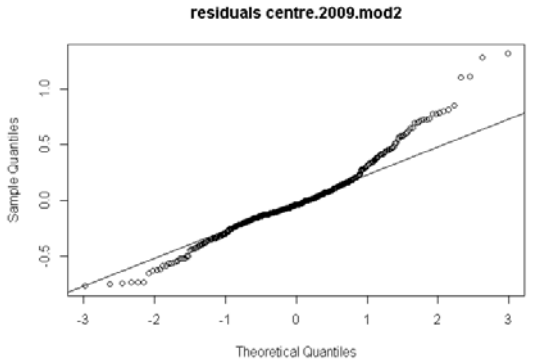
Semivariogram



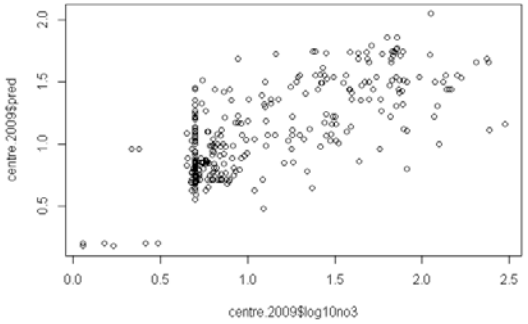
Residuals vs predictions

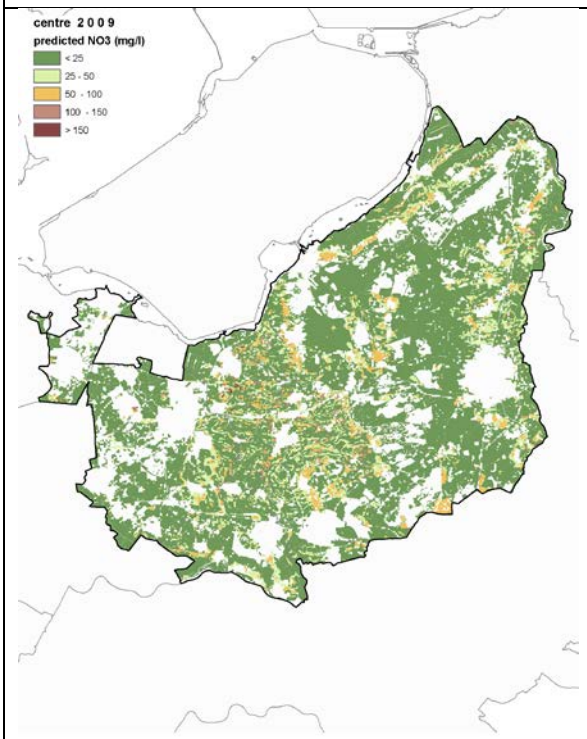


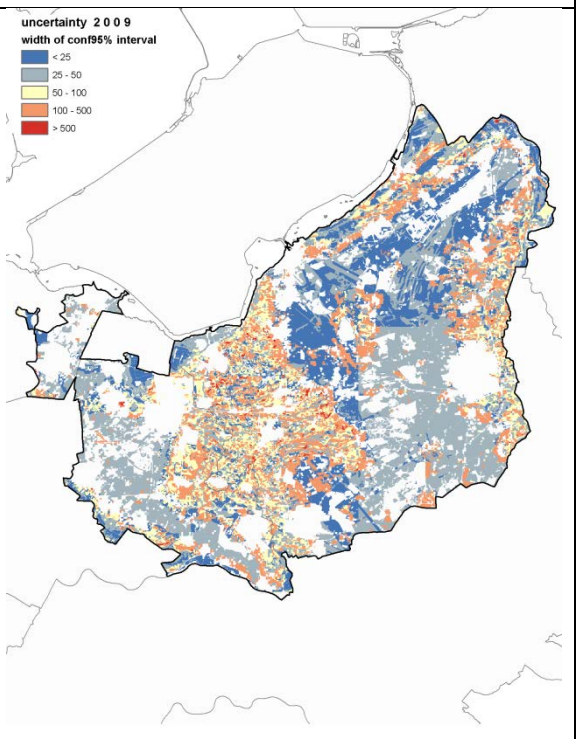
QQ-plot



Predictions vs observations

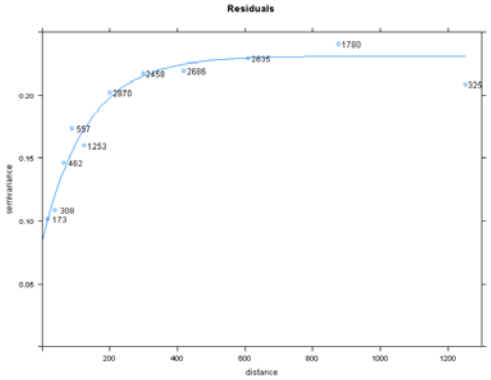




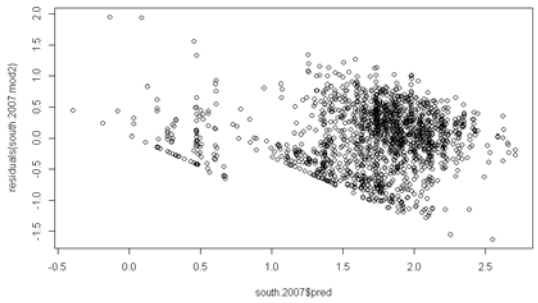


Data south 2007

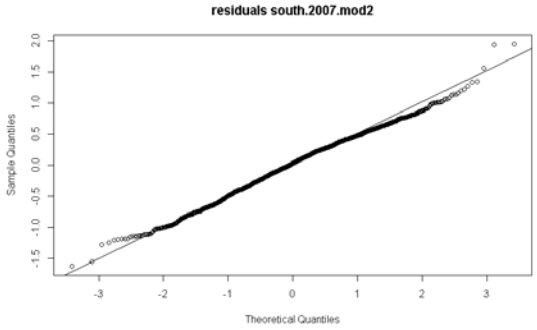
Semivariogram



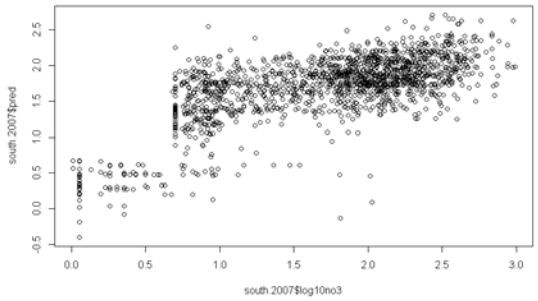
Residuals vs predictions

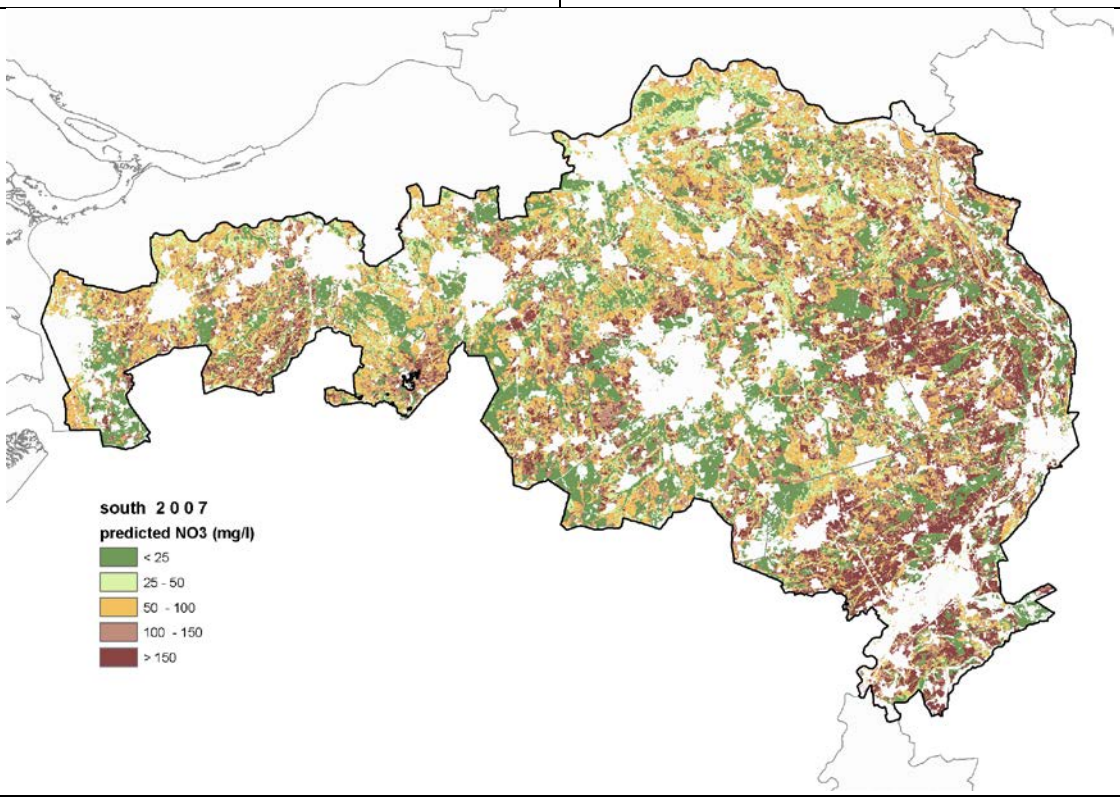


QQ-plot



Predictions vs observations

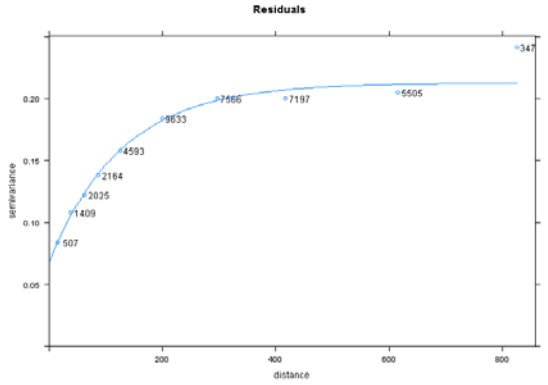




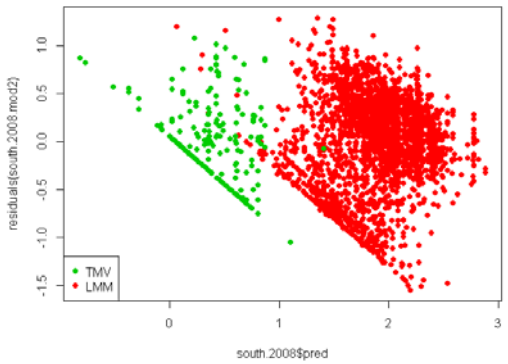
For 95%confidentiality map, see after data South 2009

Data south 2008

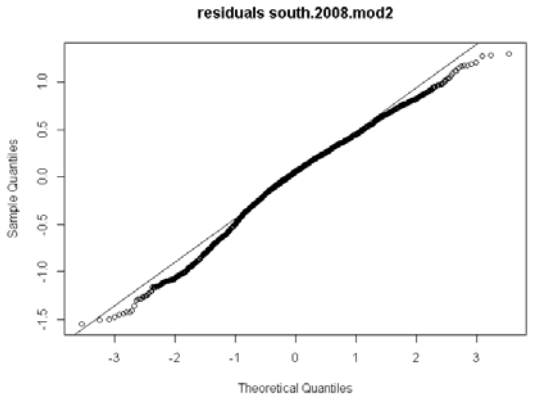
Semivariogram



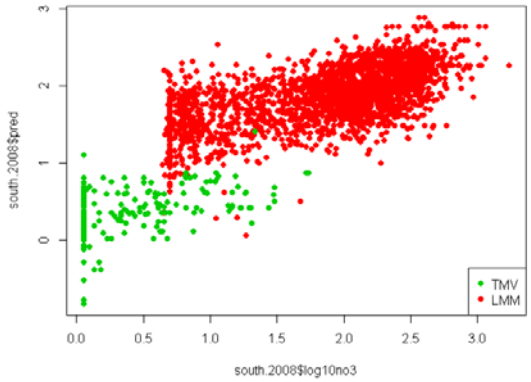
Residuals vs predictions

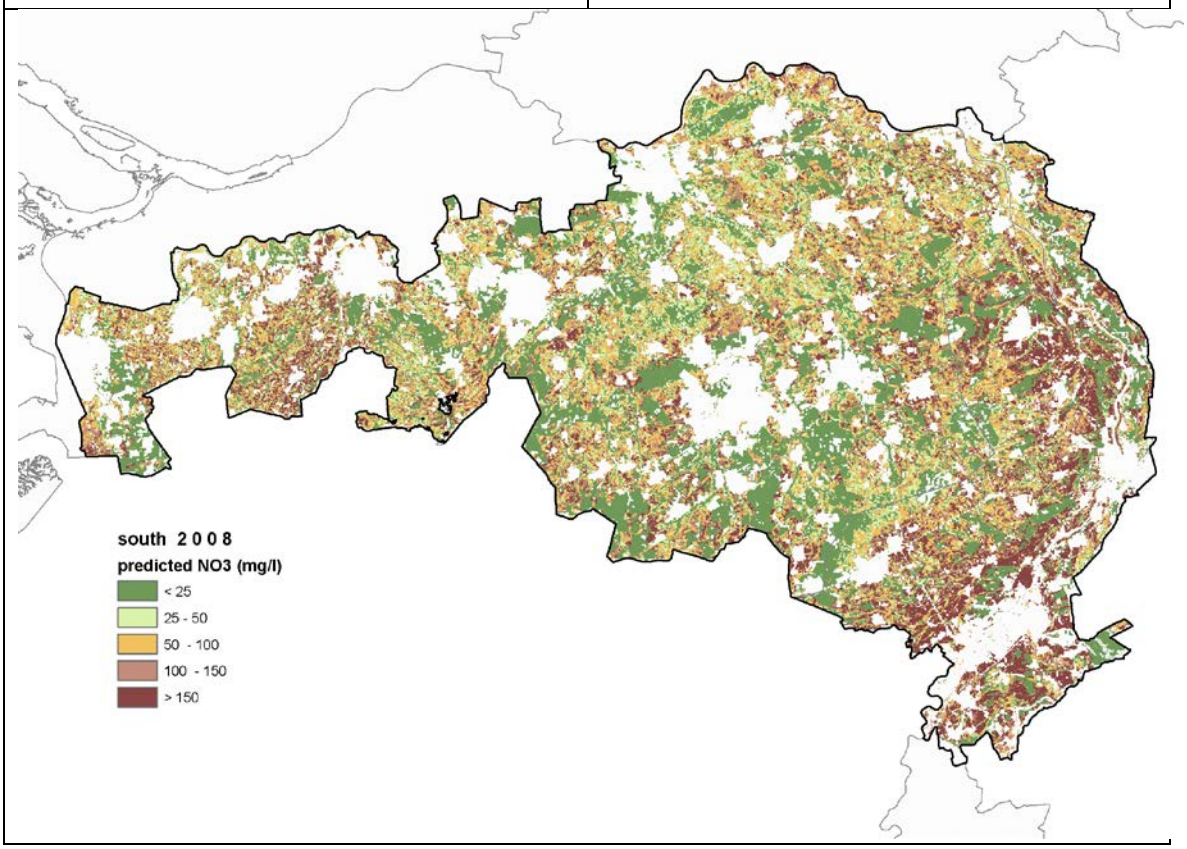


QQ-plot



Predictions vs observations

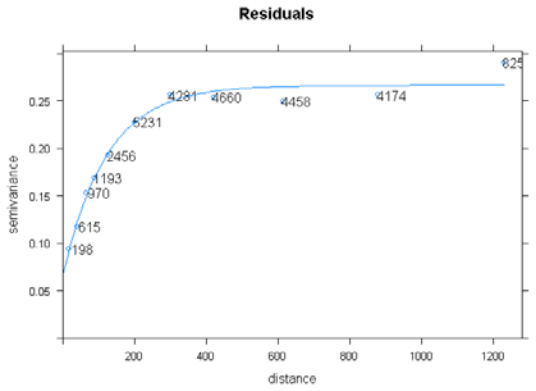




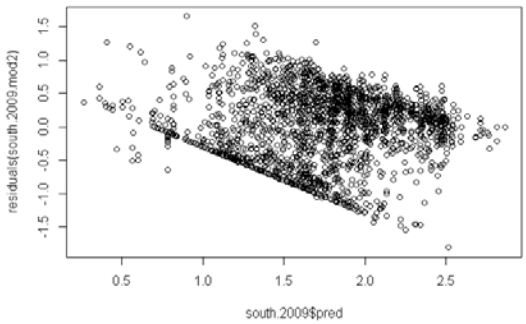
For 95%confidentiality map, see after data South 2009

Data south 2009

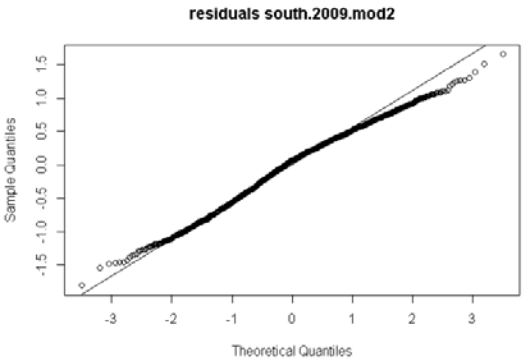
Semivariogram



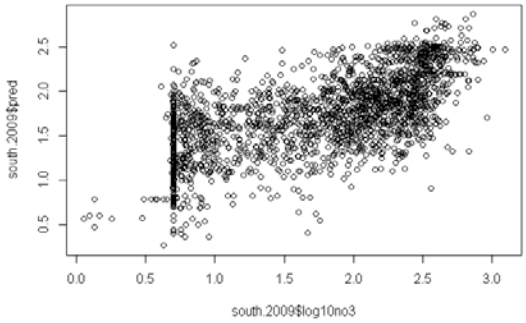
Residuals vs predictions

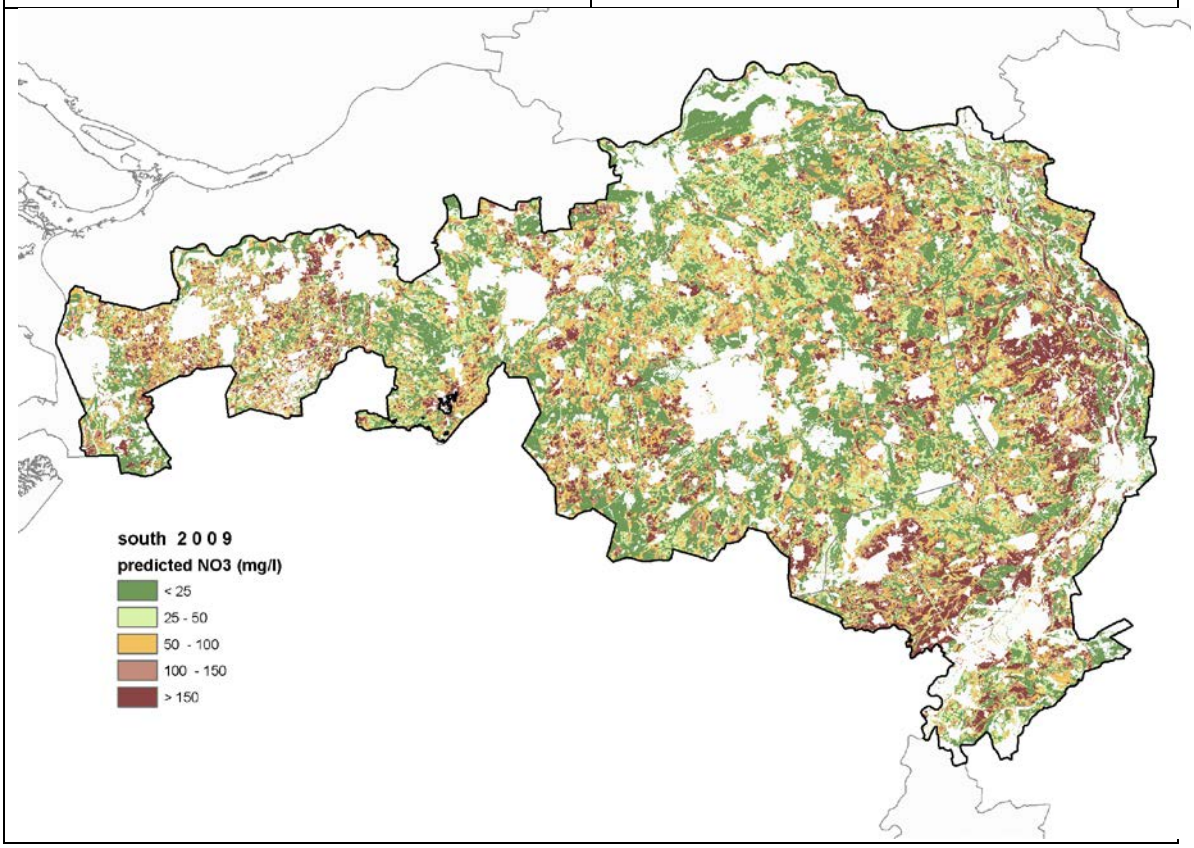


QQ-plot

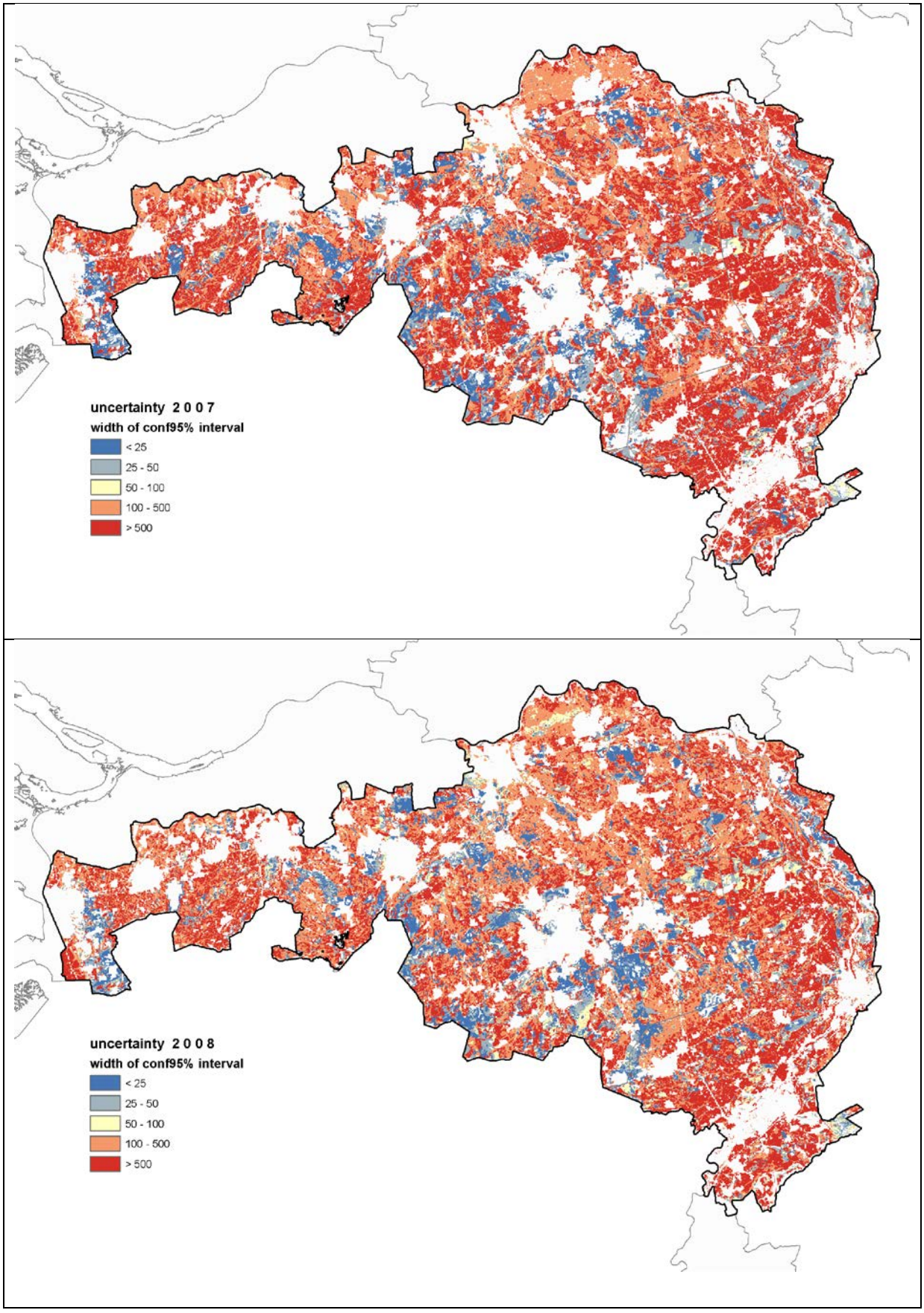


Predictions vs observations

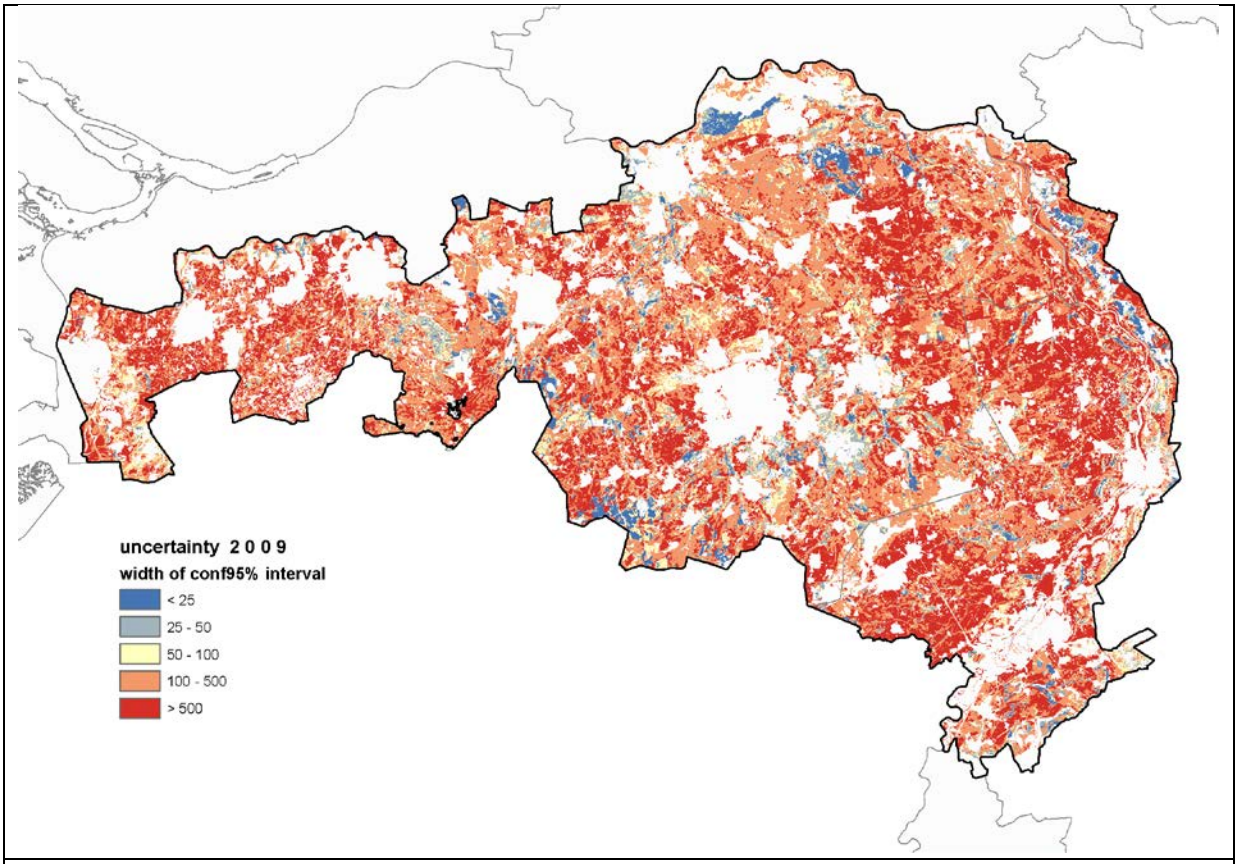




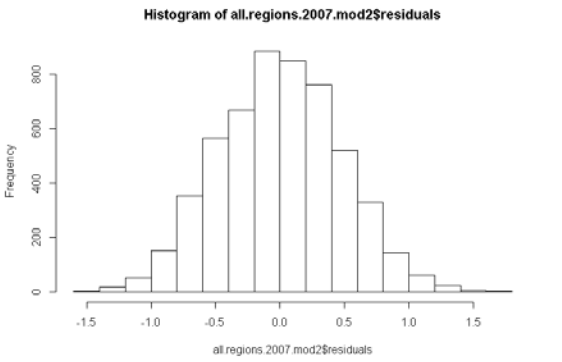
Data South 2007 and South 2008



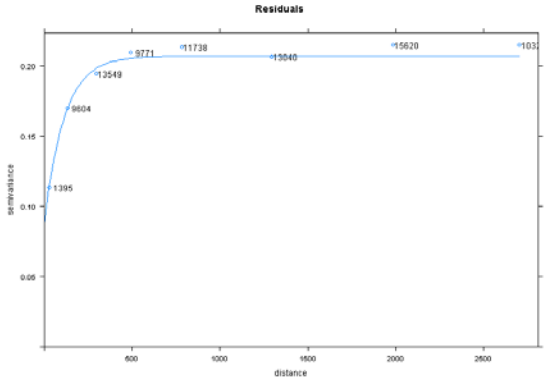
Data South 2009



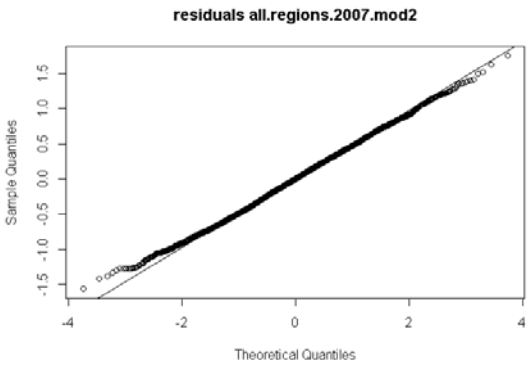
Data all regions 2007



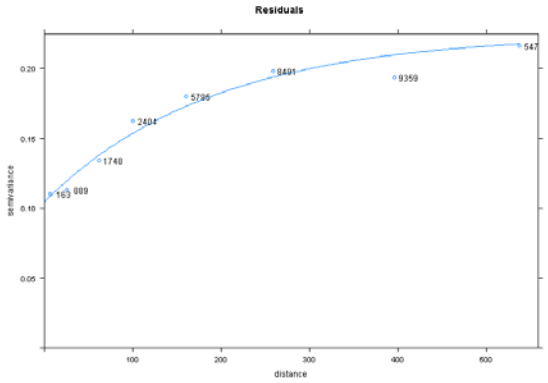
Semivariogram long range

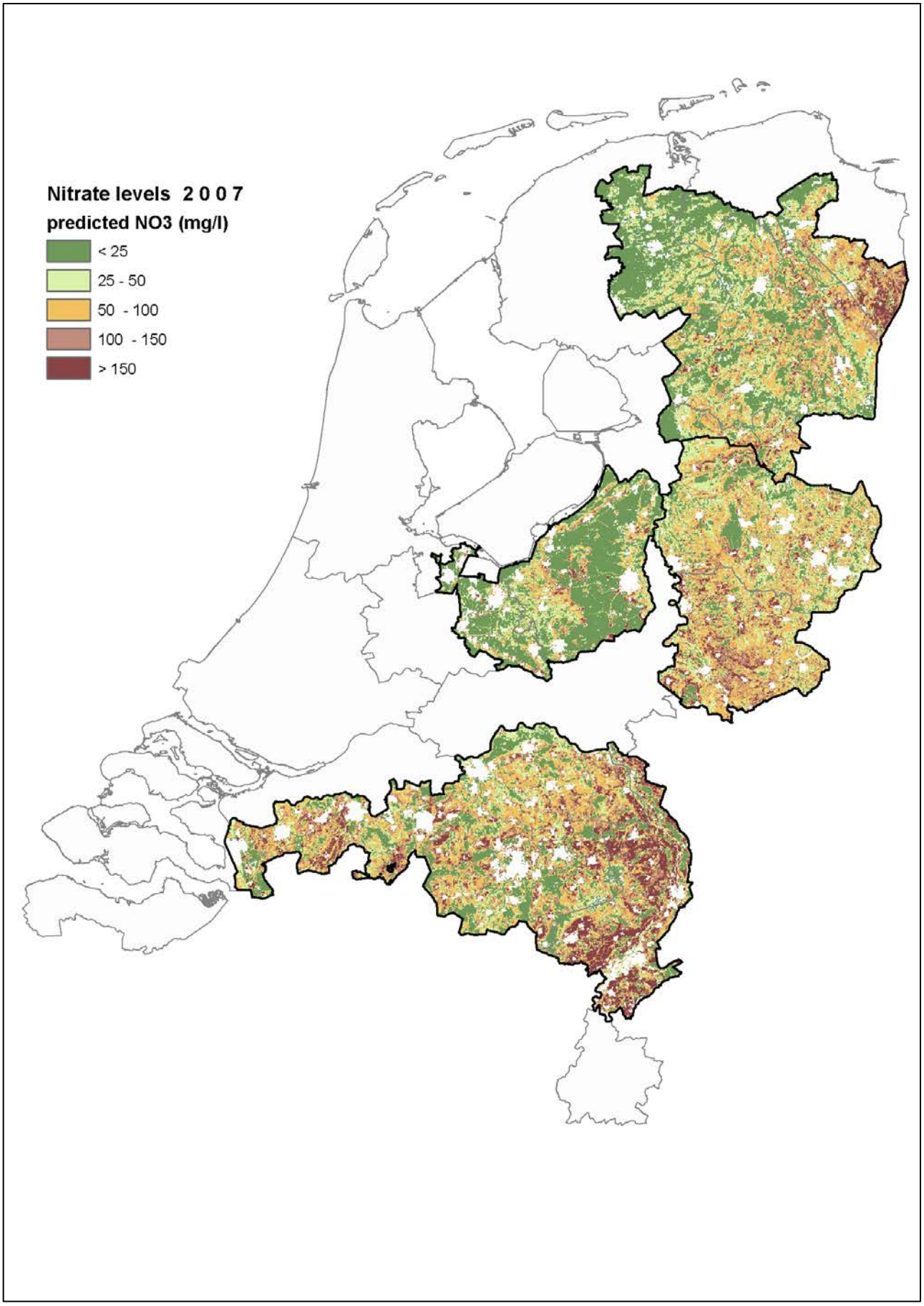


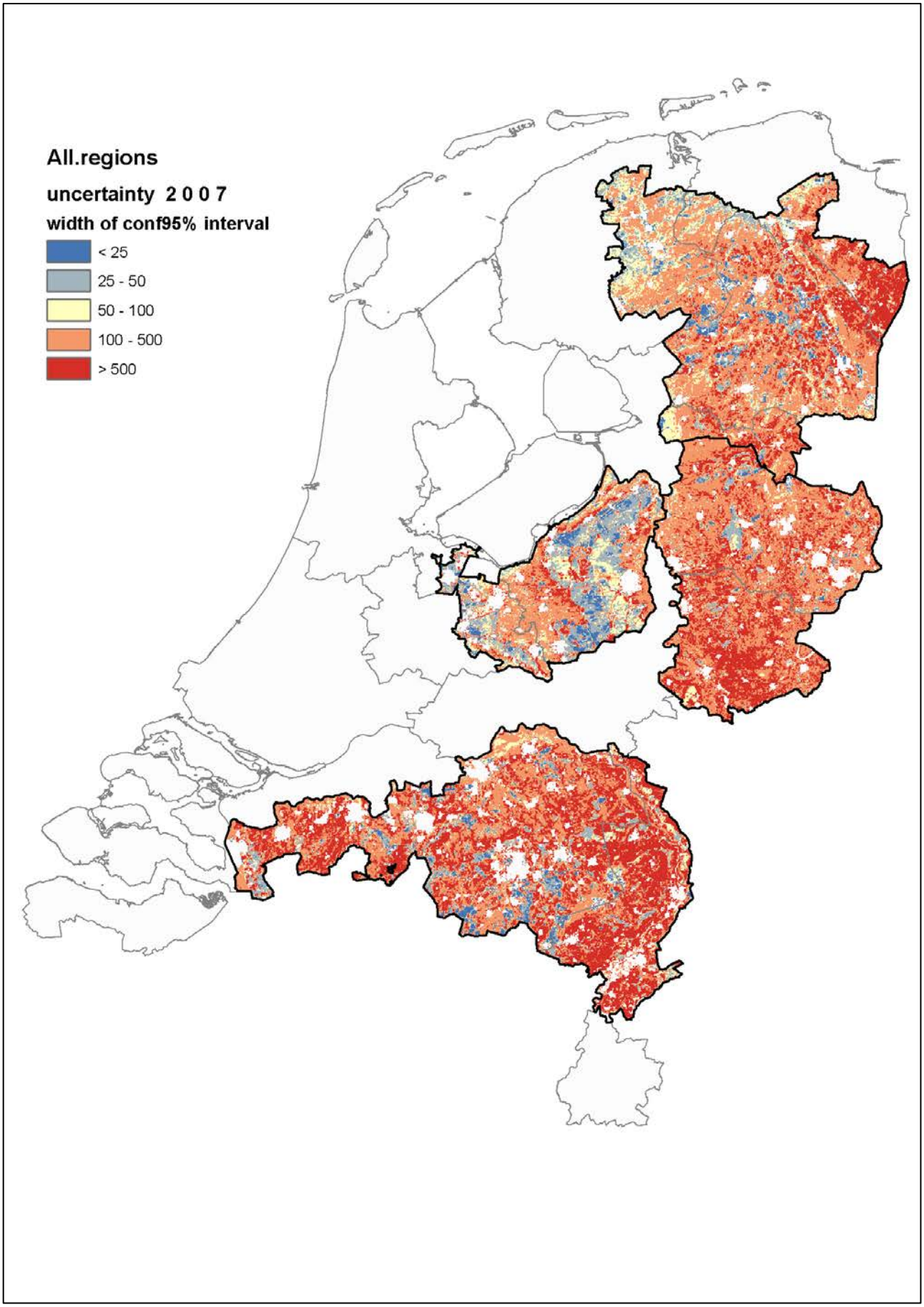
QQ-plot



Semivariogram short range

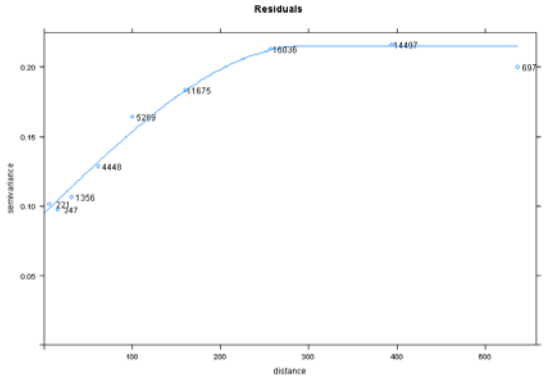



Nitrate levels 2 0 0 7
predicted NO3 (mg/l)
< 25
25 - 50
50 - 100
100 - 150
> 150

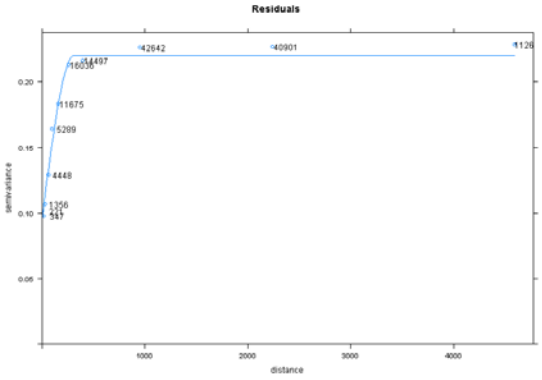

All.regions
uncertainty 2007
width of conf95% interval
< 25
25 - 50
50 - 100
100 - 500
> 500

Data all regions 2008

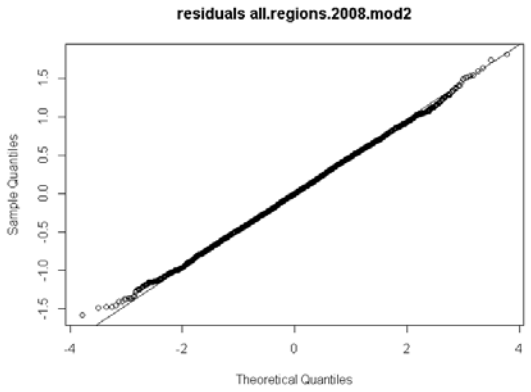
Semivariogram short range



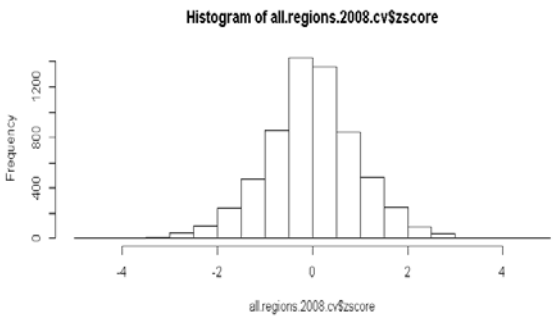
Long range

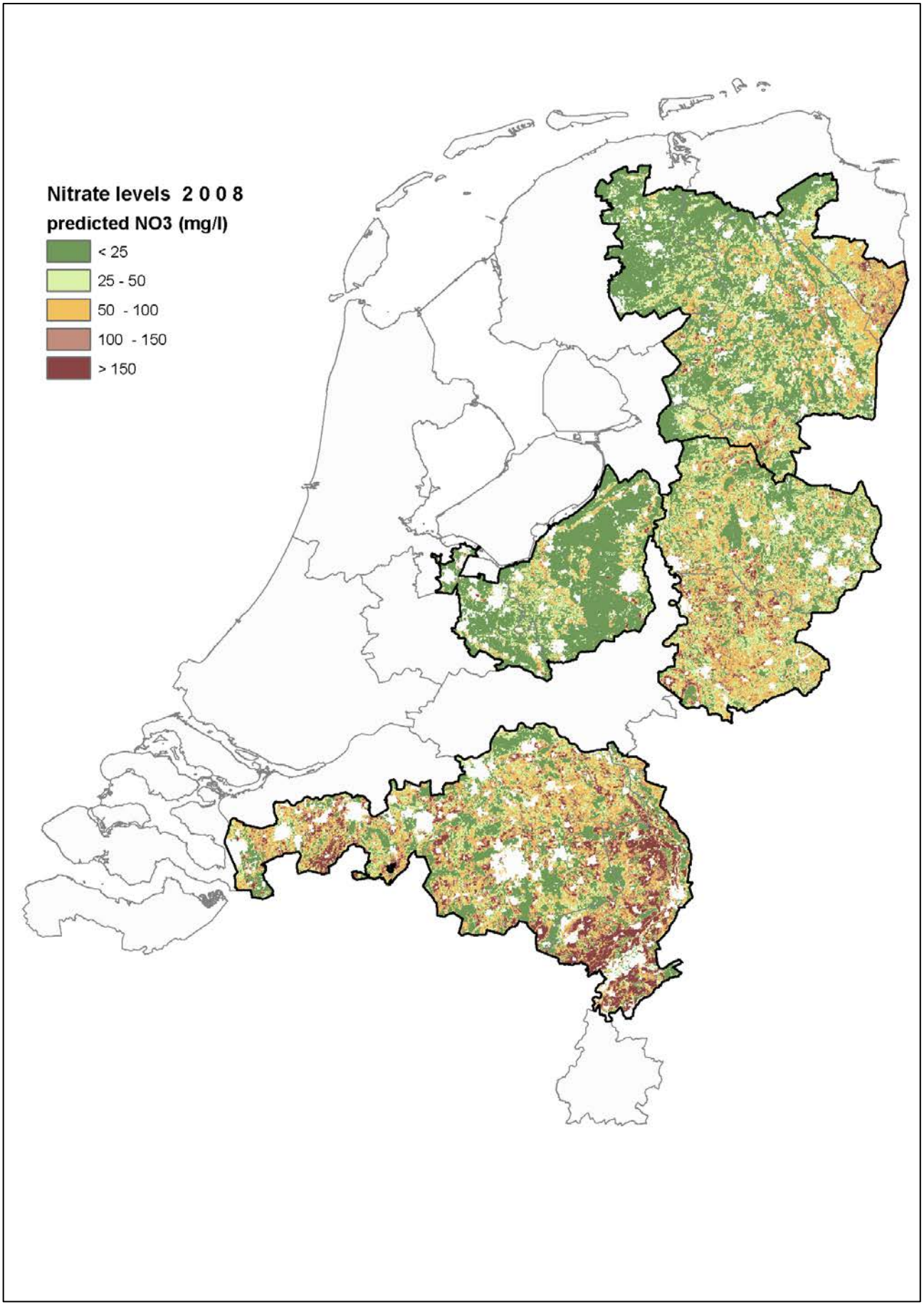


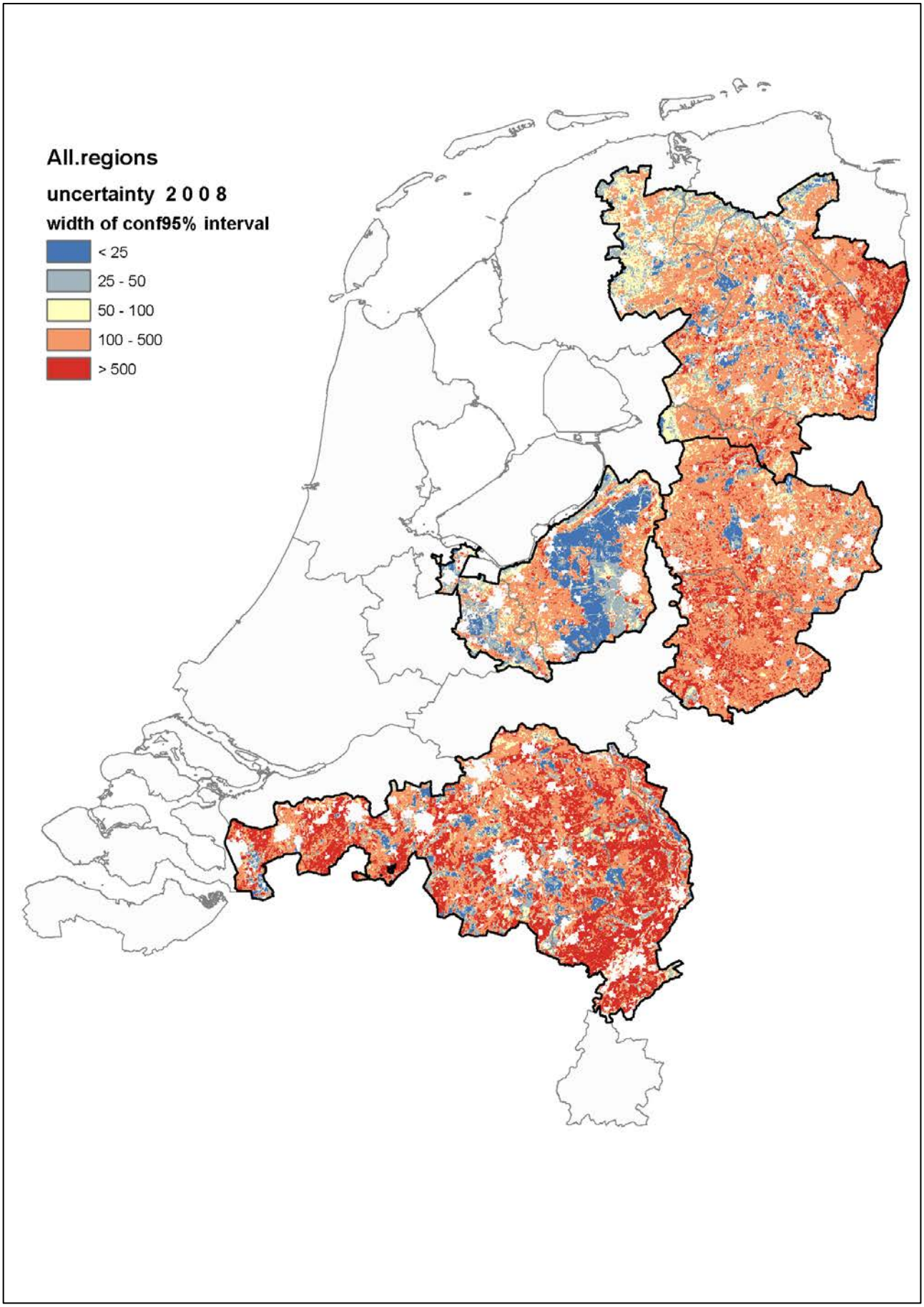
QQ-plot



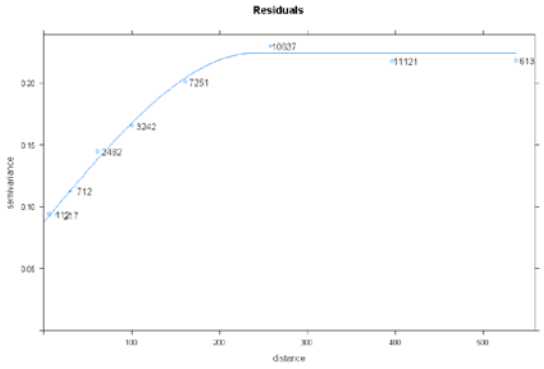
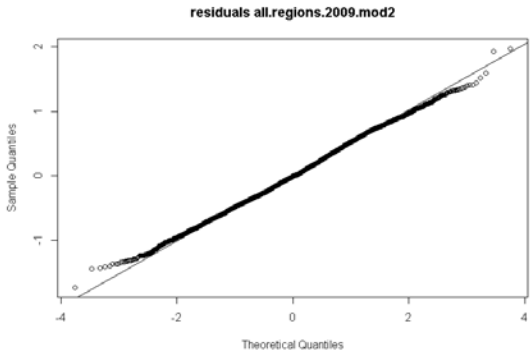
Histogram of crossvalidation-zscore

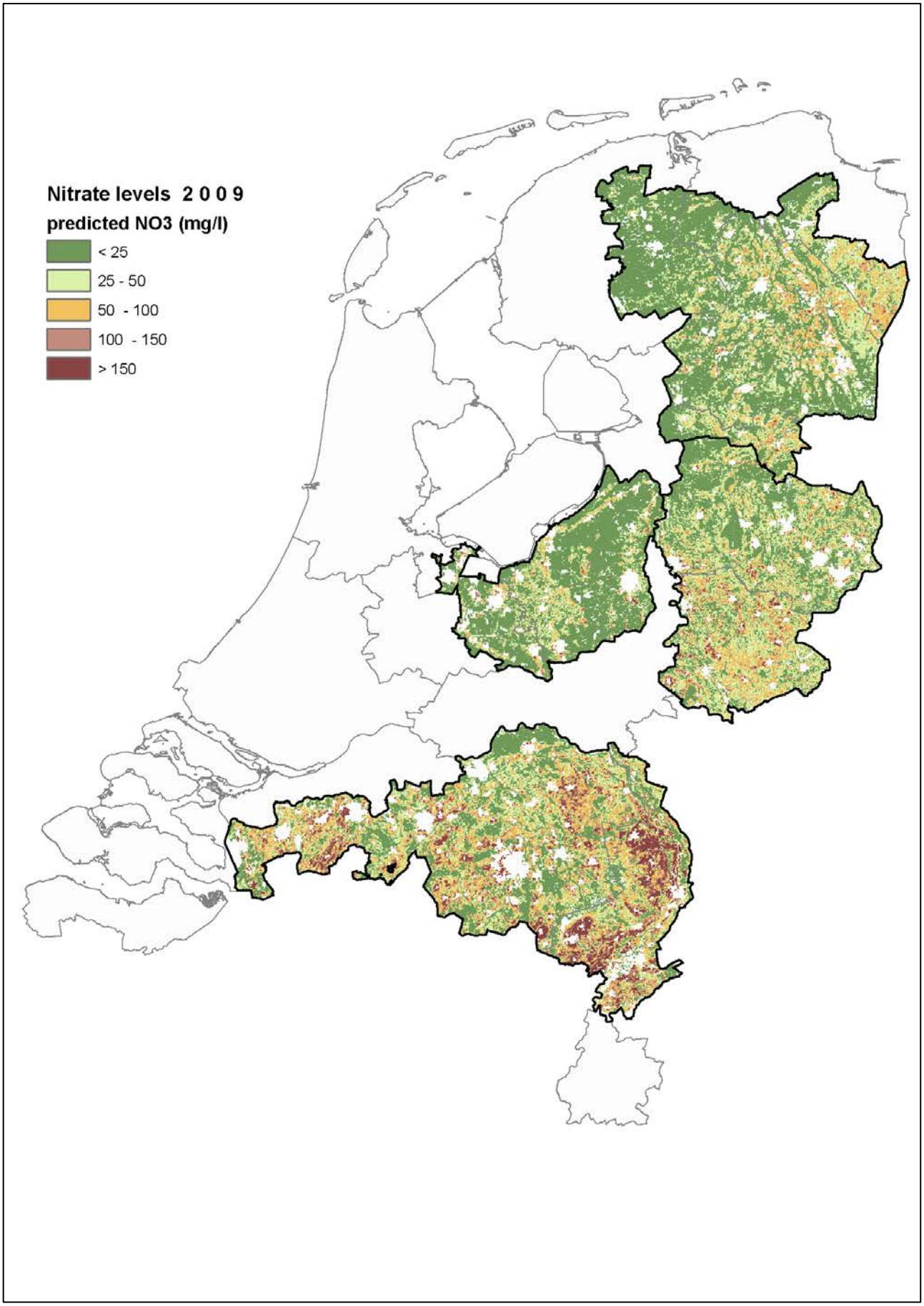


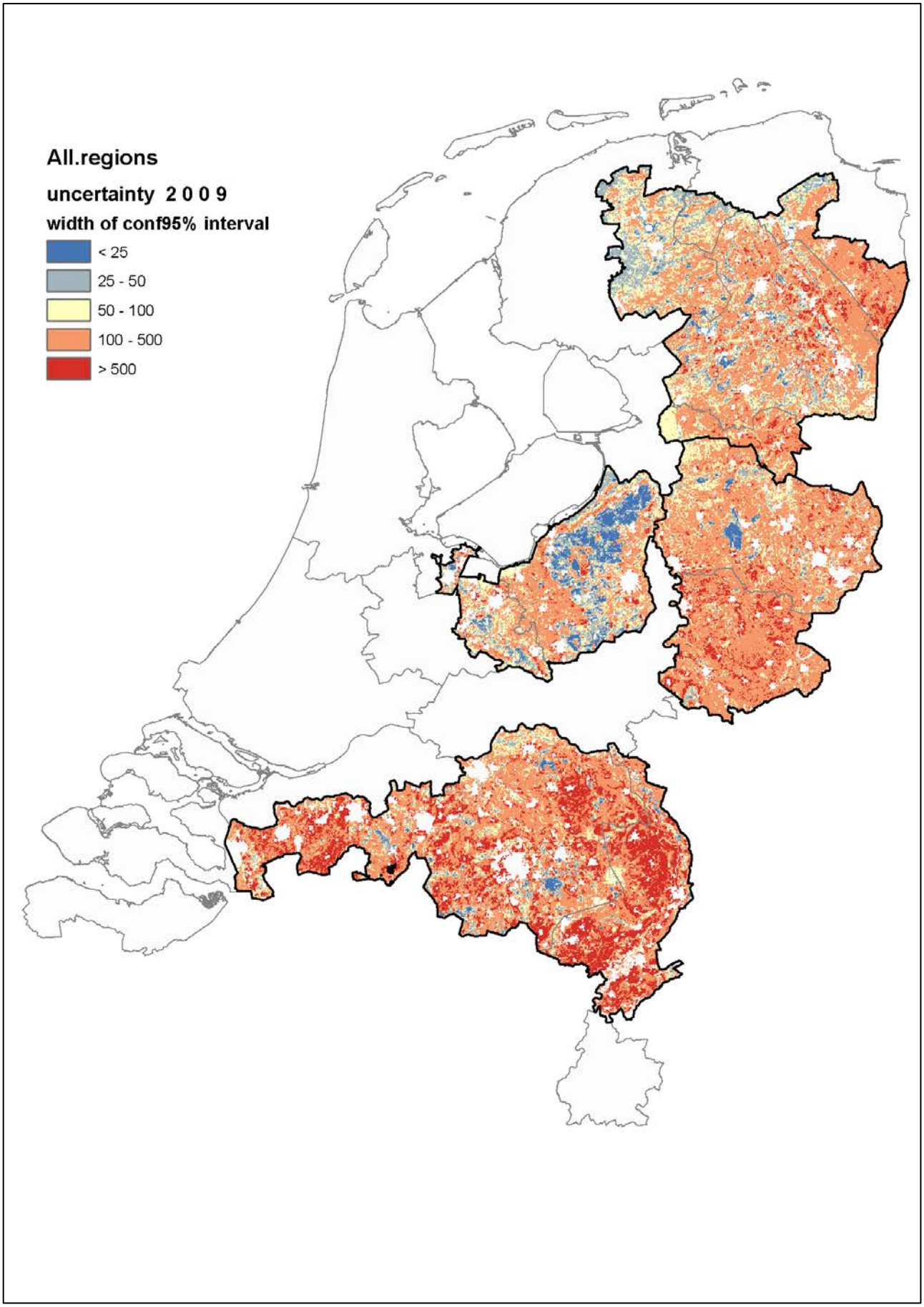

Nitrate levels 2 0 0 8
predicted NO3 (mg/l)
< 25
25 - 50
50 - 100
100 - 150
> 150


All.regions
uncertainty 2 0 0 8
width of conf95% interval
< 25
25 - 50
50 - 100
100 - 500
> 500

Data all regions 2009

Semivariogram	QQ-plot
	


Nitrate levels 2 0 0 9
predicted NO3 (mg/l)
< 25
25 - 50
50 - 100
100 - 150
> 150


All.regions
uncertainty 2009
width of conf95% interval
< 25
25 - 50
50 - 100
100 - 500
> 500

Appendix Va - Data description of covariates

In this appendix a short description of the covariate maps referred to in this report is listed. Since only selected parts of the Netherlands were used for prediction (only the sandy soils), the clipping method of generating input data may have omitted some categories in the variables. A distinction is made between categorical and continuous data. Grid codes correspond with those listed in the verbose model listings, found in Appendix VI. For instance, in the bbg06 covariate map listing, bbg0660 corresponds with grid code 60, “bos” or forest in English. Images of the covariate maps are available in Appendix Vb.

Categorical data (model parameter name, official map name)

1. bbg06, BBG06 (CBS)
GRID map, 25m resolution
Statistical interpretation of satellite imagery, aligned with TOP10vector map of the Netherlands. Aimed at urban areas, also with land use categories. In Dutch: Bestand Bodemgebruik 2006.

Table 19. Classification of BBG06 after (CBS, 2008).

Group	**Group name**	**Grid code**	**Description** (in Dutch)
1	Traffic area	10	spoorterrein
		11	wegverkeersterrein
		12	vliegveld
2	Urban area (dense)	20	woonterrein
		21	detailhandel/horeca
		22	openbare voorzieningen
		23	sociaal-culturele voorzieningen
		24	bedrijventerrein
3	Semi-urban area	30	stortplaats
		31	wrakkenopslagplaats
		32	begraafplaats
		33	delfstofwinplaats
		34	bouwterrein
		35	semiverhard overig
4	Recreational area	40	park en plantsoen
		41	sportterrein
		42	volkstuin
		43	dagrecreatief terrein
		44	verblijfsrecreatief terrein
5	Agricultural area	50	glastuinbouw
		51	overig agrarisch terrein
6	Forested area and Nature	60	bos
		61	open droog natuurlijk terrein
		62	open nat natuurlijk terrein
7	Fresh water	70	IJsselmeer/Markermeer
		71	afgesloten zeearm
		72	Rijn en Maas
		73	Randmeer
		74	spaarbekken
		75	recreatief binnenwater
		76	binnenwater delfstofwinning

		77	vloei- en/of slibveld
		78	overig binnenwater
8	Tidal water (marine)	80	Waddenzee, Eems, Dollard
		81	Oosterschelde
		82	Westerschelde
		83	Noordzee
9	Foreign (non-NL)	90	buitenland

2. geom, GKN (Alterra)
Vector map, converted to GRID map, 25m resolution
Geomorphology map (Geomorfologische Kaart Nederland) of the Netherlands at 1:50000 scale, made in 2003. This map is a simplification of a much more detailed version, having 429 possible subdivisions of 18 different categories. Reference can be found in (Koomen & Maas, 2004). The simplified geomorphology map has 25 units.

Table 20. grid codes for simplified geomorphology map

Grid Code	**Main class**	**Description (in Dutch)**
0	A	Wanden
1	B	(hoge) Geïsoleerde verhogingen
2	BEB	Bebouwde kom
3	C	Hoge heuvels, ruggen, welvingen
4	D	Plateaus
5	Db	
6	Dijk	
7	E	Terrasvormen
8	F	Plateau-achtige vormen
9	G	Waaiervormige glooiingen
10	H	Niet-waaiervormige glooiingen
11	Hw	
12	K	(lage) Geïsoleerde verhogingen
13	L	Lage heuvels, ruggen, welvingen
14	M	Vlakten
15	N	Niet-dalvormigen laagten
16	R	Ondiepe dalen
17	Recr	
18	S	Matig diepe dalen
19	T	Zeer diepe dalen
20	Terp	Terplichamen
21	Vv	
22	W	Water
23	Zee	Zee
24	Vib	

3. gronds, Simplified Soil map (Alterra)
Vector map, converted to GRID map, 25m resolution
Scale 1:50000, grouped soil type map, derived from (detailed) soil map (Alterra/Stiboka, 2006).

Table 21. grid codes for simplified soil map

Grid code	**Category** (Dutch)	**English**
10	Veen	Peat
20	Zand	Sand
21	Moerig op zand	Humic sand

30	Lichte zavel	Light sandy clay
40	Zware zavel	Heavy sandy clay
50	Lichte klei	Light clay
60	Zware klei	Heavy clay
70	Leem	Loam
98	Bebouwing	Built-up area
99	Water	Water

4. vds, Aggregated soil map (RIVM, based on (Alterra/Stiboka, 2006))
Vector map, converted to GRID map, 25 m resolution. Original scale 1:50000

Table 22. grid codes for aggregated soil map

Grid code	**Name** (Dutch)	**English**
0	Bebouwd	Built-up area
4	Löss	Loess
6	Moerig	Humic sand
7	Oude klei	Old (Tertiary) clay
8	Rivierklei	River clay
10	Veen	Peat
12	Water	Water
13	Zand	Sand
14	Zeeklei	Marine clay
Recoded as "other":		
1	Bebouwd/dijk	Dyke
2	Bovenland	No translation
3	Groeve	Mine/pit
5	Mijnstort	Mine debris dump site
9	Terp	Man-made hill
11	Vergraven	Disturbed soil

5. pawn, PAWN map (NHI)
Vector map, converted to GRID map, 25m resolution
PAWN is an acronym denoting 'Policy Analysis for the Water management of the Netherlands'. It refers to a clustering of soil units, originally at a scale 1:250 000 (Wösten, de Vries, Denneboom, & van Holst, 1998). The version in use for this research project is the translation for the 1:50 000 scale soil map (Alterra/Stiboka, 2006) in 2008 during the NHI-project (Nationaal Hydrologisch Instrumentarium) by commission of Deltares. Below is an abridged table of the units that were distinguished. Soil depth is limited to 1,20 meter. Not all codes are present in the selection of sandy soil regions of the Netherlands, used for this Regression Kriging study.

Table 23. Grid codes for the PAWN-classification. Abridged after (Wösten et al., 1998)

Grid code	**Description** (in Dutch)	**English**
1	Veraarde bovengrond op diep veen	Peaty earthy topsoil on peaty subsoil
2	Veraarde bovengrond op veen op zand	Peaty earthy topsoil on peat on sandy subsoil
3	Kleidek op veen	Clayey topsoil on peaty subsoil
4	Kleidek op veen op zand	Clayey topsoil on peat on sandy subsoil
5	Zanddek op veen op zand	Sandy topsoil on peat on sandy subsoil
6	Veen op ongerijpte klei	Peat on non-ripened clayey subsoil
7	Stuifzand	Dune sand
8	Leemarm zand	Loam-poor sand

9	Zwaklemig fijn zand	Slightly loamy sand
10	Zwaklemig fijn zand op grof zand	Slightly loamy sand on coarse sandy subsoil
11	Sterk lemig fijn zand op (kei-)leem	Loamy sand on boulder clay
12	Enkeerdgronden (fijn zand)	Sandy man-made thick earth soils
13	Sterk lemig zand	Loamy sand
14	Grof zand	Coarse sand
15	Zavel met homogeen profiel	Sandy (clay) loam
16	Lichte klei met homogeen profiel	Clay (loam)
17	Klei met zware tussenlaag of ondergrond	Clay (loam) on heavy clayey subsoil
18	Klei op veen	Clay (loam) on peaty subsoil
19	Klei op zand	Clay loam on fine sandy subsoil
20	Klei op grof zand	Clay loam on coarse sandy subsoil
21	Leem	Loam
22	Water	Water and marshy land
23	Stedelijk gebied	Built-up area

6. gt06, GT2006 (Alterra)

SWAP Model output GRID map, 25m resolution

Watertable depth class grid map. Based upon soil map (4) 1:50000, but improved by (van der Gaast, Massop, Vroon, & Staritsky, 2006). GHG stands for the (average) upper limit where the groundwater table is situated in centimeters below the field surface level. GLG means the (average) deepest limit. Updates from field scale research have improved the original map. 'Opinio generalis' was that water tables are lower since the first maps were made in 1960s to 1980s. Accuracy is described in (van der Gaast et al., 2006).

Table 24. Groundwater table classification with upper and lower limits of water levels.

Grid code	Groundwater table code	GHG (upper limit) -cm	GLG (lower limit) -cm
0	-	Not defined	Not defined
1	I	-	< 50
2	II	-	50 – 80
3	IIb	25 - 40	50 – 80
4	III	< 40	80 – 120
5	IIIb	25 – 40	80 – 120
6	IV	> 40	80 – 120
7	V	< 40	> 120
8	Vb	25 – 40	> 120
9	VI	40 – 80	> 120
10	VII	80 – 140	> 120
11	VIII	> 140	> 120

7. lgn6, LGN6 (Alterra)
GRID map, 25m resolution
Land use grid map in 39 classes, based on satellite imagery from 2007 and 2008. Geometry and thematic division were harmonized with BBG03 and Top10-vector (2006). Small changes in classification with LGN5 (20,21,22,44 and 46). Aggregation files for main classes were used. Crop accuracy was estimated at 84.5%. References can be found in (Hazeu, Schuiling, Dorland, Oldengarm, & Gijsbertse, 2010).

Table 25. LGN6 classification table

Grid code	Group name	Sub group	Class + description
1	Agricultural area		Agricultural grassland
2			Maize
3			Potatoes
4			Sugar beet / Field beet
5			Cereals
6			Other crops
61			Tree nursery
62			Fruit orchard
8			Greenhouse horticulture
9			Orchard (other)
10			Bulb growing
26			Built-up countryside
11	Forest		Deciduous forest
12			Coniferous forest
16	Water		Fresh water
17			Saltwater
18	Built-up area		Buildings in primary built-up area
19			Buildings in secondary built-up area
20			Forest within primary built-up area
22			Forest within secondary built-up area
23			Grassland within primary built-up area
24			Fallow within rural built-up area
28			Grassland within secondary built-up area
25	Infrastructure		Main roads & railroads
30	Natural area	Coastal	Salt marshes
31			Open sand in coastal areas
32			Dunes, low vegetation (<1m)
33			Dunes, high vegetation (>1m)
34			Dunes covered with heather
35		Heathland	Shifting sands or river sand
36			Moor / Heather
37			Moderately grassy moorland
38			Highly grassy moorland
39		Peat moor	Peat moor
40			Forest on peat moor
41		Swamp	Other swamp vegetation
42			Reedy land / canebrake
43			Forest in swamp area
45			Natural grassland

8. various, Hydrological parameters (Alterra)

Various GRID maps, 25m resolution

In Alterra project 1339 "Hydrologie op basis van karteerbare kenmerken" (van der Gaast et al., 2006), several grid maps for hydrological properties have been constructed, including:

1. Length of ditches per grid cell:
   `laf` (summed length of ditches, superficial discharge) 0-46 m/grid cell
   `lont` (summed length of ditches, profound discharge) 0-46 m/grid cell

2. Distance to nearest discharge medium:
   `slaf` (distance to discharge by ditch, superficial) 0-10000m
   `slont` (distance to discharge by ditch, profound) 0-10000m

3. Resistance maps: amount of time before water reaches the nearest water discharge unit
   `draf` (drainage resistance for superficial discharge) <50d – 25000d
   `dront` (drainage resistance for profound drainage) <50d – 25000d

Numbers correspond with numbers in Figure 27. Ranges of units are indicated.

The maps have nine different classes, ranging from close range or small (0 or 1) to far/highest (9). The grids have either a proximity meaning, for instance the distance to the nearest ditch or the nearest tube drainage system, or a density meaning. Superficial discharge is not the same as run-off, which is very fast compared to drainage.

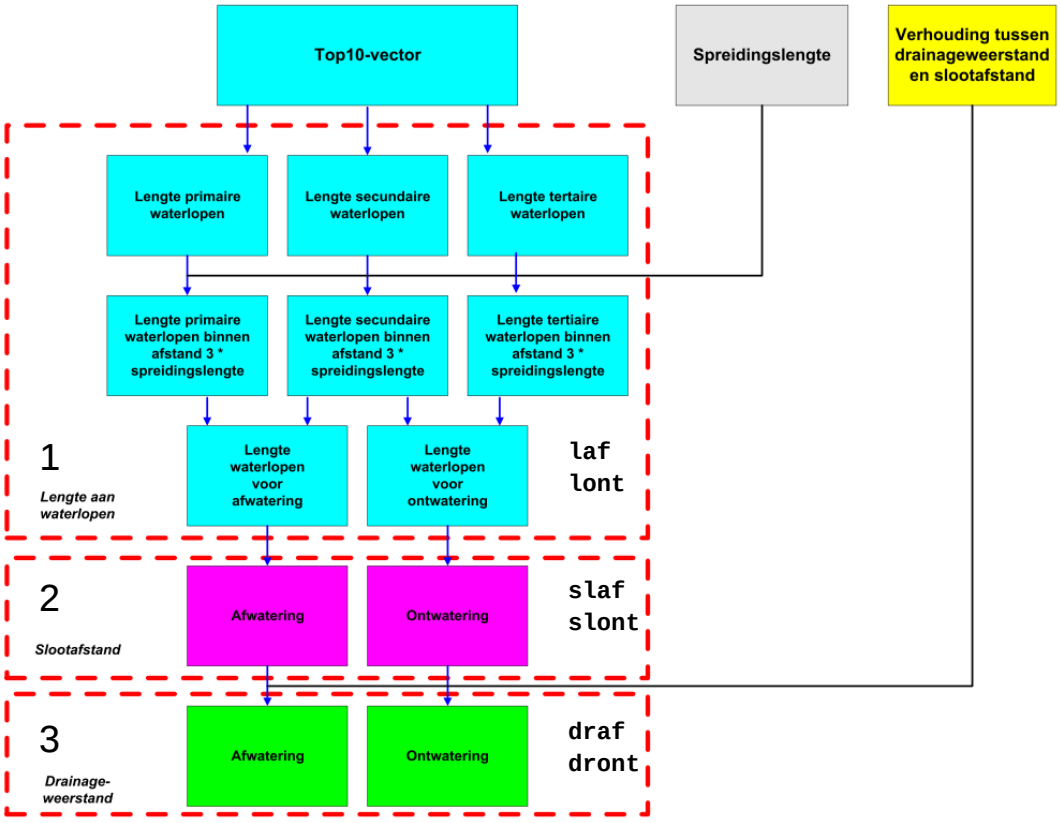


Figure 27. Scheme (modified) from the Alterra modellers for deducing the hydrological properties maps. Taken from Alterra report 1139, p41.

laf = lengte waterlopen voor afwatering
slaf = slootafstand afwatering
draf = drainageweerstand afwatering

lont = lengte waterlopen voor ontwatering
slont = slootafstand ontwatering
dront = drainageweerstand ontwatering

Continuous data

9. ahn, AHN-1 digital elevation model (RWS)
GRID map, 25m resolution
Digital elevation Model for the Netherlands, by LIDAR-measurements, originally at 5 meter resolution. Values in cm-1, ranging from -1230 to 33913 cm. In the selected regions, these values might be in a slightly moderate range.

10. kwel2, Seepage and infiltration map (Alterra)
GRID map, 25m resolution
Modeled results using the tool 'Hydromap' from the Alterra project 1339, "Hydrologie op basis van karteerbare kenmerken" (van der Gaast et al., 2006). Grid map giving vertical velocity of groundwater displacement in mm/day. This can have positive (upward or seepage) or negative (downward or infiltration) direction.) Accuracy unknown. *(Note; the dutch abbreviation indicates only upward transport, this was chosen poorly)*

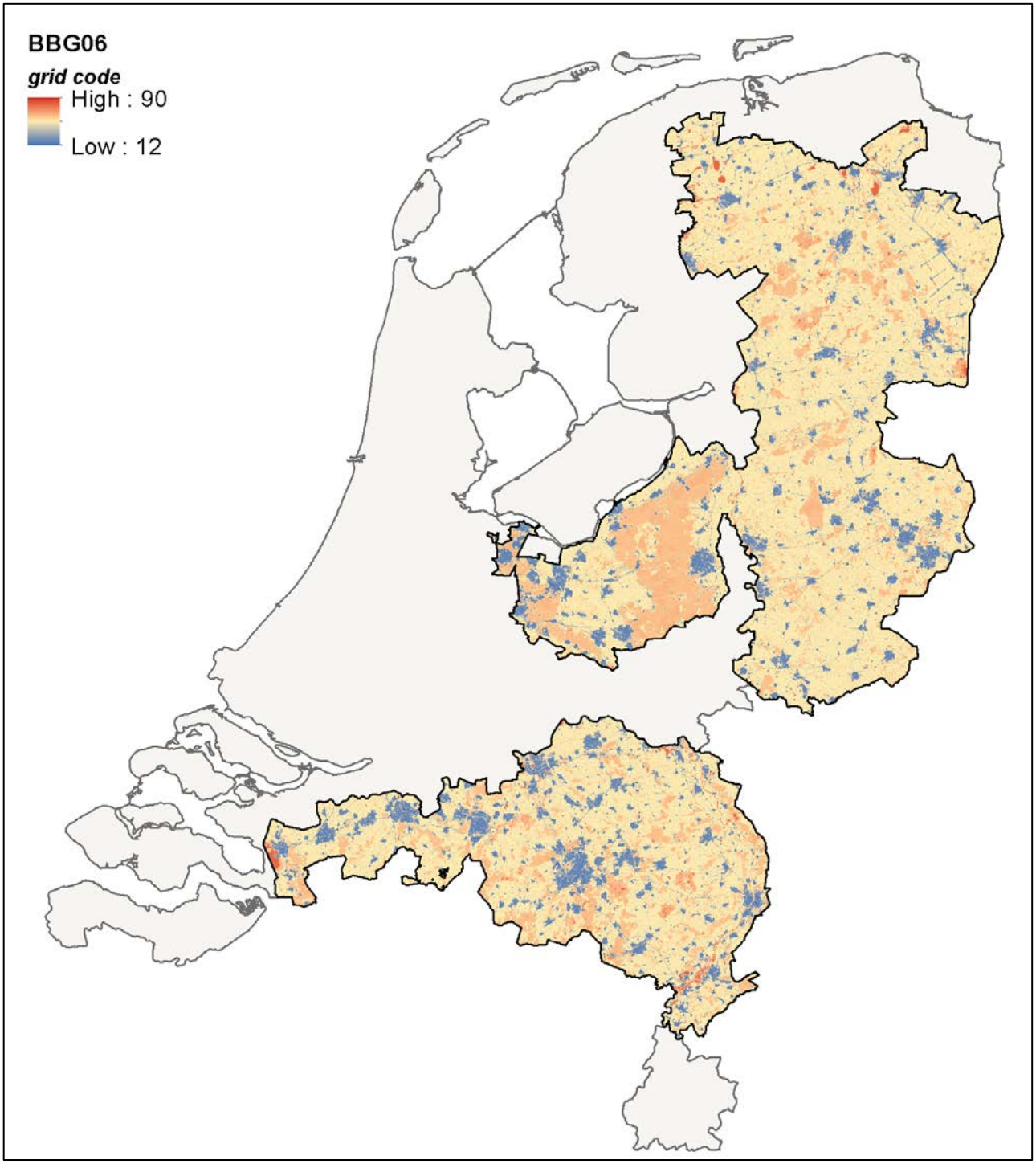
11. nhx, NHx-deposition in 2010 (RIVM/PBL)
GRID map, 1000m resolution
Model outcome for NHx (reduced nitrogen compounds) in mole N/ha in 2010 from the OPS dispersion model[6]. This model calculates average concentrations of substances in the atmosphere and deposition from there based on registered emission sources in Europe. The other years 2007, 2008 and 2009 are linear derivations via a scale parameter, yielding exactly the same patterns and therefore these were not considered for input separately. Accuracy unknown.

12. omxx, Organic matter maps (Alterra)
GRID maps, 25m resolution
Set of maps, depicting organic matter content in mass % at eight soil depths (5-10-25-40-60-80-100-120cm). Data from the construction of maps of the physical-chemistry conditions of Dutch soils (de Vries, 1999), updated with newer data from BIS[7]. Accuracy unknown.

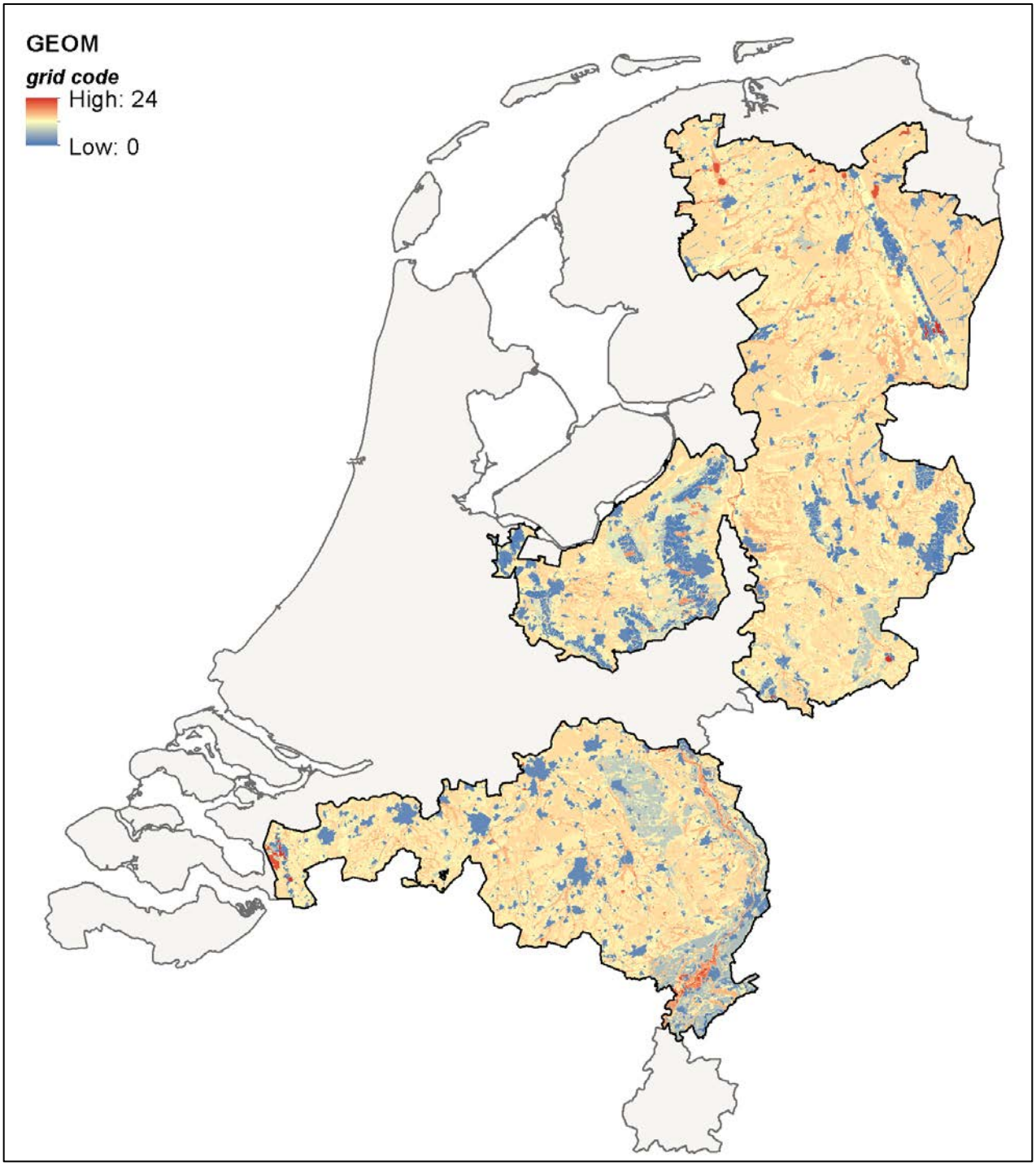
13. stone5 .. stone8, Manure and fertilizer addition map for 2005, 2006, 2007, 2008 (Alterra/RIVM)
GRID map, 250m resolution in STONE-plot format
(kg Nitrogen per grid cell per annum). Data from the MAMBO[8] model (link), adding up artificial fertilizer N and animal manure N for each of the 6505 STONE-plots[9]. These STONE-model plots are a combination of land use, soil type and hydrological conditions. Each plot is considered as unique and homogeneous. The actual attribute data was derived by WUR-LEI (Agricultural Economics Institute) and CBS (Netherlands Statistics) from their census data sources. Aggregated data to prevent privacy issues. *(Note: this covariabele should have been named like "Nman" or "Nfert", but since this format was prepared for STONE-grid cells, that name was given to it.)*

[6] OPS is an acronym for "Operationele Prioritaire Stoffen" model, see http://www.rivm.nl/ops
[7] BIS is an acronym for "Bodemkundig Informatie Systeem", a soil database by WUR/Alterra
[8] MAMBO is an acronym for "Mest en Ammoniak Model Beleids Ondersteuning", (Kruseman, G. et al, 2011)
[9] STONE is an acronym for "Samen Te Ontwikkelen Nutriënten Emissiemodel", (Wolf, J. et al, 2003)

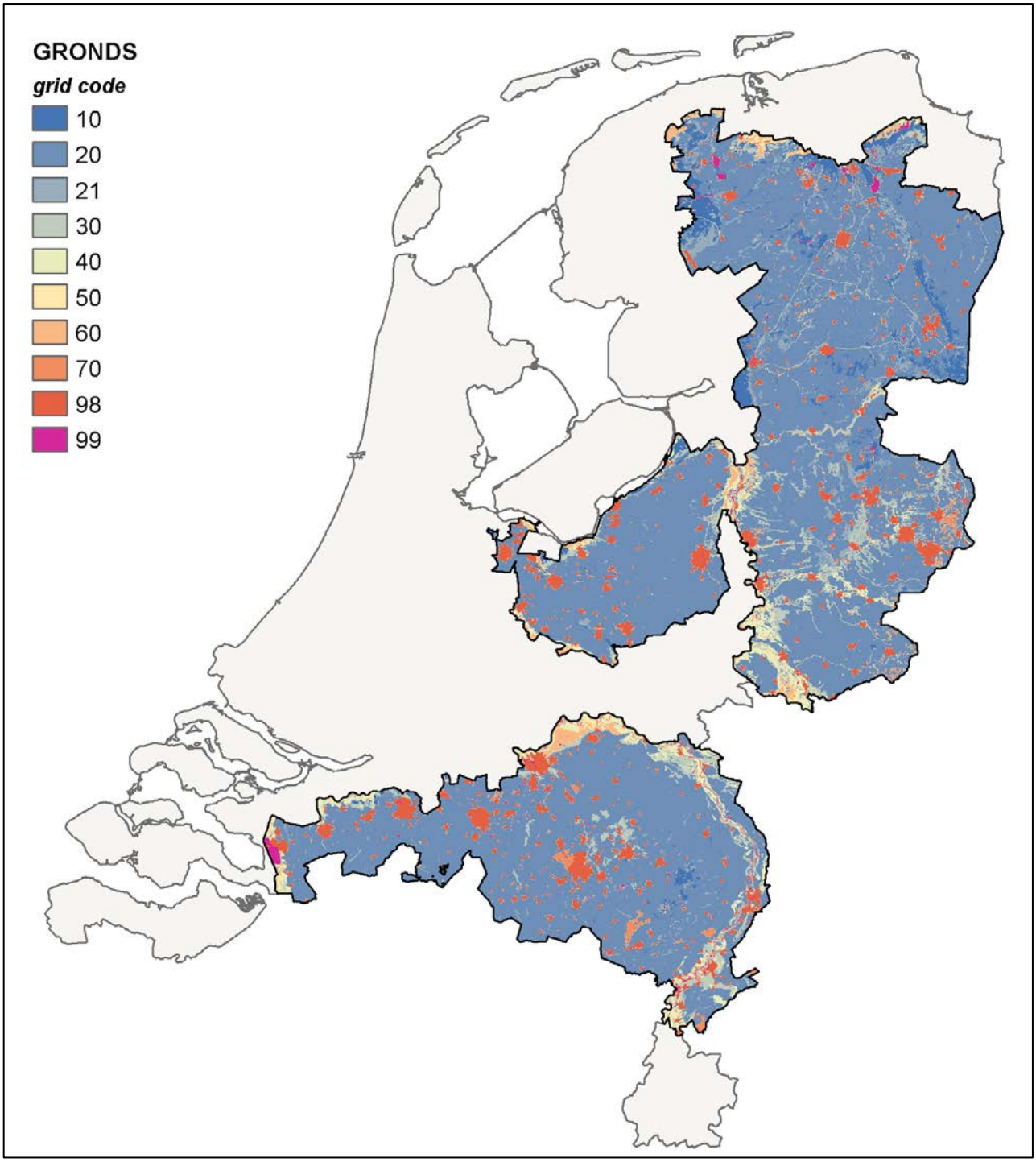
Appendix Vb - Data description of covariates - images



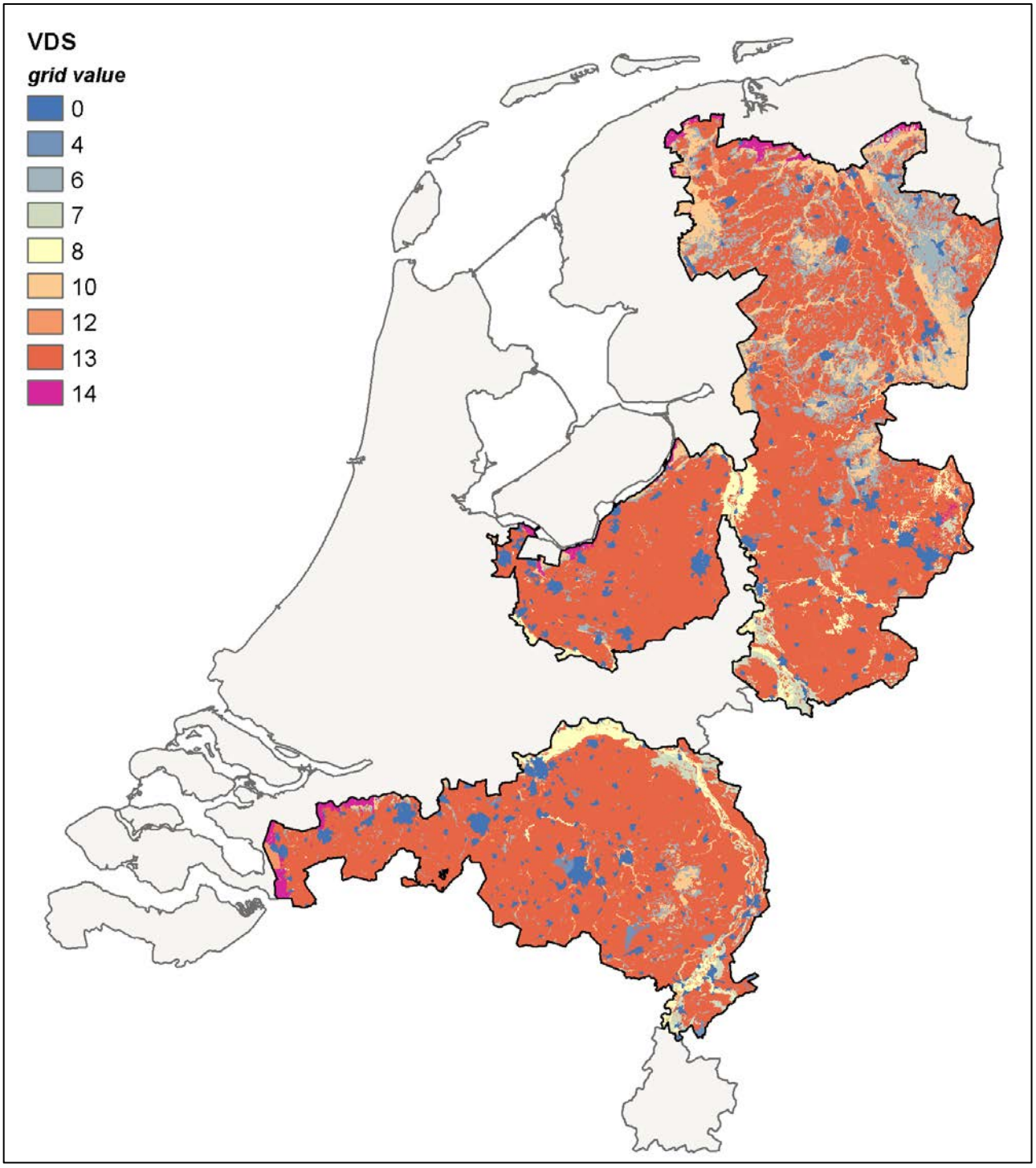
1. BBG06



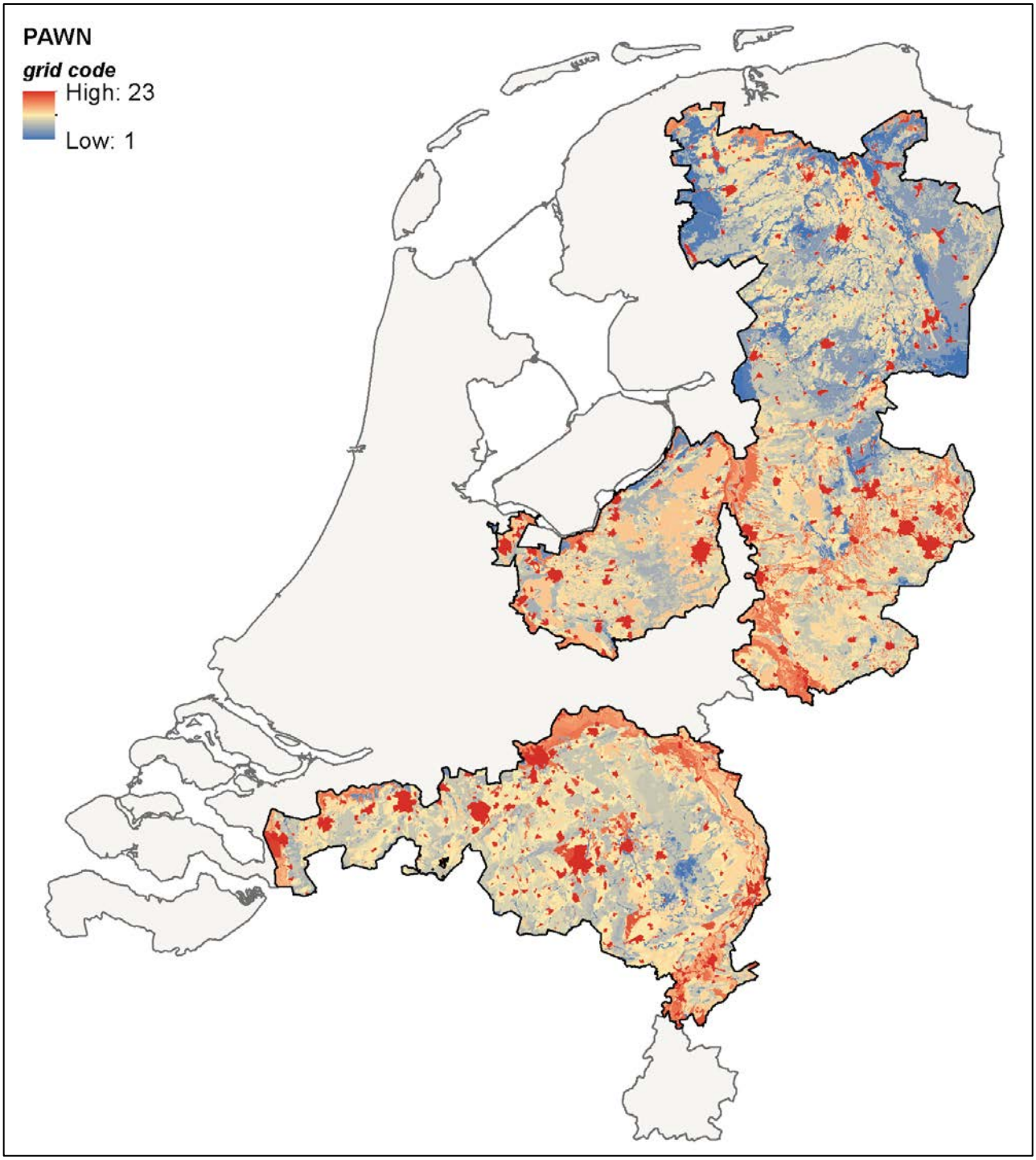
2. geom



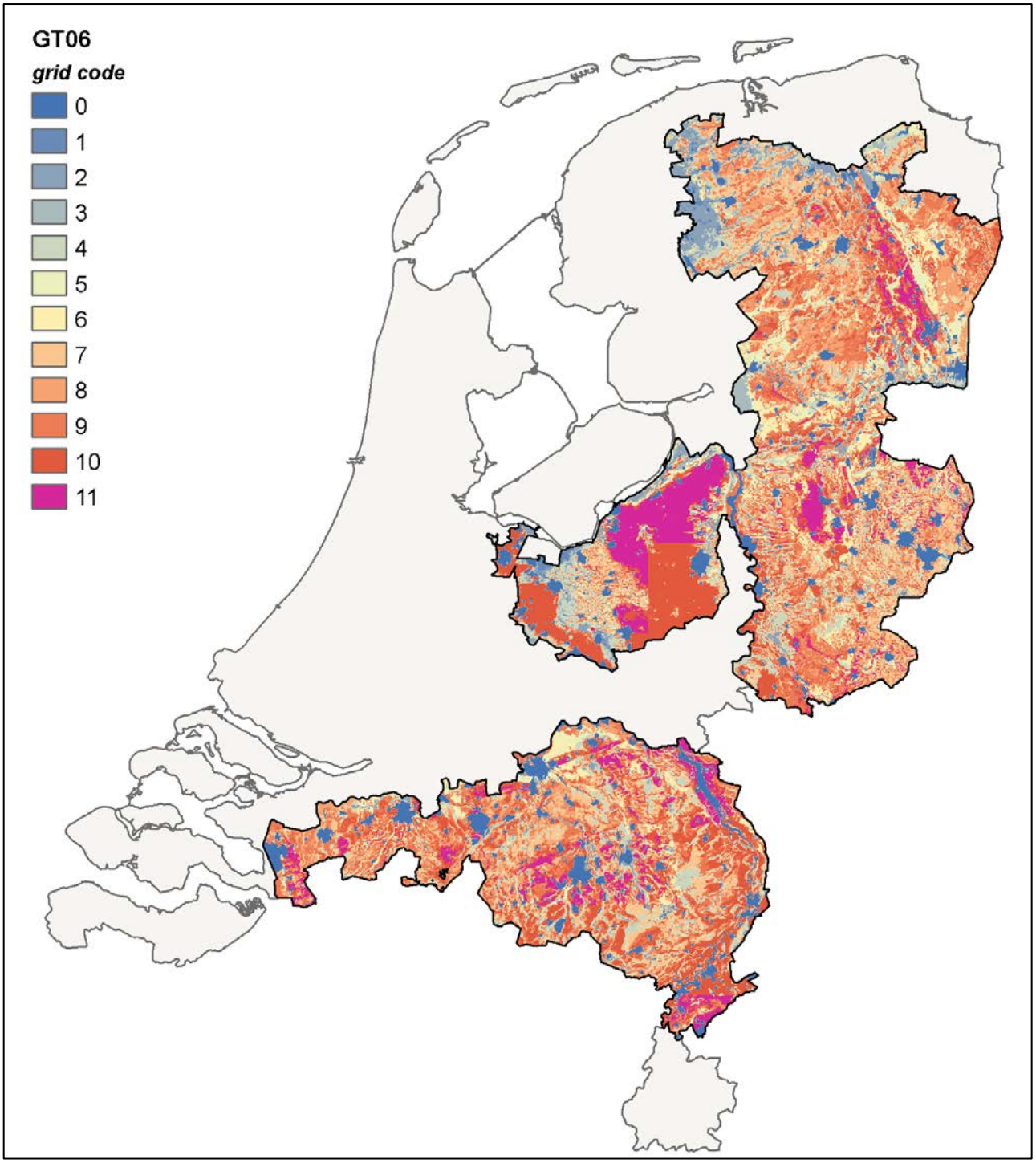
3. gronds



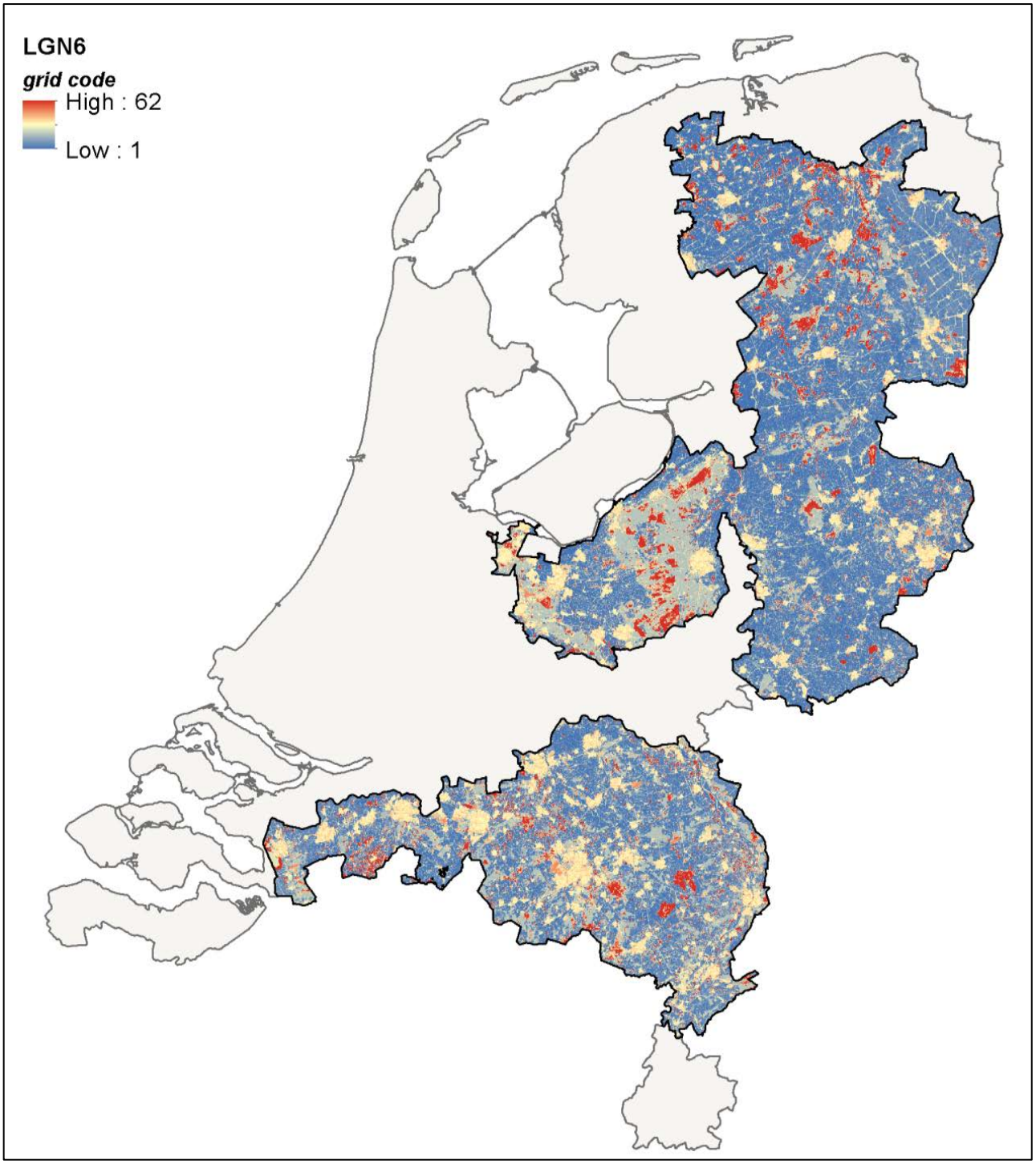
4. vds



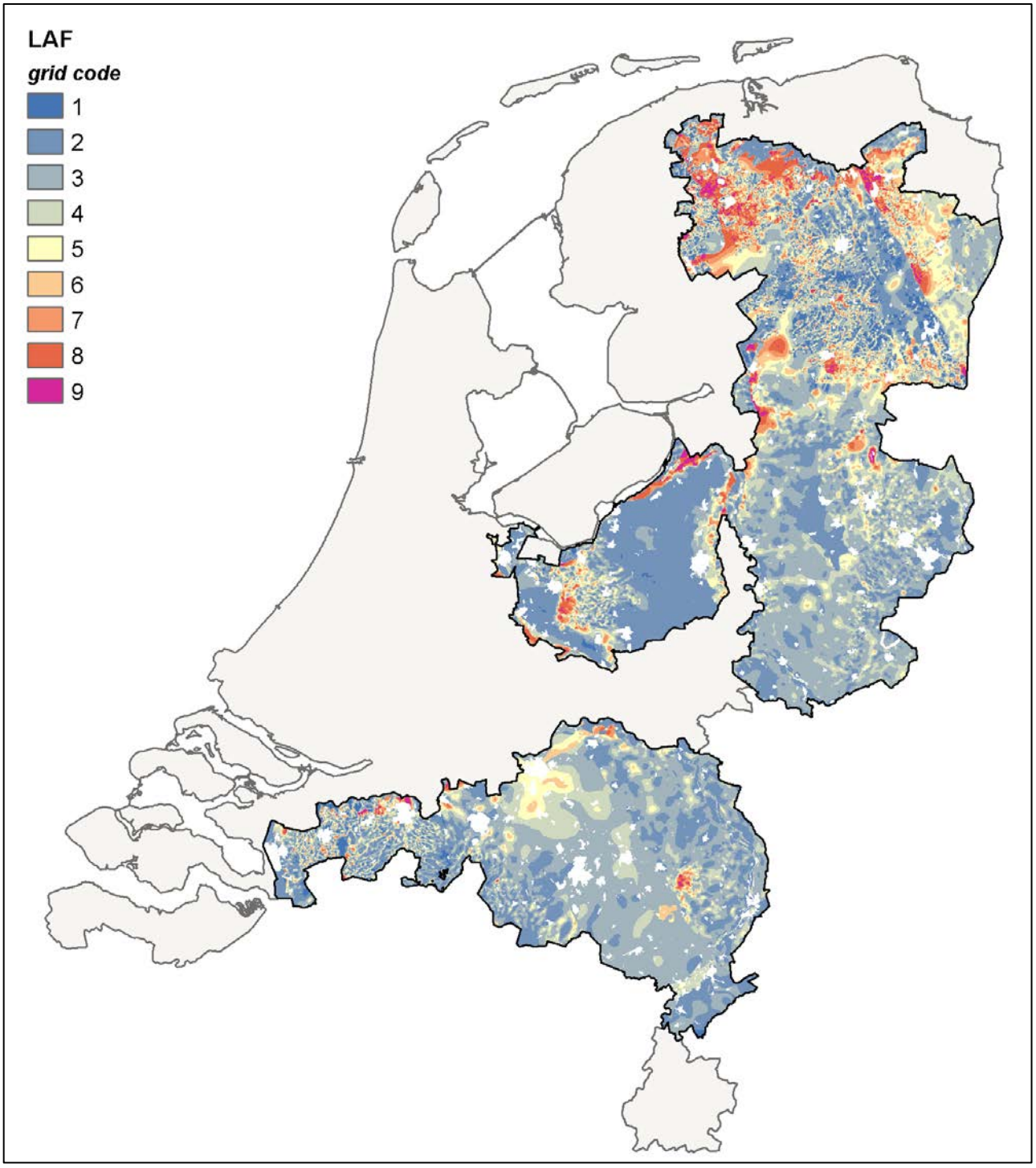
5. pawn



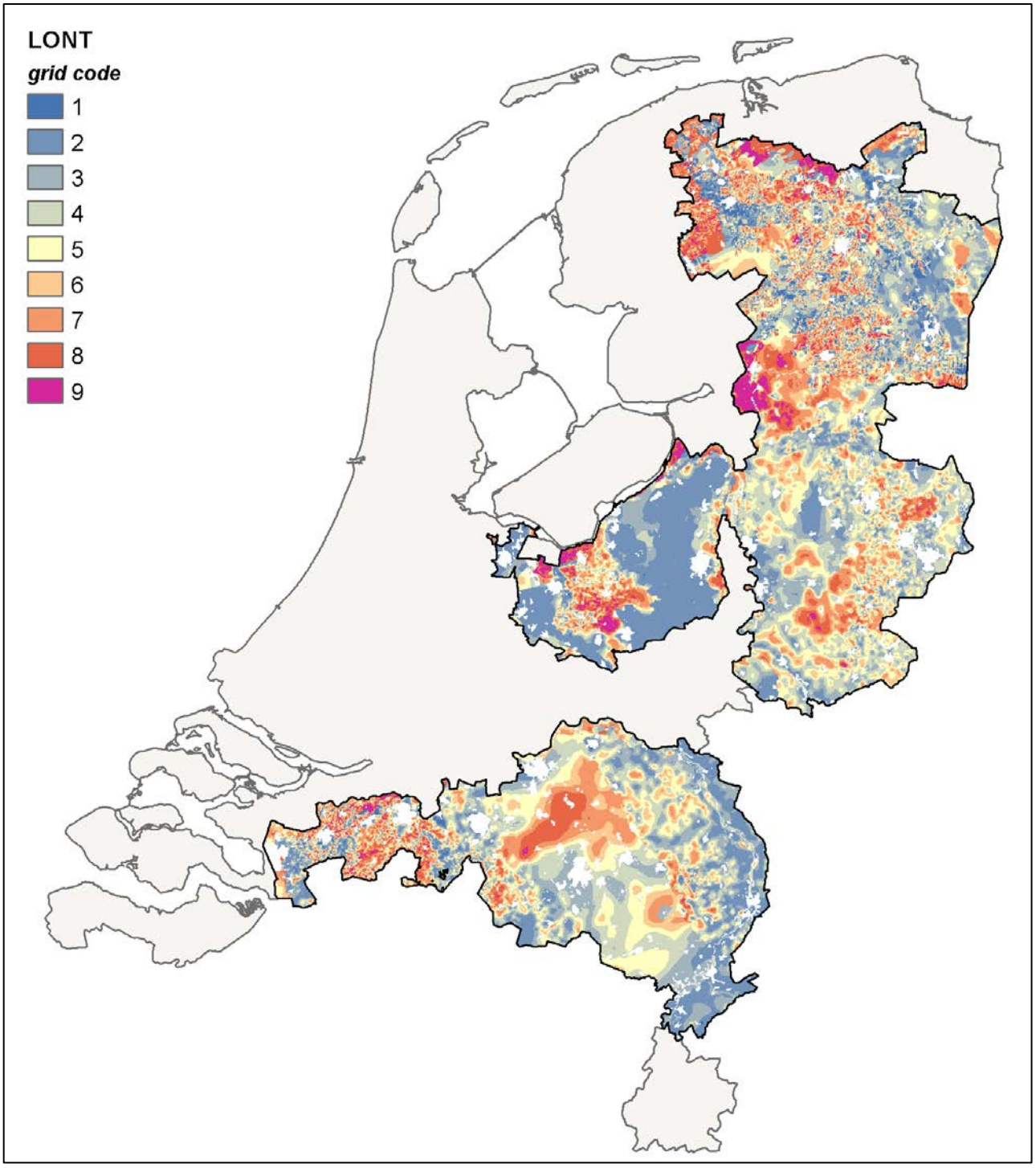
6. gt06



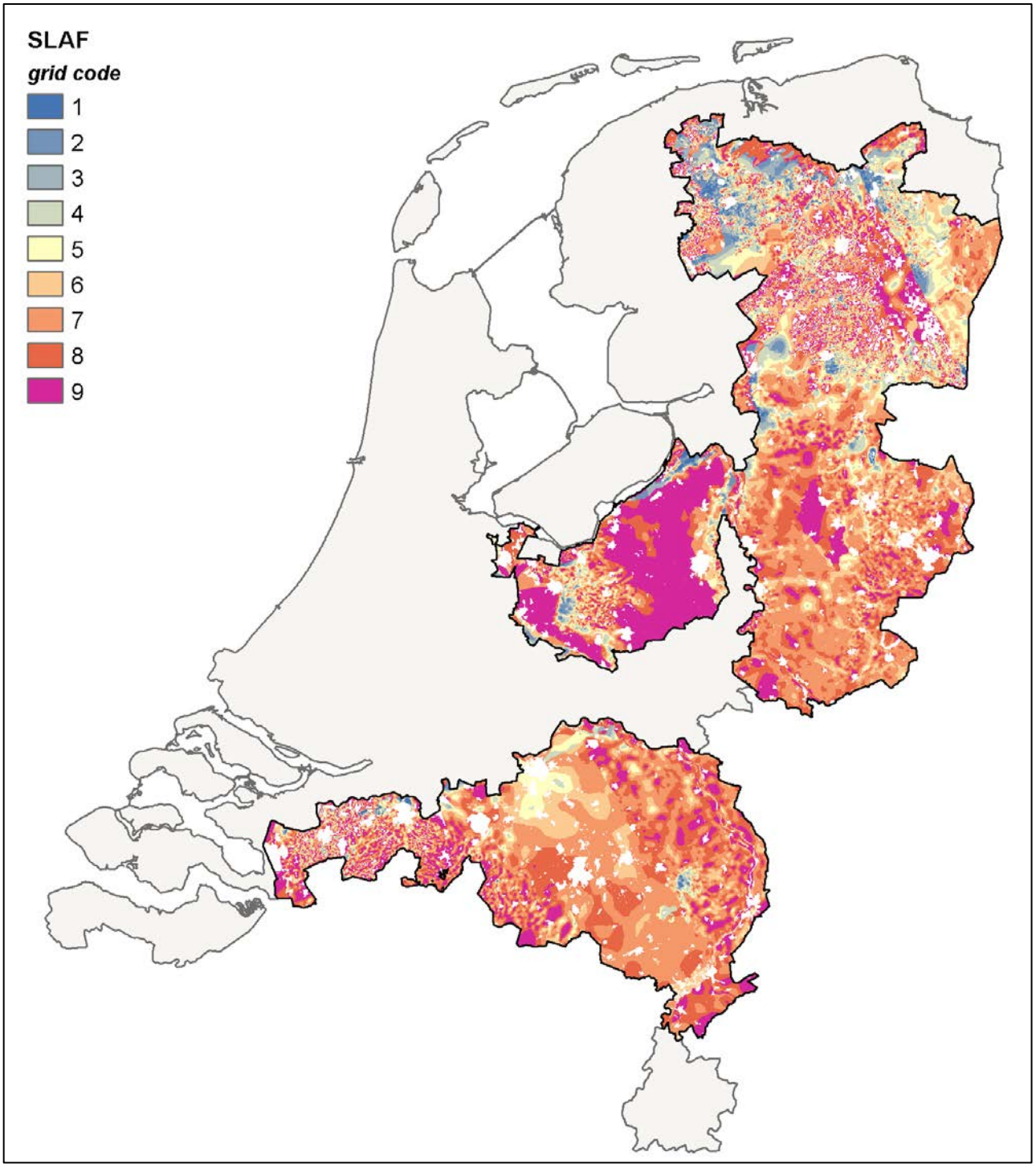
7. lgn6



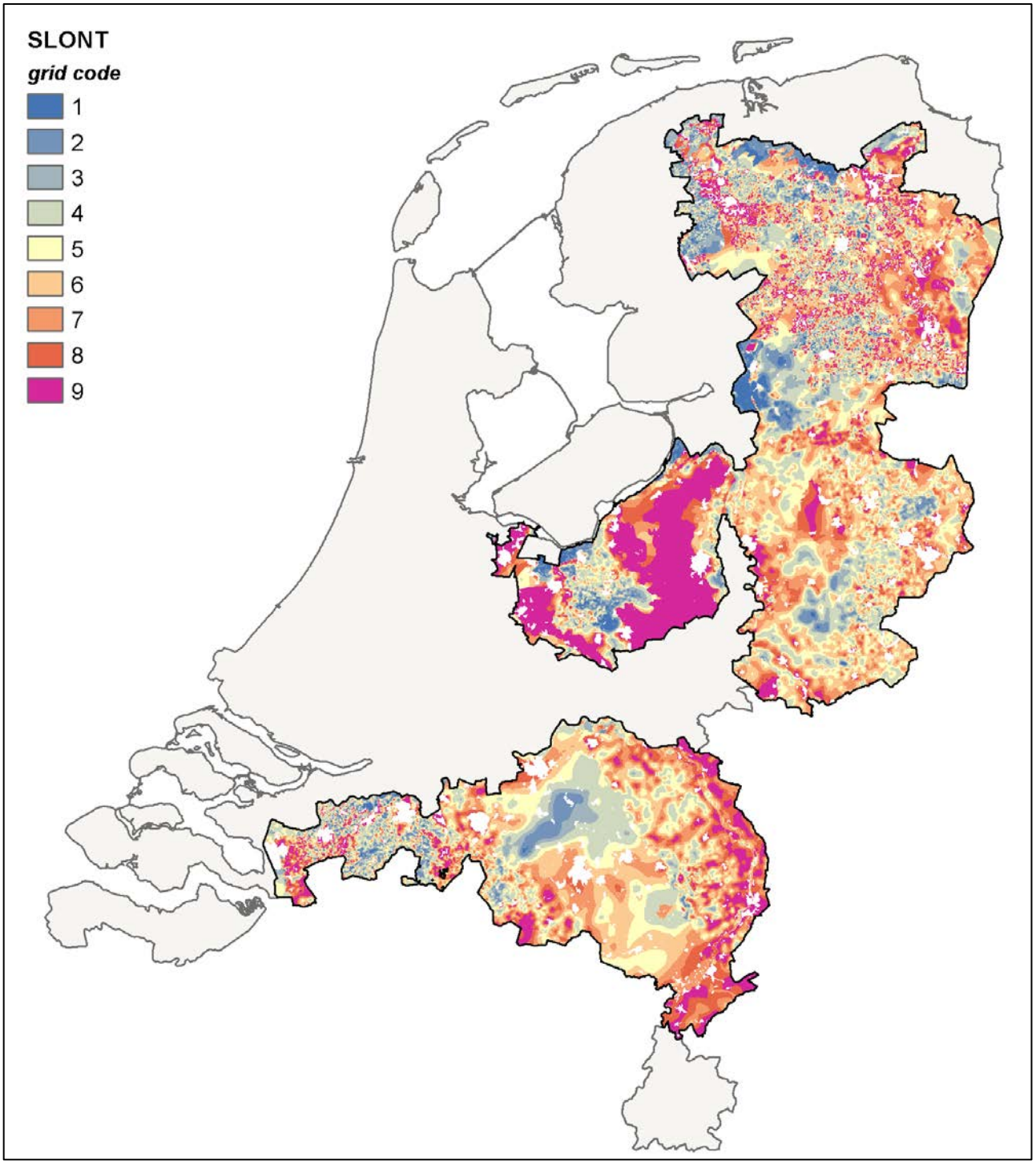
8a. laf



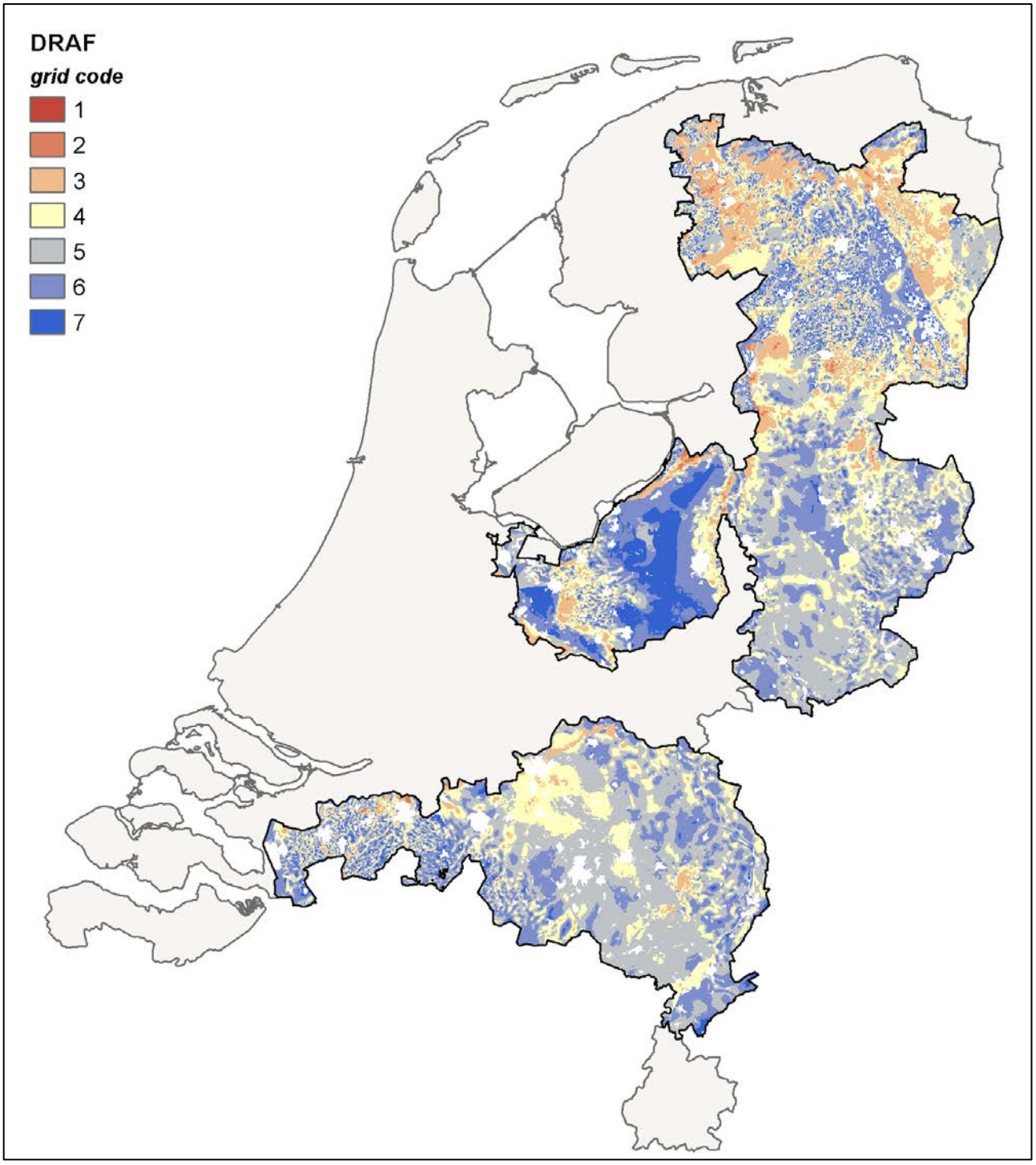
8b. lont



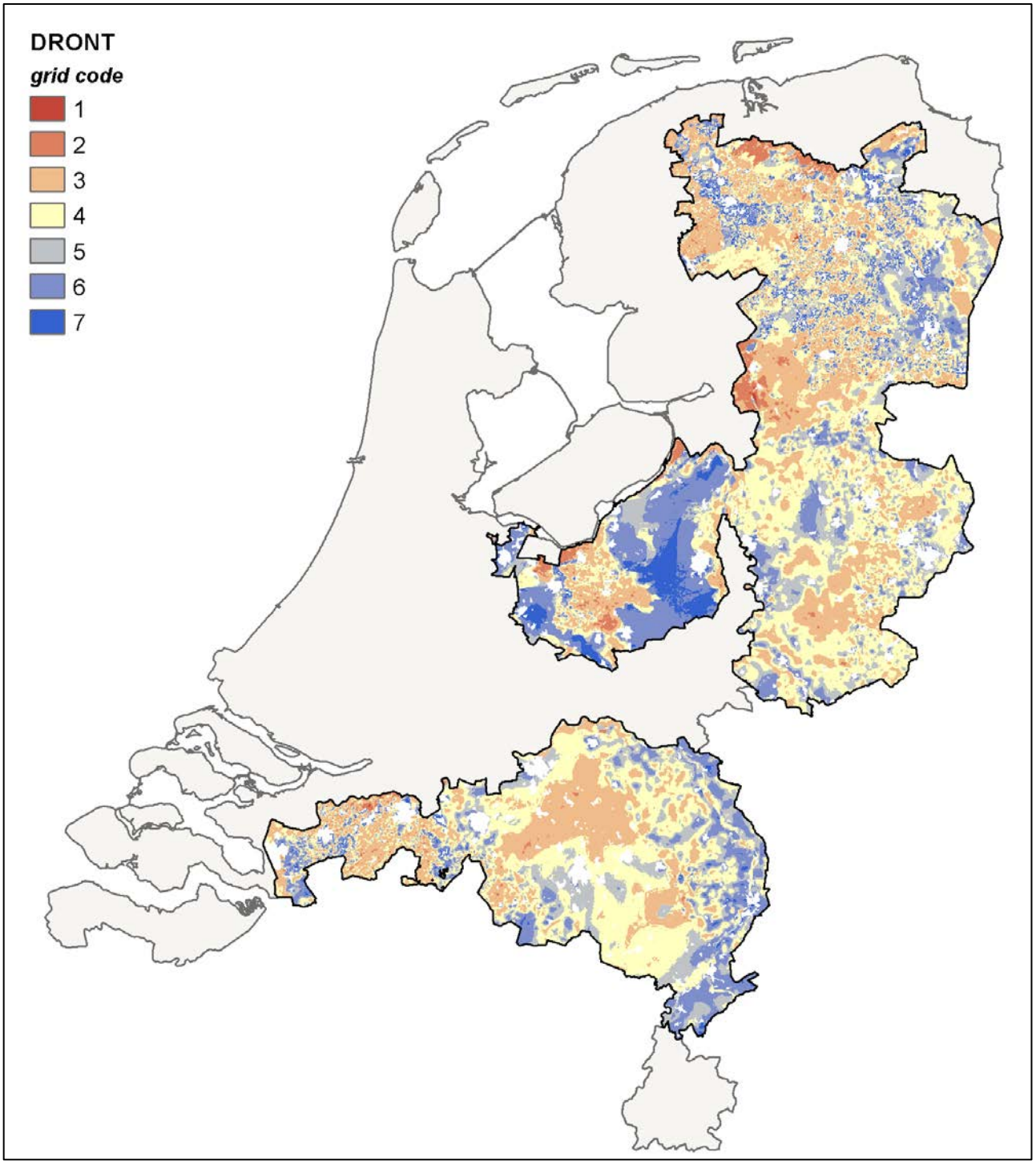
8c. slaf



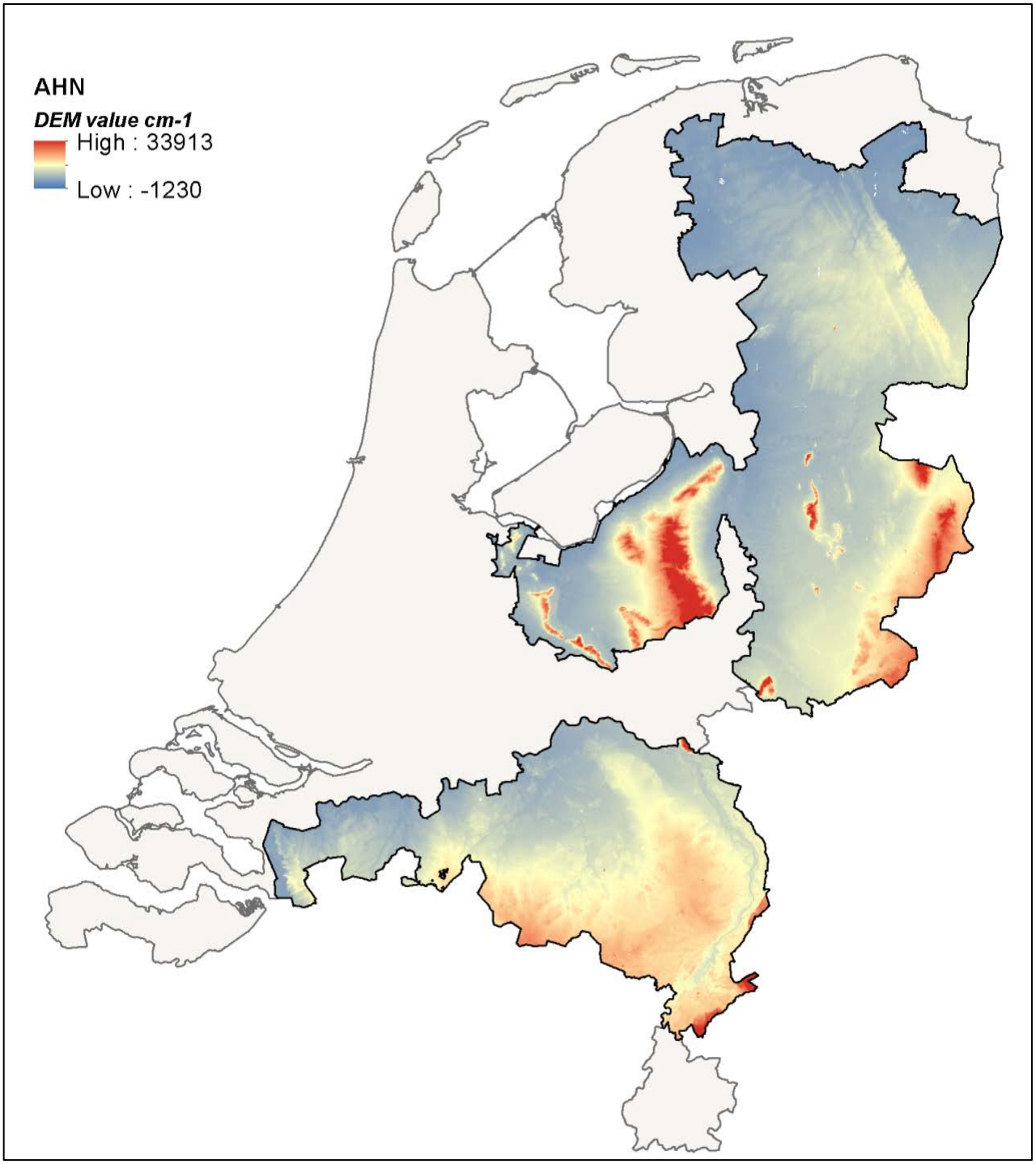
8d. slont



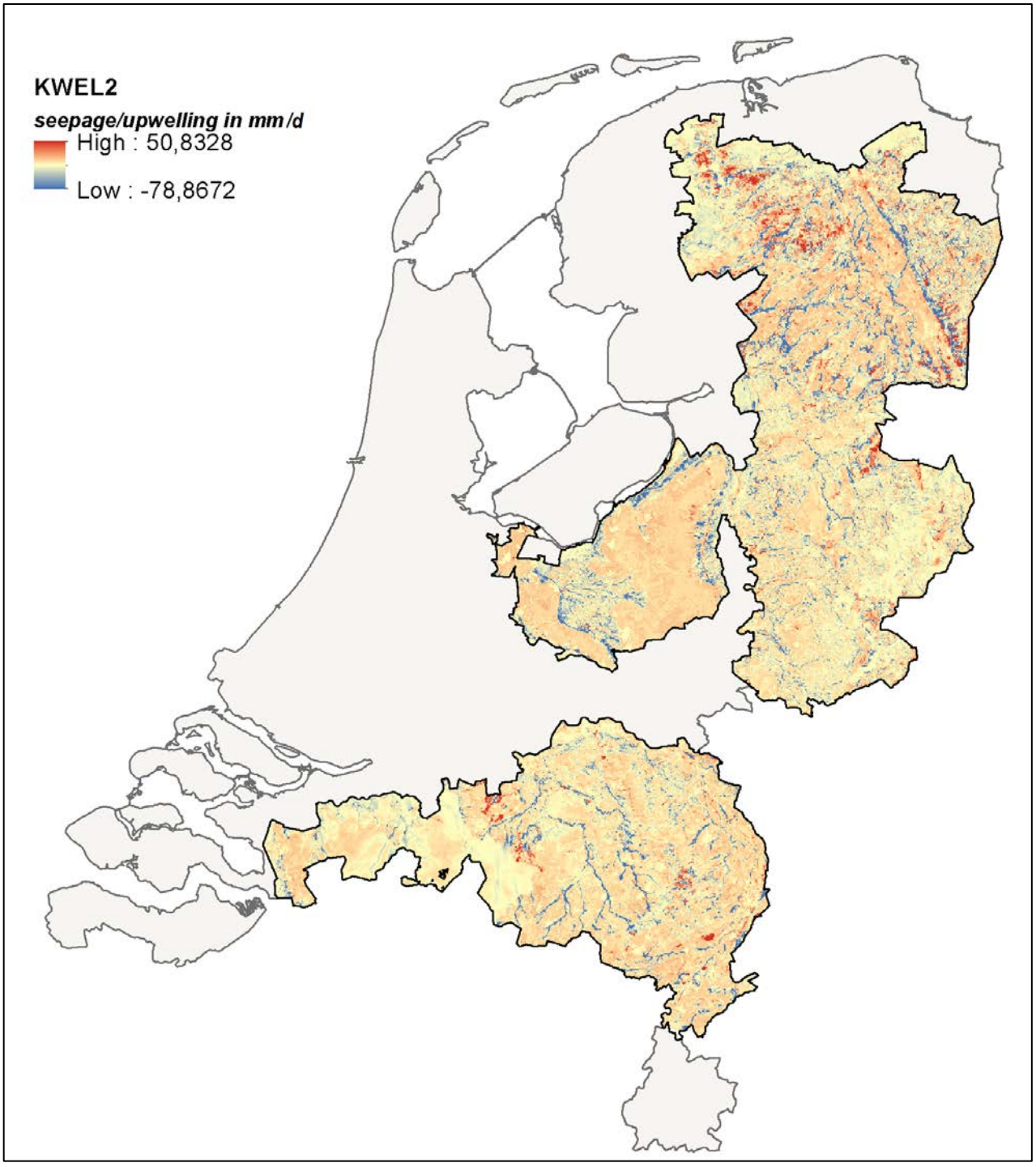
8e. draf



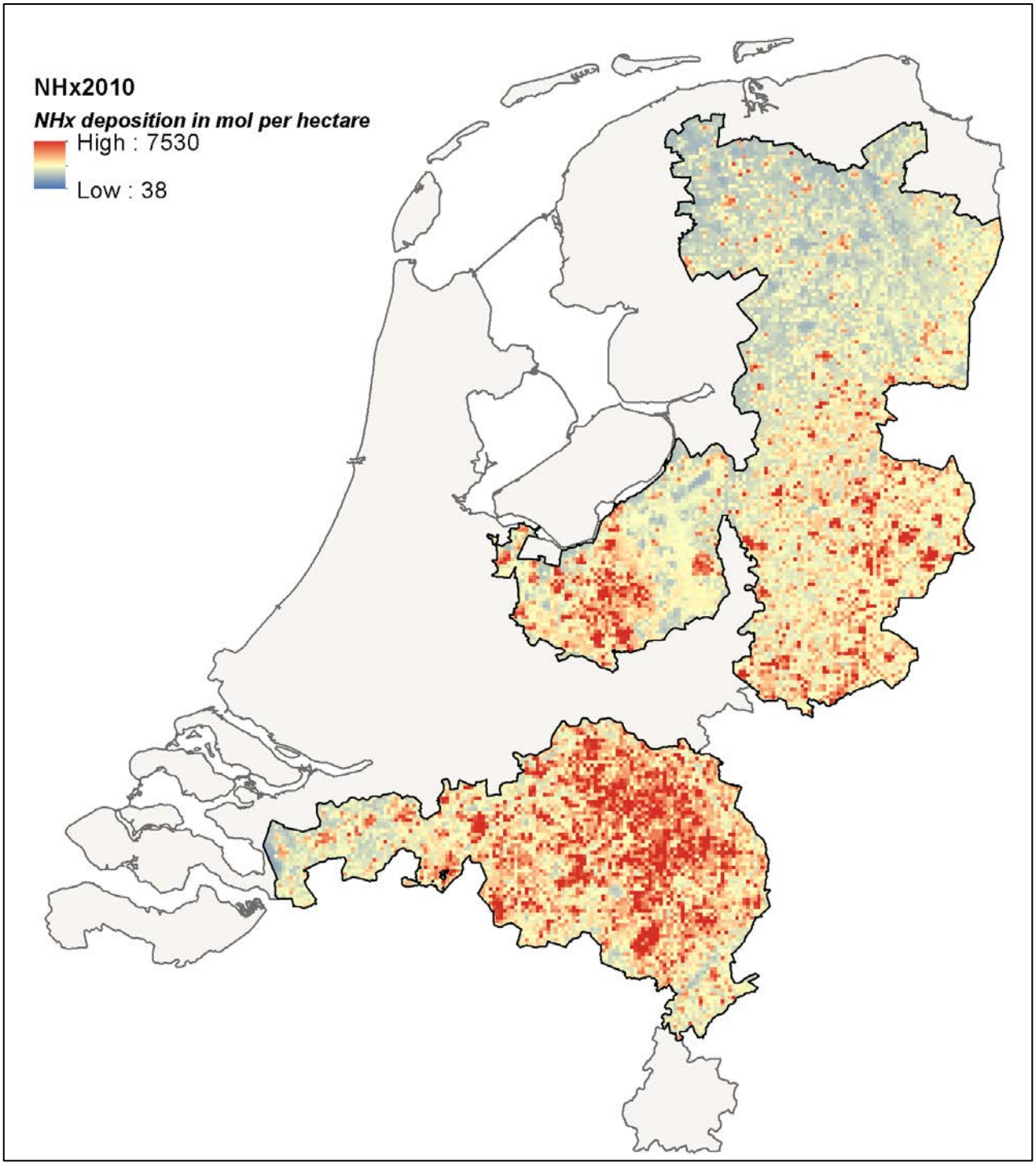
8f. dront



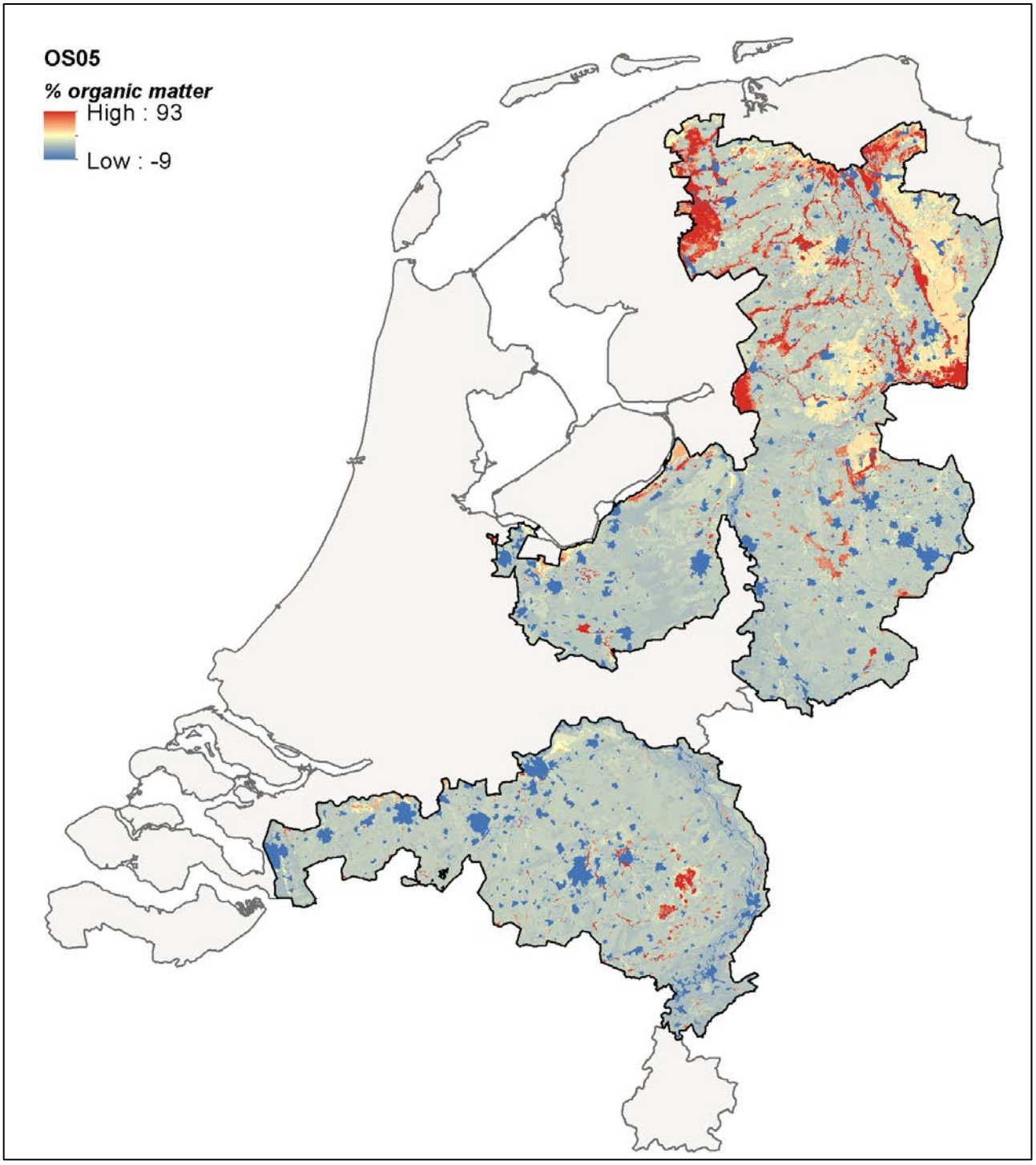
9. ahn



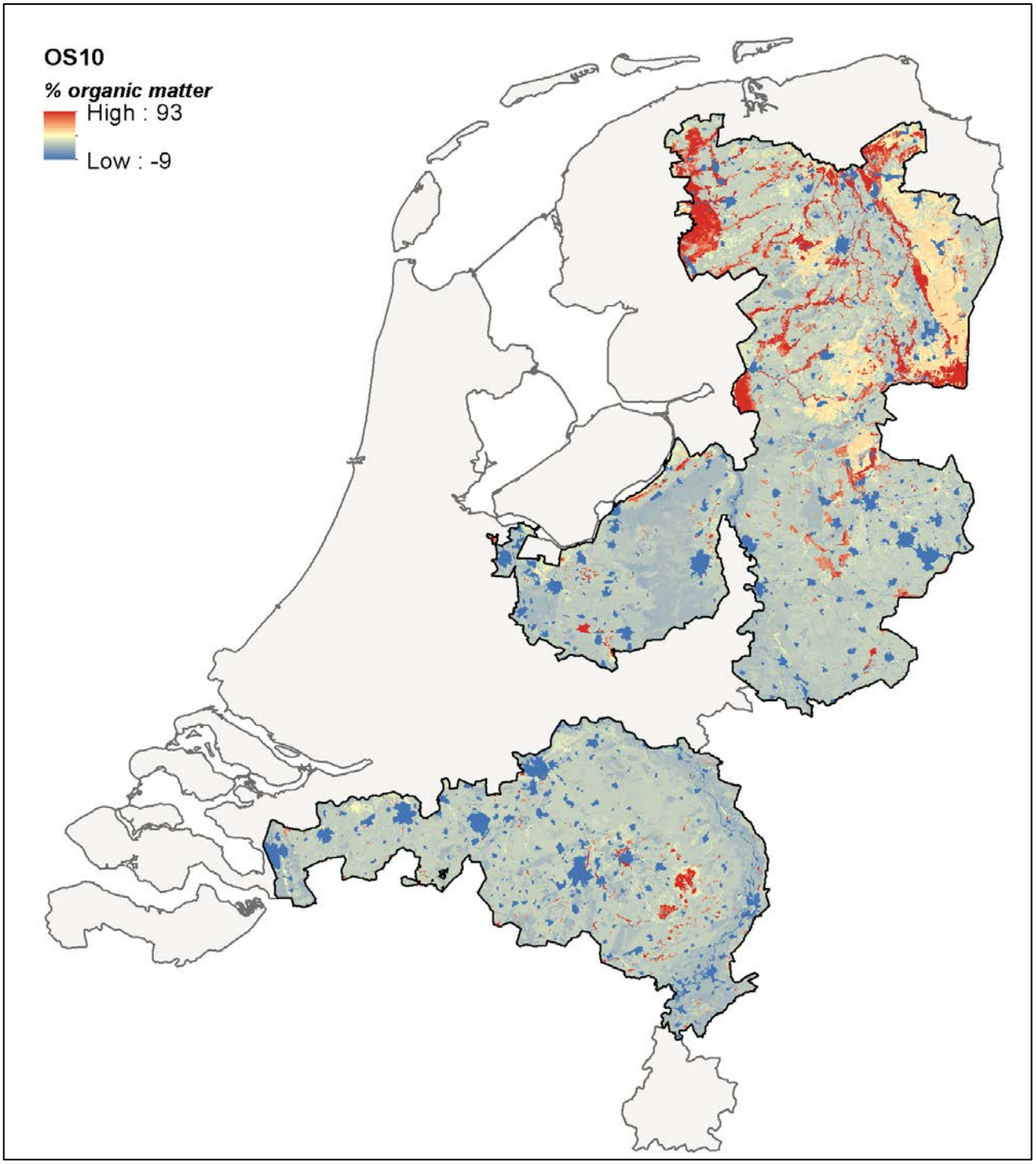
10. kwel2



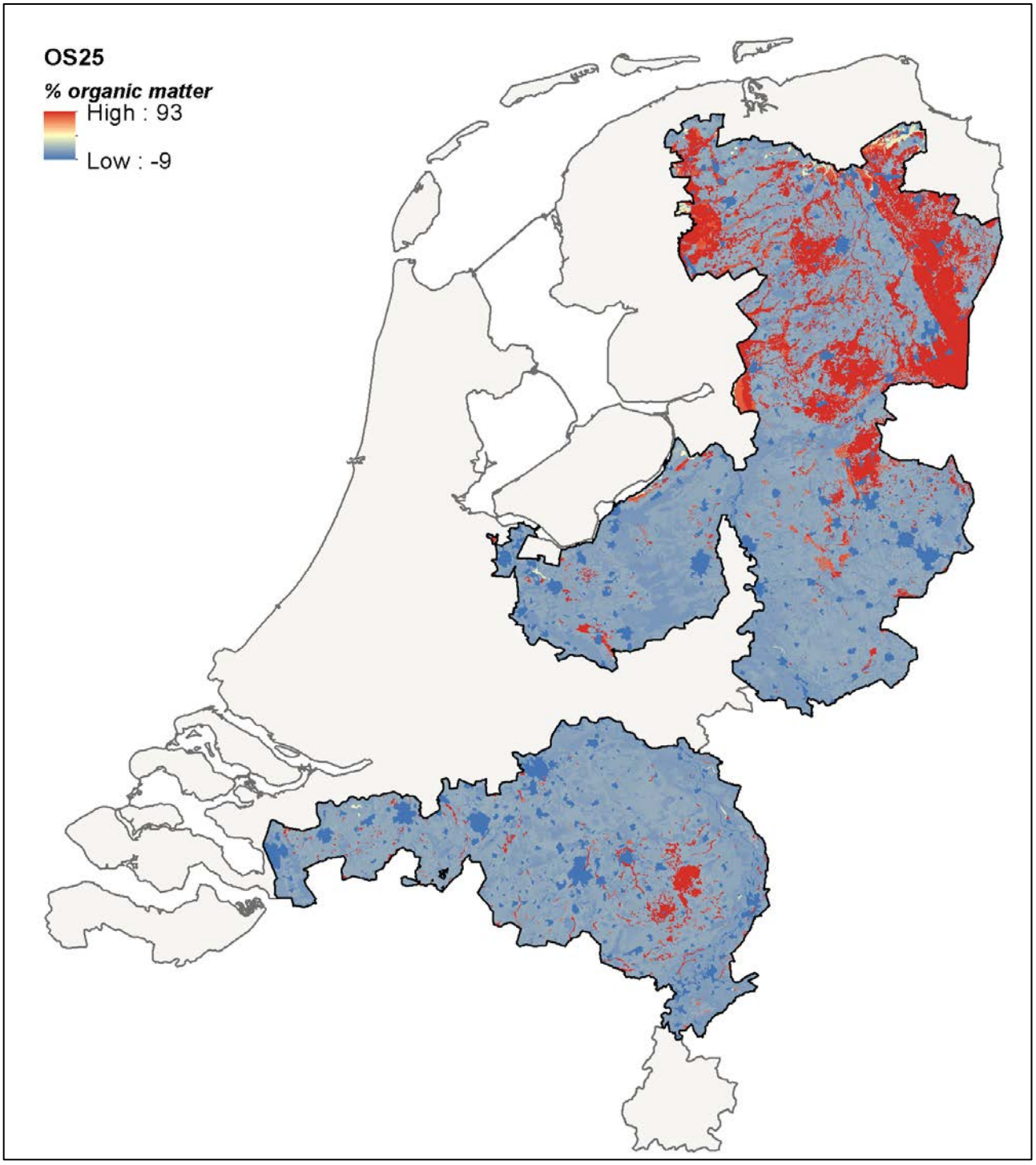
11. nhx



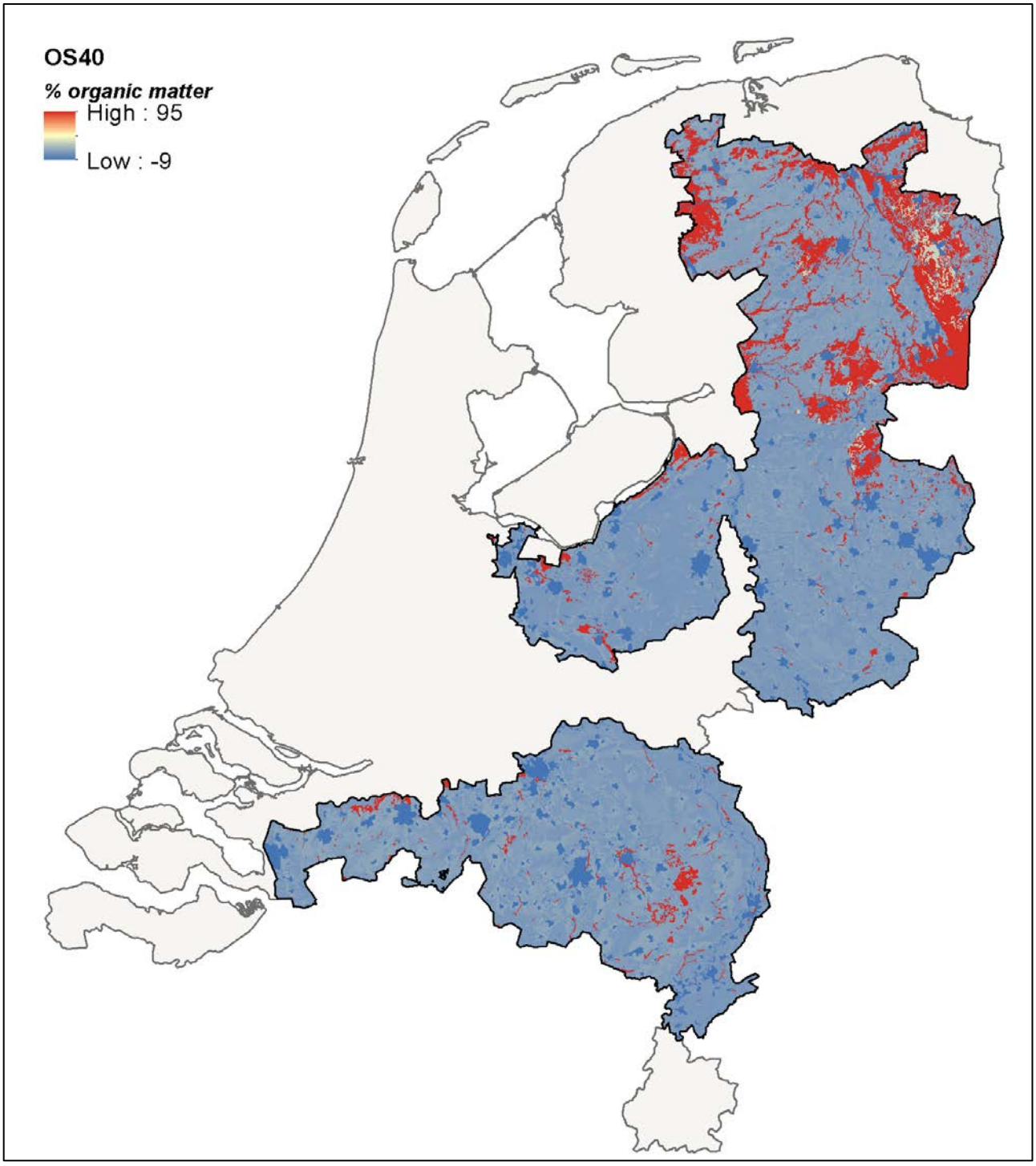
12a. om05



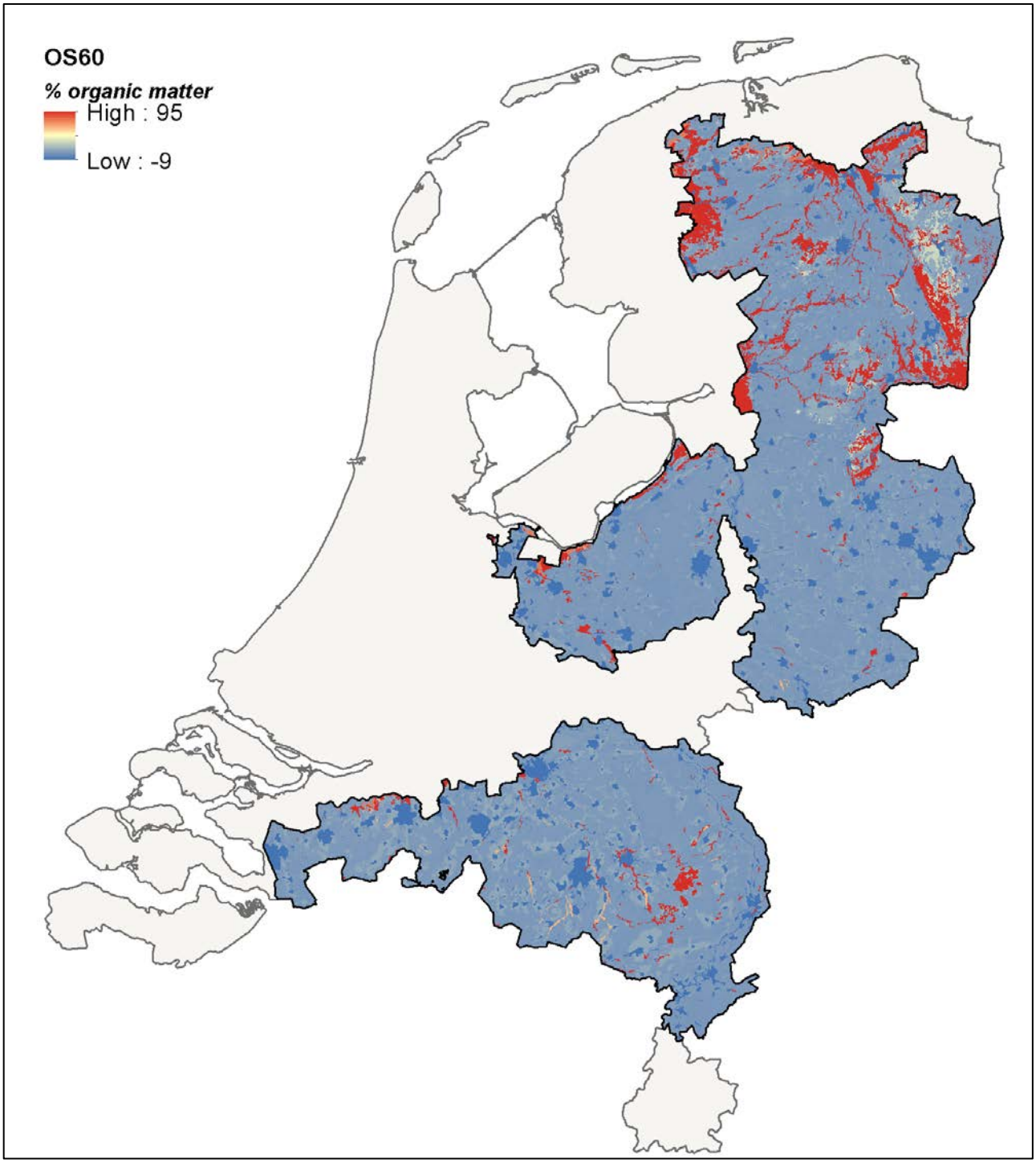
12b. om10



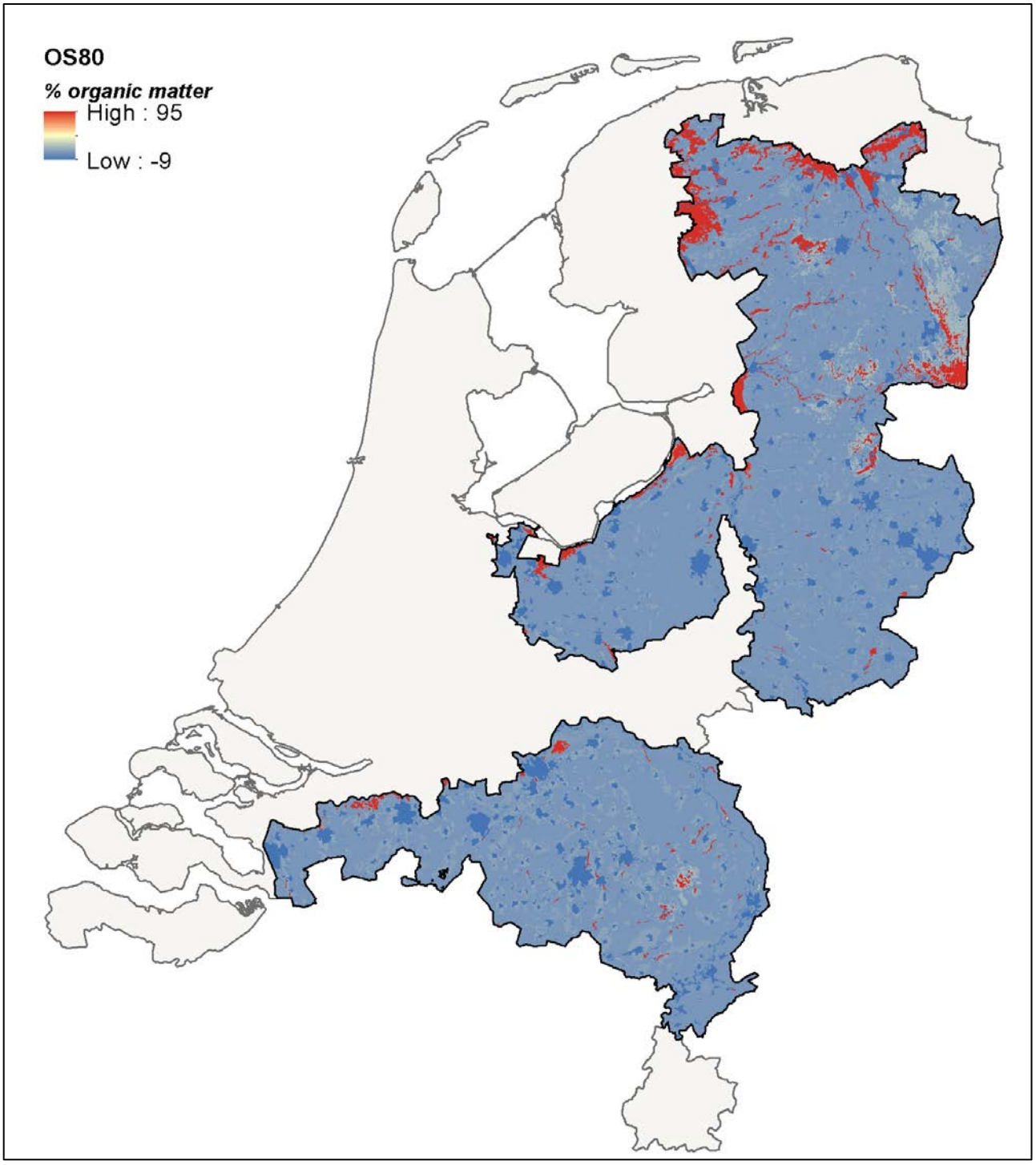
12c. om25



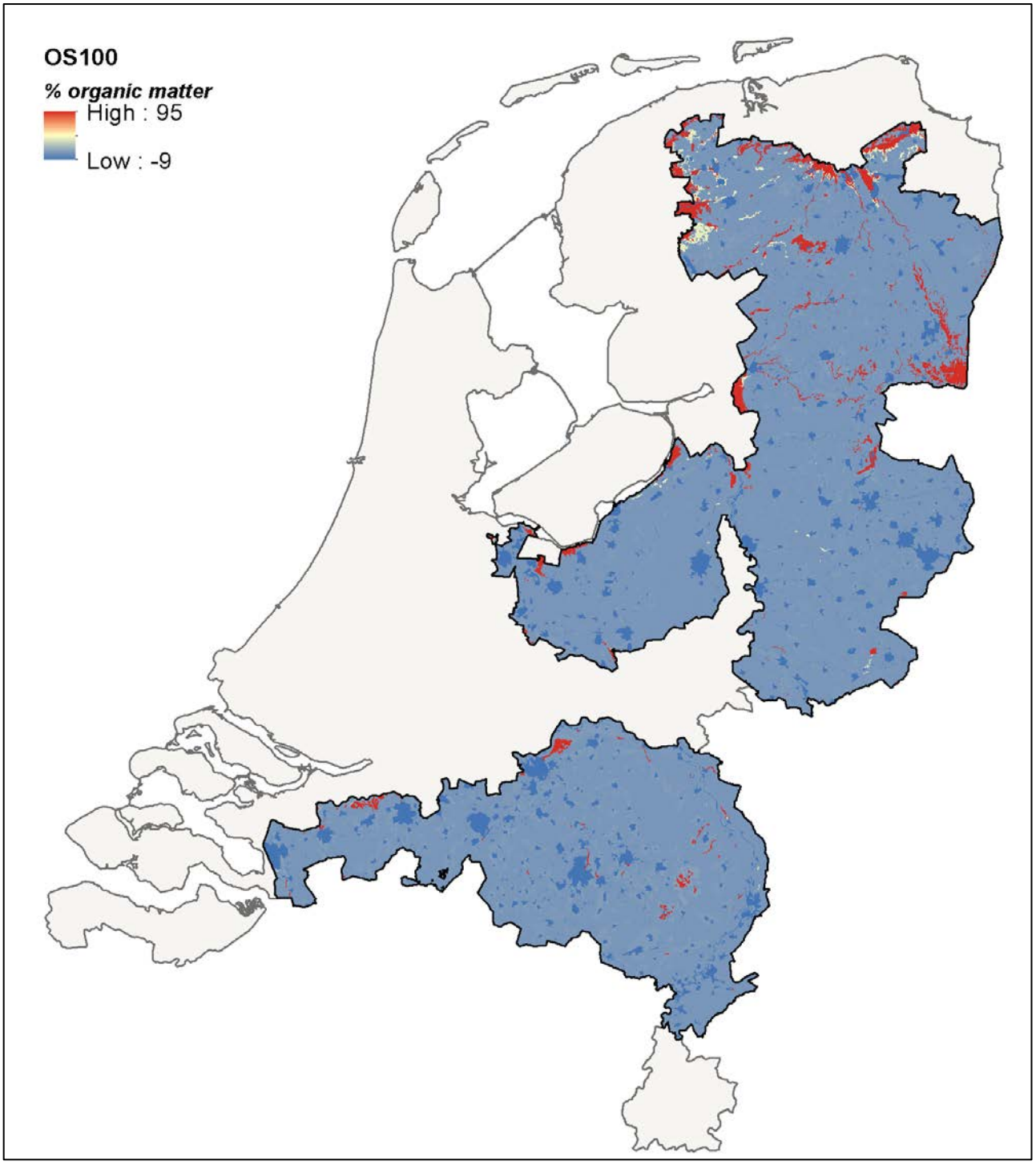
12d. om40



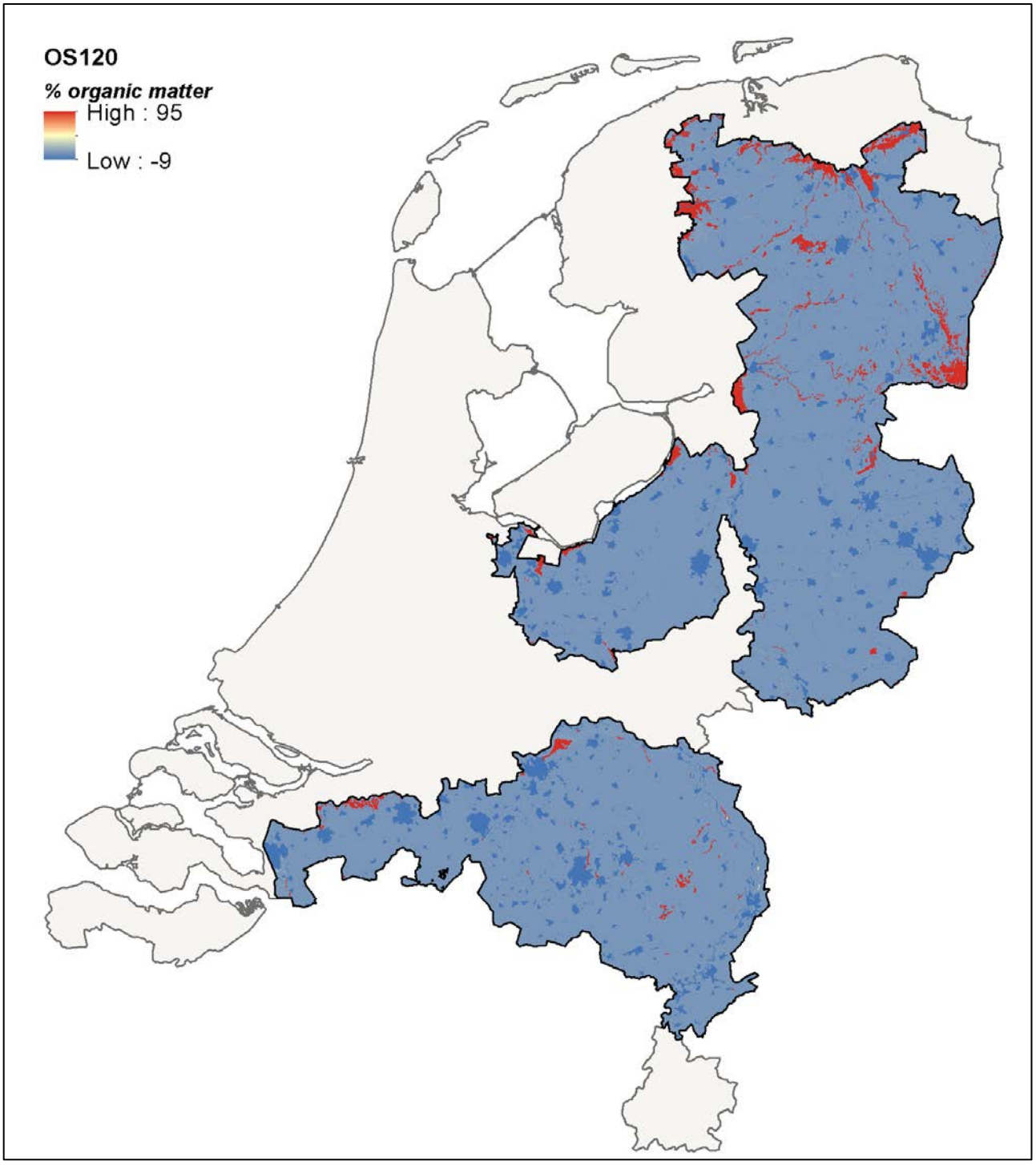
12e. om60



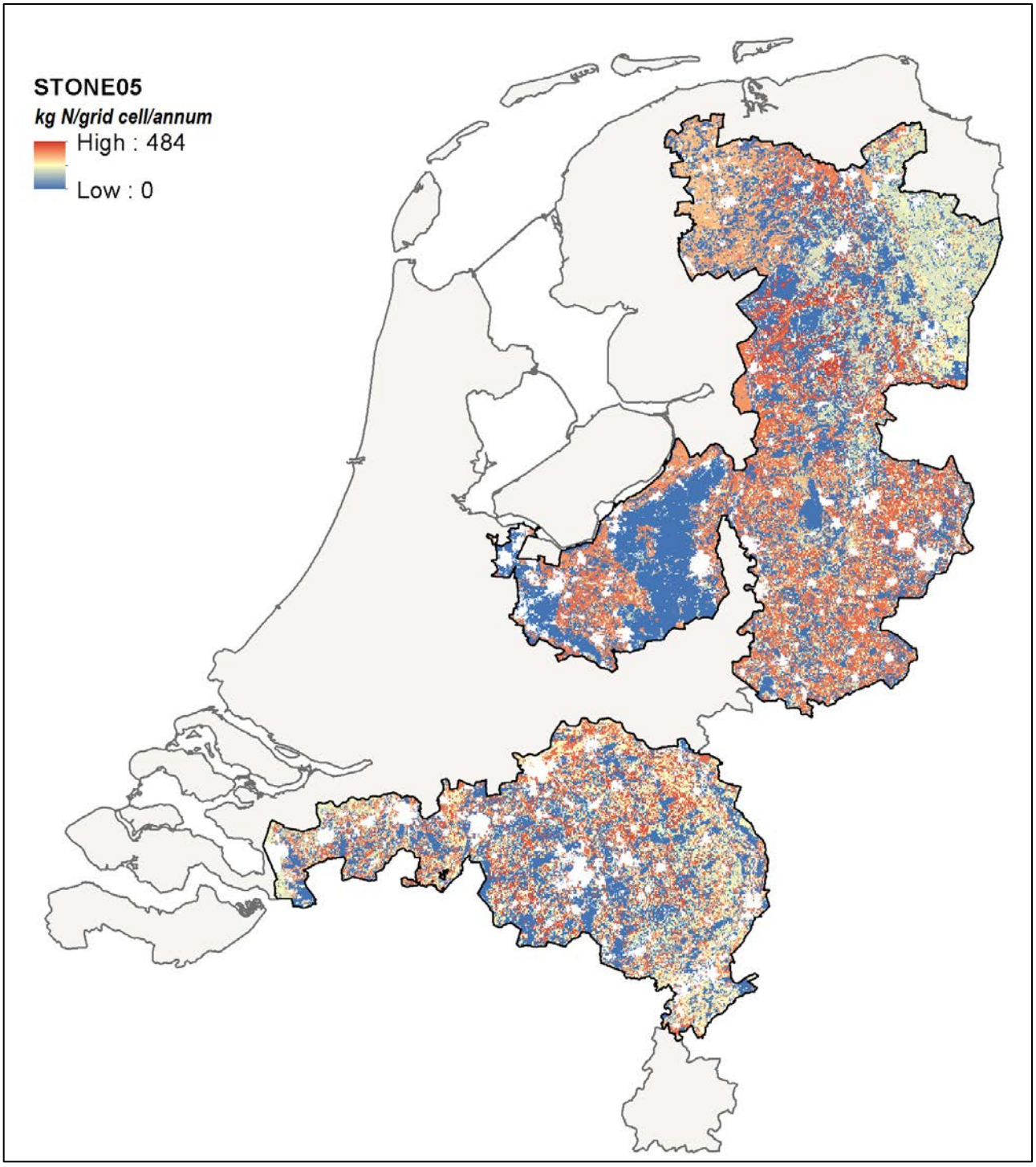
12f. om80



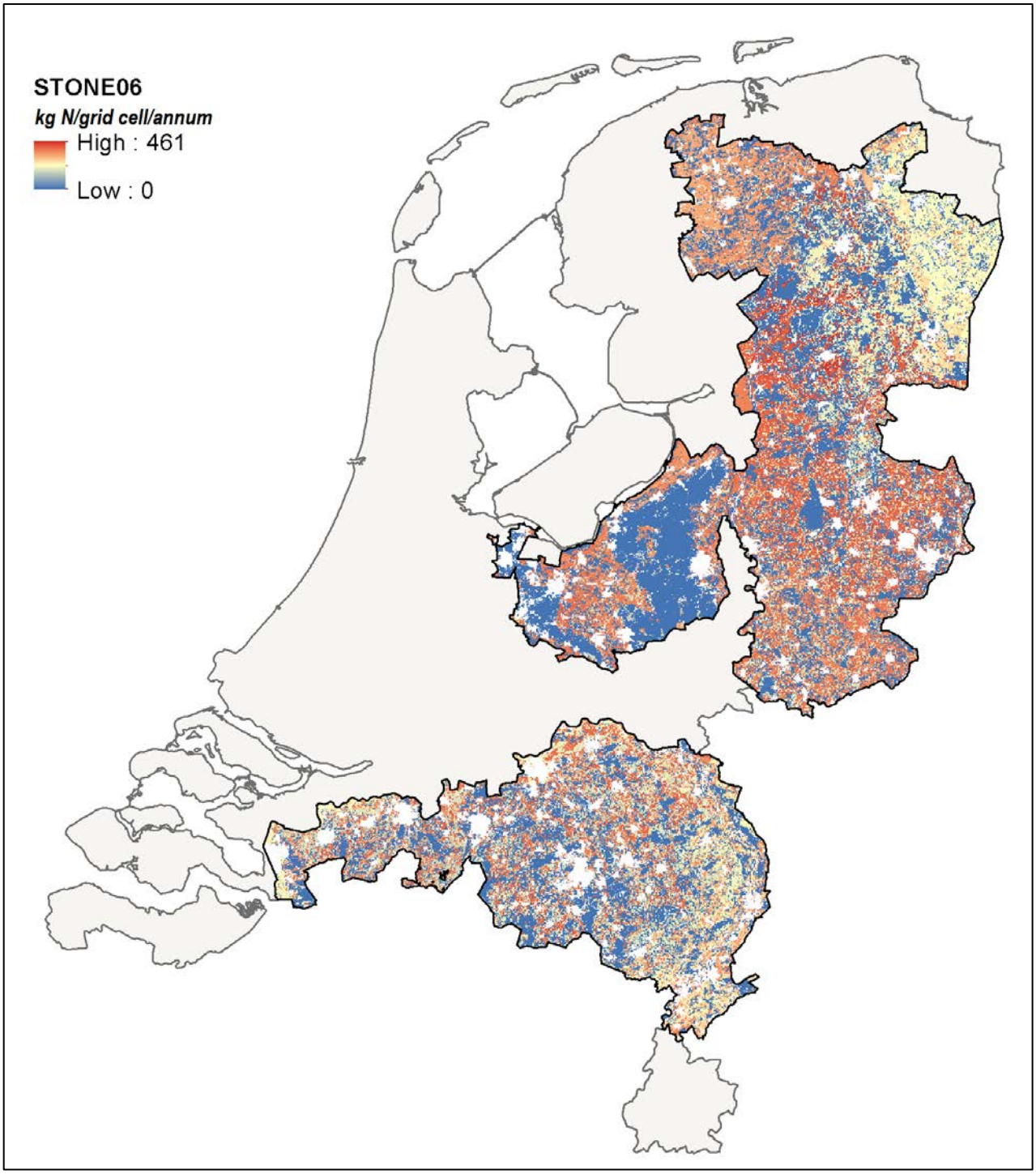
12g. om100



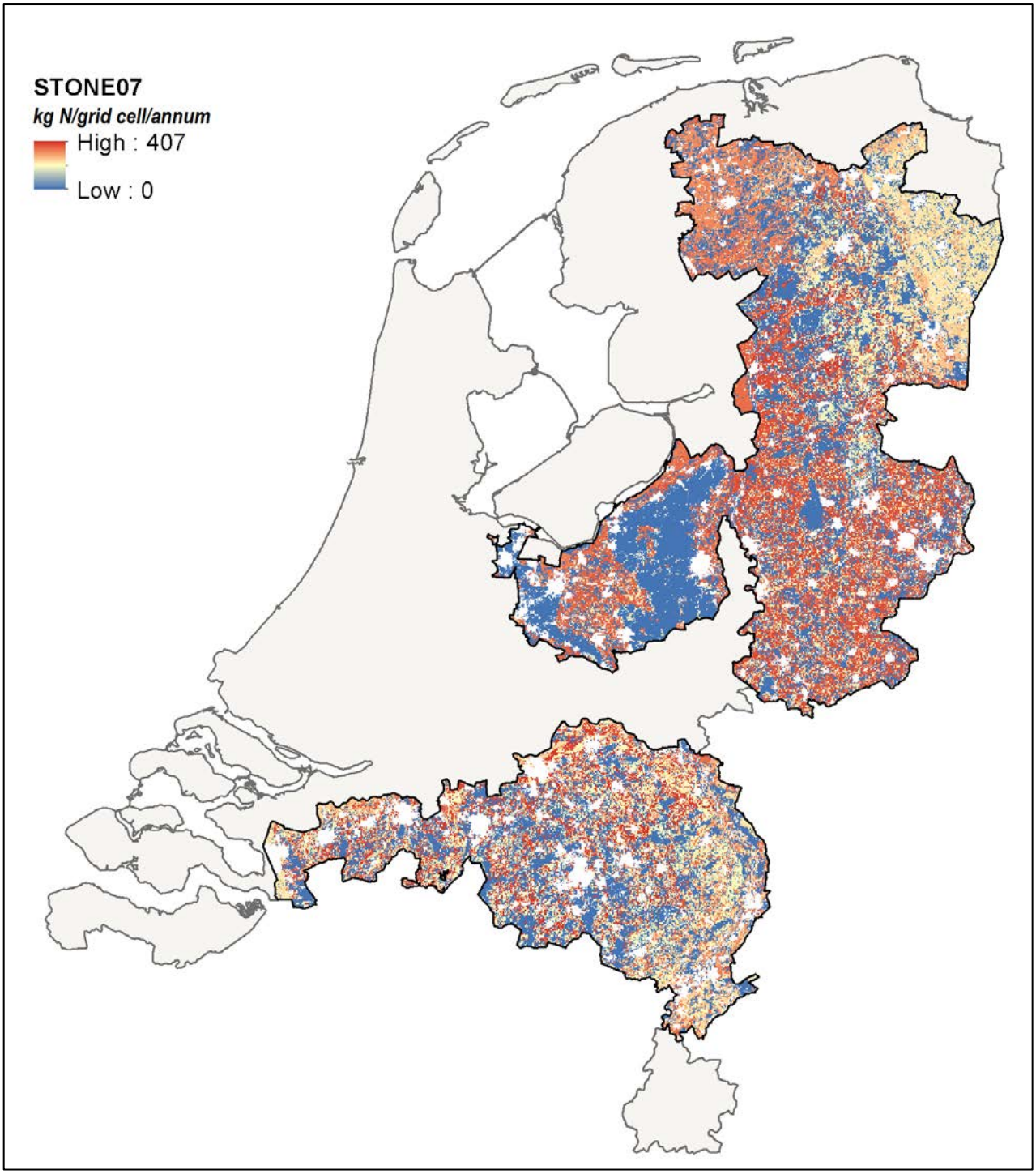
12h. om120



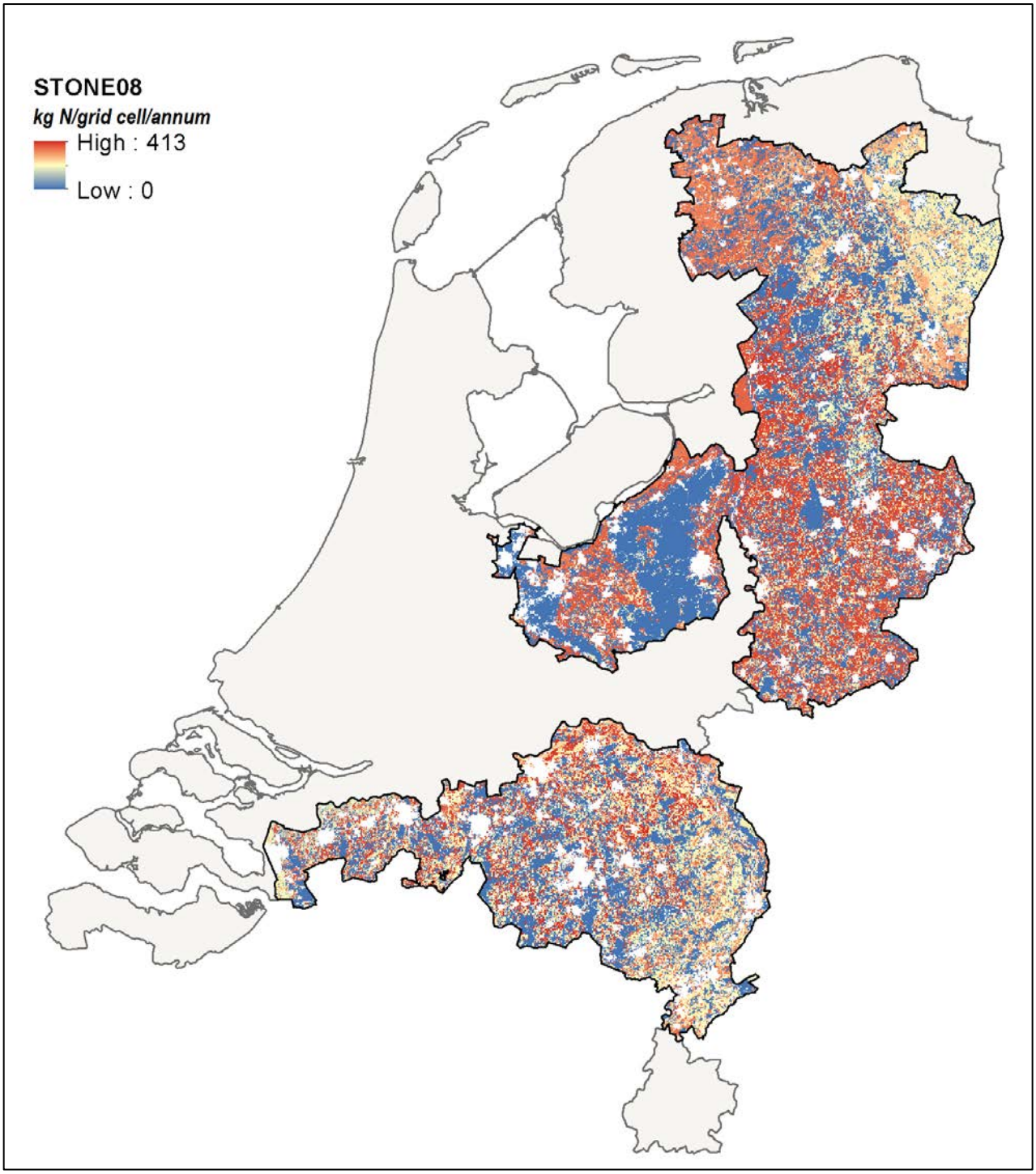
13a. stone05



13b. stone06



13c. stone07



13d. stone08

Appendix VI - Verbose model summary

In this appendix the model listings are presented, as given by the R-command 'summary'.

NORTH.2007

```
Call:
lm(formula = log10no3 ~ om60 + om100 + om120 + gt06 + bbg06 +
    gronds + kwel2 + pawn + lgn6 + laf + slont, data = north.2007)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26347 -0.30383 -0.00808  0.27774  1.39253

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0517496  0.6957270   1.512  0.13077
om60        -0.0014179  0.0005963  -2.378  0.01752 *
om100       -0.0107017  0.0037426  -2.859  0.00429 **
om120        0.0116977  0.0037848   3.091  0.00203 **
gt062       -0.1209652  0.4684257  -0.258  0.79625
gt063       -0.0987239  0.4756487  -0.208  0.83560
gt064       -0.1095778  0.4701918  -0.233  0.81575
gt065       -0.1417886  0.4703844  -0.301  0.76312
gt066        0.1086236  0.4715884   0.230  0.81786
gt067        0.0112239  0.4707210   0.024  0.98098
gt068        0.0471567  0.4707773   0.100  0.92022
gt069        0.1267895  0.4705560   0.269  0.78762
gt0610       0.0438451  0.4728976   0.093  0.92614
gt0611      -0.0002595  0.4809721  -0.001  0.99957
bbg0611      0.0872628  0.4942344   0.177  0.85987
bbg0612     -0.7060962  0.5672392  -1.245  0.21336
bbg0651      0.0963622  0.4656476   0.207  0.83608
bbg0660     -1.3996853  0.5143468  -2.721  0.00656 **
bbg0661     -1.8122761  0.5694748  -3.182  0.00148 **
bbg0662     -0.9686347  0.6863513  -1.411  0.15832
bbg0678     -0.3598651  0.6942149  -0.518  0.60426
gronds20     0.2351863  0.0512643   4.588 4.77e-06 ***
gronds21     0.0794647  0.0665425   1.194  0.23255
gronds30     0.3624874  0.1773714   2.044  0.04112 *
gronds40     0.5170535  0.4599234   1.124  0.26106
gronds50     0.6165498  0.3021409   2.041  0.04143 *
gronds60     0.6128928  0.3225856   1.900  0.05759 .
kwel2        0.0176427  0.0041678   4.233 2.41e-05 ***
pawn2        0.1542847  0.1133014   1.362  0.17345
pawn3       -0.0550947  0.1271875  -0.433  0.66494
pawn4       -0.2197801  0.2213453  -0.993  0.32087
pawn5        0.0173020  0.1092898   0.158  0.87423
pawn6       -0.8527258  0.4194220  -2.033  0.04218 *
pawn7       -0.1580441  0.1738940  -0.909  0.36354
pawn8        0.0818073  0.1559927   0.524  0.60004
pawn9        0.2558073  0.1139656   2.245  0.02491 *
pawn11      -0.0571943  0.1145380  -0.499  0.61759
pawn12       0.3132706  0.1561139   2.007  0.04492 *
pawn13       0.0599806  0.1163425   0.516  0.60623
pawn15      -0.7358530  0.6622158  -1.111  0.26662
pawn16      -0.8257450  0.3527683  -2.341  0.01935 *
pawn17      -0.7527901  0.3455890  -2.178  0.02951 *
pawn18      -0.7430296  0.3343847  -2.222  0.02639 *
pawn19      -0.3313232  0.2702993  -1.226  0.22044
lgn62        0.3263205  0.0396062   8.239 3.18e-16 ***
lgn63        0.3682191  0.0358075  10.283  < 2e-16 ***
lgn64        0.2902438  0.0548065   5.296 1.32e-07 ***
```

```
lgn65        0.2504449  0.0414512   6.042 1.83e-09 ***
lgn66        0.3203473  0.1087952   2.944  0.00327 **
lgn610       0.0295097  0.2065048   0.143  0.88638
lgn611       0.5354774  0.2077763   2.577  0.01004 *
lgn612       0.4162022  0.2317963   1.796  0.07272 .
lgn616       0.2209583  0.2318962   0.953  0.34080
lgn623       0.6961387  0.4575516   1.521  0.12831
lgn625      -0.2364387  0.4594733  -0.515  0.60690
lgn626      -0.3667322  0.3280249  -1.118  0.26371
lgn636       0.7336303  0.3314756   2.213  0.02700 *
lgn637       0.7626899  0.4089600   1.865  0.06234 .
lgn645      -0.3154253  0.1243609  -2.536  0.01128 *
laf2        -0.1606095  0.2303050  -0.697  0.48565
laf3        -0.1744700  0.2300197  -0.759  0.44825
laf4        -0.2574901  0.2306431  -1.116  0.26439
laf5        -0.2471460  0.2312128  -1.069  0.28524
laf6        -0.4404772  0.2327098  -1.893  0.05853 .
laf7        -0.4140158  0.2322515  -1.783  0.07481 .
laf8        -0.5370517  0.2358012  -2.278  0.02286 *
laf9        -0.5591825  0.2446204  -2.286  0.02237 *
slont2       0.0150049  0.0914816   0.164  0.86973
slont3       0.0478255  0.0893470   0.535  0.59252
slont4       0.1534357  0.0884558   1.735  0.08297 .
slont5       0.1312583  0.0910744   1.441  0.14969
slont6       0.2707116  0.0888157   3.048  0.00234 **
slont7       0.2094812  0.0908158   2.307  0.02118 *
slont8       0.2650114  0.0950278   2.789  0.00534 **
slont9       0.2979247  0.0973616   3.060  0.00224 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4548 on 1908 degrees of freedom
Multiple R-squared:  0.4926,  Adjusted R-squared:  0.4729
F-statistic: 25.03 on 74 and 1908 DF,  p-value: < 2.2e-16
```

NORTH.2008

```
Call:
lm(formula = log10no3 ~ om60 + om100 + gt06 + ahn + bbg06 + kwel2 +
    pawn + lgn6 + lont + slaf, data = north.2008)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25915 -0.30723 -0.03577  0.27739  1.89346

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.530e+00  6.414e-01   2.386 0.017126 *
om60        -2.441e-03  5.005e-04  -4.878 1.15e-06 ***
om100       -1.782e-03  8.170e-04  -2.181 0.029316 *
gt061       -6.238e-01  3.209e-01  -1.944 0.052085 .
gt062       -3.782e-01  1.008e-01  -3.752 0.000180 ***
gt063       -5.434e-01  1.784e-01  -3.046 0.002351 **
gt064       -4.700e-01  1.048e-01  -4.484 7.71e-06 ***
gt065       -5.859e-01  1.023e-01  -5.725 1.18e-08 ***
gt066       -5.068e-01  1.068e-01  -4.744 2.23e-06 ***
gt067       -4.805e-01  1.070e-01  -4.490 7.51e-06 ***
gt068       -3.965e-01  1.081e-01  -3.669 0.000250 ***
gt069       -4.075e-01  1.053e-01  -3.870 0.000112 ***
gt0610      -3.746e-01  1.105e-01  -3.390 0.000712 ***
gt0611      -3.476e-01  1.514e-01  -2.296 0.021777 *
ahn          5.377e-05  2.240e-05   2.400 0.016477 *
bbg0611      2.237e-02  4.766e-01   0.047 0.962561
bbg0651      6.242e-02  4.431e-01   0.141 0.887997
bbg0660     -1.381e+00  4.845e-01  -2.851 0.004398 **
bbg0661     -1.495e+00  5.645e-01  -2.648 0.008153 **
bbg0662     -1.734e+00  5.816e-01  -2.981 0.002904 **
kwel2        1.383e-02  3.481e-03   3.973 7.32e-05 ***
pawn2       -1.277e-01  9.774e-02  -1.306 0.191641
pawn3       -2.020e-01  1.091e-01  -1.851 0.064280 .
pawn4       -1.026e-01  1.427e-01  -0.719 0.472236
pawn5       -8.579e-02  9.255e-02  -0.927 0.354035
pawn6       -2.388e-01  1.802e-01  -1.325 0.185163
pawn7       -3.028e-01  1.601e-01  -1.891 0.058762 .
pawn8       -5.413e-03  1.221e-01  -0.044 0.964647
pawn9        8.784e-02  9.763e-02   0.900 0.368363
pawn10       2.031e-01  1.773e-01   1.145 0.252287
pawn11      -1.252e-01  9.765e-02  -1.282 0.199935
pawn12       3.024e-01  1.447e-01   2.089 0.036805 *
pawn13       5.812e-02  9.963e-02   0.583 0.559725
pawn15      -3.641e-01  3.262e-01  -1.116 0.264483
pawn16      -2.677e-01  1.424e-01  -1.880 0.060264 .
pawn17      -3.633e-01  1.113e-01  -3.265 0.001113 **
pawn18      -1.833e-01  1.010e-01  -1.816 0.069569 .
pawn19      -3.844e-01  1.491e-01  -2.578 0.010016 *
lgn62        3.613e-01  3.799e-02   9.511  < 2e-16 ***
lgn63        4.524e-01  3.324e-02  13.608  < 2e-16 ***
lgn64        2.915e-01  5.019e-02   5.808 7.27e-09 ***
lgn65        3.047e-01  3.724e-02   8.182 4.79e-16 ***
lgn66        2.501e-01  7.004e-02   3.571 0.000364 ***
lgn610       2.286e-01  2.220e-01   1.030 0.303164
lgn611       4.993e-01  1.901e-01   2.627 0.008679 **
lgn612       5.246e-01  2.039e-01   2.572 0.010168 *
lgn625       8.063e-01  6.247e-01   1.291 0.196930
lgn626       7.013e-01  4.416e-01   1.588 0.112417
lgn636       5.777e-01  3.314e-01   1.743 0.081475 .
lgn637       8.655e-01  3.744e-01   2.312 0.020875 *
lgn638       1.170e+00  3.669e-01   3.188 0.001454 **
lgn641      -2.274e-01  3.194e-01  -0.712 0.476667
lgn645      -7.536e-02  1.205e-01  -0.625 0.531721
lont2       -1.543e-01  4.422e-01  -0.349 0.727091
```

```
lont3       -2.015e-01  4.424e-01  -0.456 0.648762
lont4       -1.544e-01  4.427e-01  -0.349 0.727313
lont5       -1.976e-01  4.430e-01  -0.446 0.655683
lont6       -2.170e-01  4.433e-01  -0.490 0.624462
lont7       -3.220e-01  4.434e-01  -0.726 0.467702
lont8       -3.512e-01  4.436e-01  -0.792 0.428605
lont9       -3.071e-01  4.481e-01  -0.685 0.493114
slaf2        3.252e-02  1.042e-01   0.312 0.754940
slaf3        4.905e-03  1.023e-01   0.048 0.961771
slaf4        2.341e-01  9.448e-02   2.478 0.013279 *
slaf5        2.853e-01  9.496e-02   3.004 0.002692 **
slaf6        3.241e-01  9.338e-02   3.470 0.000530 ***
slaf7        3.392e-01  9.595e-02   3.535 0.000416 ***
slaf8        3.016e-01  9.880e-02   3.053 0.002297 **
slaf9        3.217e-01  9.987e-02   3.221 0.001297 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4374 on 2098 degrees of freedom
Multiple R-squared:  0.5038,  Adjusted R-squared:  0.4877
F-statistic: 31.32 on 68 and 2098 DF,  p-value: < 2.2e-16
```

NORTH.2009

```
Call:
lm(formula = log10no3 ~ stone7 + stone8 + gt06 + ahn + bbg06 +
    kwel2 + lgn6 + laf + slont + vds, data = north.2009)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07065 -0.28395 -0.04377  0.26296  1.92579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.969e+00  5.149e-01   3.823 0.000136 ***
stone7       3.853e-03  1.602e-03   2.405 0.016280 *
stone8      -3.687e-03  1.551e-03  -2.378 0.017496 *
gt062       -4.380e-01  9.605e-02  -4.560 5.44e-06 ***
gt063       -4.679e-01  1.343e-01  -3.483 0.000506 ***
gt064       -4.847e-01  9.583e-02  -5.058 4.63e-07 ***
gt065       -4.850e-01  8.937e-02  -5.427 6.46e-08 ***
gt066       -3.369e-01  8.974e-02  -3.754 0.000179 ***
gt067       -4.397e-01  9.319e-02  -4.718 2.56e-06 ***
gt068       -3.619e-01  9.324e-02  -3.881 0.000107 ***
gt069       -2.224e-01  9.075e-02  -2.450 0.014365 *
gt0610      -2.162e-01  9.910e-02  -2.182 0.029246 *
gt0611      -5.642e-02  1.599e-01  -0.353 0.724288
ahn          5.812e-05  2.186e-05   2.658 0.007927 **
bbg0611     -3.513e-01  4.651e-01  -0.755 0.450101
bbg0612     -3.461e-01  6.119e-01  -0.565 0.571802
bbg0651      1.508e-02  4.325e-01   0.035 0.972181
bbg0660     -9.107e-01  4.759e-01  -1.914 0.055796 .
bbg0662     -1.869e+00  6.617e-01  -2.824 0.004790 **
bbg0678     -2.520e-02  6.509e-01  -0.039 0.969125
kwel2        1.567e-02  3.754e-03   4.176 3.10e-05 ***
lgn62        2.913e-01  3.553e-02   8.200 4.35e-16 ***
lgn63        2.976e-01  3.244e-02   9.174  < 2e-16 ***
lgn64        2.367e-01  5.174e-02   4.574 5.09e-06 ***
lgn65        2.404e-01  3.671e-02   6.548 7.45e-11 ***
lgn66        2.319e-01  7.656e-02   3.029 0.002484 **
lgn610       2.600e-01  2.177e-01   1.194 0.232606
lgn611      -1.319e-01  1.618e-01  -0.815 0.415024
lgn612      -3.068e-01  2.356e-01  -1.302 0.193129
lgn616       2.219e-01  2.174e-01   1.021 0.307617
lgn625       7.331e-01  6.125e-01   1.197 0.231501
lgn626       5.611e-01  3.064e-01   1.832 0.067165 .
lgn636       5.914e-01  5.147e-01   1.149 0.250658
lgn637       5.365e-01  5.720e-01   0.938 0.348438
lgn641      -1.246e-01  3.166e-01  -0.394 0.693877
lgn645      -2.497e-01  2.172e-01  -1.149 0.250516
laf2        -4.942e-01  2.527e-01  -1.956 0.050614 .
laf3        -4.908e-01  2.525e-01  -1.944 0.052014 .
laf4        -5.952e-01  2.528e-01  -2.354 0.018666 *
laf5        -5.614e-01  2.531e-01  -2.218 0.026700 *
laf6        -7.271e-01  2.547e-01  -2.855 0.004350 **
laf7        -6.316e-01  2.549e-01  -2.478 0.013290 *
laf8        -6.887e-01  2.569e-01  -2.680 0.007416 **
laf9        -8.237e-01  2.631e-01  -3.131 0.001771 **
slont2      -1.206e-01  7.938e-02  -1.520 0.128785
slont3      -8.540e-02  7.785e-02  -1.097 0.272756
slont4      -2.953e-02  7.669e-02  -0.385 0.700291
slont5      -2.381e-02  7.988e-02  -0.298 0.765637
slont6       1.266e-01  7.847e-02   1.613 0.106865
slont7       3.953e-02  8.036e-02   0.492 0.622876
slont8       7.430e-02  8.282e-02   0.897 0.369773
slont9       4.226e-02  8.858e-02   0.477 0.633335
vds8         3.311e-01  2.525e-01   1.311 0.189885
vds10       -1.885e-01  3.332e-02  -5.656 1.78e-08 ***
```

```
vds13        3.480e-02  3.238e-02   1.075 0.282625
vds14       -2.244e-01  5.291e-02  -4.241 2.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4302 on 1923 degrees of freedom
Multiple R-squared:  0.3899,  Adjusted R-squared:  0.3724
F-statistic: 22.34 on 55 and 1923 DF,  p-value: < 2.2e-16
```

EAST.2007

```
Call:
lm(formula = log10no3 ~ om10 + om40 + om80 + om120 + gt06 + bbg06 +
    kwel2 + geom + draf + slont, data = east.2007)

Residuals:
     Min       1Q   Median       3Q      Max
-1.34417 -0.32417  0.00993  0.31260  1.38341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1627065  0.3658019   0.445 0.656541
om10        -0.0153918  0.0035055  -4.391 1.22e-05 ***
om40        -0.0026568  0.0008872  -2.995 0.002799 **
om80         0.0388828  0.0110427   3.521 0.000444 ***
om120       -0.0327584  0.0105945  -3.092 0.002030 **
gt062        0.4128320  0.1570043   2.629 0.008652 **
gt064        0.4181232  0.1235981   3.383 0.000738 ***
gt065        0.3791895  0.1223440   3.099 0.001980 **
gt066        0.5841136  0.1362538   4.287 1.94e-05 ***
gt067        0.3566603  0.1293930   2.756 0.005924 **
gt068        0.7116877  0.1250868   5.690 1.57e-08 ***
gt069        0.6746497  0.1208836   5.581 2.90e-08 ***
gt0610       0.6828732  0.1252212   5.453 5.89e-08 ***
gt0611       0.6844391  0.1357799   5.041 5.28e-07 ***
bbg0634     -0.8863613  0.5265835  -1.683 0.092566 .
bbg0651     -0.3822667  0.2362498  -1.618 0.105888
bbg0660     -0.8580007  0.2603191  -3.296 0.001007 **
kwel2        0.0300312  0.0105890   2.836 0.004637 **
geom4        0.0418590  0.3038533   0.138 0.890451
geom6       -0.8059465  0.2356915  -3.419 0.000646 ***
geom8       -0.1075786  0.1429072  -0.753 0.451712
geom10      -0.0701548  0.1958883  -0.358 0.720298
geom11       0.1014728  0.4879738   0.208 0.835302
geom12       0.0116317  0.1327098   0.088 0.930170
geom13      -0.0670924  0.1299158  -0.516 0.605640
geom14      -0.1422969  0.1303999  -1.091 0.275368
geom15      -0.5150570  0.1801359  -2.859 0.004313 **
geom16      -0.2513288  0.1338076  -1.878 0.060562 .
geom22      -0.1931407  0.3573618  -0.540 0.588969
draf4        0.2439962  0.0814126   2.997 0.002777 **
draf5        0.3421993  0.0804635   4.253 2.26e-05 ***
draf6        0.2544499  0.0849237   2.996 0.002784 **
draf7        0.5278256  0.2559507   2.062 0.039382 *
slont2       1.2235775  0.2031945   6.022 2.23e-09 ***
slont3       1.0802398  0.1962516   5.504 4.44e-08 ***
slont4       1.1697987  0.1937984   6.036 2.05e-09 ***
slont5       1.1616250  0.1939724   5.989 2.72e-09 ***
slont6       1.1910057  0.1952454   6.100 1.39e-09 ***
slont7       1.2802197  0.1964717   6.516 1.02e-10 ***
slont8       1.2417021  0.1972517   6.295 4.17e-10 ***
slont9       1.1696831  0.2112489   5.537 3.71e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4682 on 1323 degrees of freedom
Multiple R-squared:  0.3307,  Adjusted R-squared:  0.3104
F-statistic: 16.34 on 40 and 1323 DF,  p-value: < 2.2e-16
```

EAST.2008

```
Call:
lm(formula = log10no3 ~ om10 + om60 + om80 + om100 + stone6 +
    stone8 + gt06 + lgn6 + draf, data = east.2008)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26780 -0.36704 -0.03534  0.33557  1.40809

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.678168   0.159069   4.263 2.18e-05 ***
om10        -0.017478   0.003931  -4.447 9.56e-06 ***
om60        -0.007670   0.002961  -2.591 0.009703 **
om80         0.111106   0.021471   5.175 2.69e-07 ***
om100       -0.100698   0.019819  -5.081 4.38e-07 ***
stone6      -0.002719   0.001301  -2.090 0.036867 *
stone8       0.002793   0.001321   2.114 0.034702 *
gt062        0.298667   0.165111   1.809 0.070730 .
gt064        0.267431   0.134737   1.985 0.047401 *
gt065        0.205116   0.136838   1.499 0.134156
gt066        0.268829   0.154638   1.738 0.082400 .
gt067        0.189603   0.139916   1.355 0.175645
gt068        0.606576   0.135713   4.470 8.61e-06 ***
gt069        0.651036   0.133983   4.859 1.34e-06 ***
gt0610       0.564433   0.138121   4.087 4.68e-05 ***
gt0611       0.465810   0.151544   3.074 0.002164 **
lgn62        0.230234   0.040299   5.713 1.41e-08 ***
lgn63        0.178674   0.105882   1.687 0.091783 .
lgn64       -0.050409   0.240983  -0.209 0.834343
lgn65        0.279372   0.069002   4.049 5.49e-05 ***
lgn66        0.094201   0.141728   0.665 0.506400
lgn611      -0.545291   0.098353  -5.544 3.66e-08 ***
lgn612      -0.564593   0.291910  -1.934 0.053342 .
lgn625       0.039776   0.351587   0.113 0.909945
lgn626       0.728139   0.249777   2.915 0.003624 **
lgn636      -0.968827   0.497515  -1.947 0.051739 .
lgn642       0.420246   0.497541   0.845 0.398485
lgn645      -0.042420   0.098499  -0.431 0.666794
draf4        0.268682   0.096012   2.798 0.005221 **
draf5        0.358341   0.093004   3.853 0.000123 ***
draf6        0.416549   0.096831   4.302 1.84e-05 ***
draf7        0.324259   0.266800   1.215 0.224478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4947 on 1148 degrees of freedom
Multiple R-squared:  0.3066,  Adjusted R-squared:  0.2879
F-statistic: 16.38 on 31 and 1148 DF,  p-value: < 2.2e-16
```

EAST.2009

```
Call:
lm(formula = log10no3 ~ om25 + om80 + om100 + gt06 + ahn + kwel2 +
    lgn6 + draf, data = east.2009)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16092 -0.37001 -0.05292  0.37028  1.50329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.959e-01  1.576e-01   4.416 1.09e-05 ***
om25        -5.103e-03  9.105e-04  -5.605 2.58e-08 ***
om80         3.971e-02  1.480e-02   2.683  0.00739 **
om100       -3.868e-02  1.426e-02  -2.712  0.00678 **
gt062        2.319e-01  1.740e-01   1.333  0.18292
gt064        1.248e-01  1.433e-01   0.871  0.38407
gt065        1.591e-01  1.437e-01   1.107  0.26834
gt066        2.070e-01  1.526e-01   1.356  0.17541
gt067        9.897e-02  1.519e-01   0.652  0.51473
gt068        3.381e-01  1.445e-01   2.340  0.01947 *
gt069        4.141e-01  1.415e-01   2.927  0.00348 **
gt0610       3.339e-01  1.450e-01   2.303  0.02143 *
gt0611       3.545e-01  1.639e-01   2.163  0.03073 *
ahn          7.659e-05  1.707e-05   4.486 7.93e-06 ***
kwel2        2.906e-02  1.125e-02   2.583  0.00992 **
lgn62        3.715e-01  4.126e-02   9.004  < 2e-16 ***
lgn63        5.333e-01  9.541e-02   5.590 2.80e-08 ***
lgn64        5.942e-01  1.228e-01   4.838 1.48e-06 ***
lgn65        2.650e-01  6.223e-02   4.258 2.22e-05 ***
lgn66        1.594e-01  1.222e-01   1.305  0.19223
lgn611       2.423e-01  1.170e-01   2.070  0.03868 *
lgn623       5.484e-01  4.911e-01   1.117  0.26436
lgn625      -3.220e-01  3.474e-01  -0.927  0.35419
lgn626       4.542e-01  2.011e-01   2.258  0.02412 *
lgn641      -6.642e-03  4.920e-01  -0.014  0.98923
lgn642      -2.540e-01  3.498e-01  -0.726  0.46780
lgn645      -5.609e-02  1.041e-01  -0.539  0.58996
draf4        1.671e-01  7.478e-02   2.235  0.02559 *
draf5        1.570e-01  7.182e-02   2.186  0.02904 *
draf6        3.188e-01  7.918e-02   4.027 6.00e-05 ***
draf7        4.885e-01  3.562e-01   1.371  0.17049
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4893 on 1218 degrees of freedom
Multiple R-squared:  0.2657,  Adjusted R-squared:  0.2476
F-statistic: 14.69 on 30 and 1218 DF,  p-value: < 2.2e-16
```

CENTRE.2007

```
Call:
lm(formula = log10no3 ~ stone5 + stone6 + gt06 + bbg06 + geom +
    vds, data = centre.2007)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15228 -0.25873 -0.00632  0.22815  1.00688

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.180475   0.510931   4.268 2.44e-05 ***
stone5       0.004487   0.001271   3.532 0.000458 ***
stone6      -0.004490   0.001324  -3.390 0.000763 ***
gt061       -0.235214   0.277501  -0.848 0.397132
gt062       -0.029519   0.231268  -0.128 0.898493
gt063       -0.276103   0.458524  -0.602 0.547392
gt064        0.163277   0.227747   0.717 0.473817
gt065       -0.091414   0.237284  -0.385 0.700245
gt066        0.244848   0.237563   1.031 0.303285
gt067        0.796598   0.321425   2.478 0.013588 *
gt068        0.248816   0.357453   0.696 0.486759
gt069        0.490190   0.230690   2.125 0.034174 *
gt0610       0.554558   0.234823   2.362 0.018647 *
gt0611       0.795352   0.255174   3.117 0.001952 **
bbg0651     -0.983879   0.390001  -2.523 0.012009 *
bbg0660     -2.255815   0.418819  -5.386 1.20e-07 ***
bbg0661     -2.580530   0.573048  -4.503 8.66e-06 ***
bbg0678     -0.909755   0.554083  -1.642 0.101350
geom9        0.127671   0.217533   0.587 0.557581
geom10      -0.111606   0.222750  -0.501 0.616607
geom12       0.117967   0.185802   0.635 0.525831
geom13      -0.125162   0.117324  -1.067 0.286667
geom14      -0.167991   0.187397  -0.896 0.370524
geom16      -0.115397   0.212632  -0.543 0.587616
vds8        -0.379208   0.156096  -2.429 0.015542 *
vds10       -0.121783   0.165788  -0.735 0.463008
vds13       -0.042249   0.147323  -0.287 0.774422
vds14       -0.012639   0.167146  -0.076 0.939759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3846 on 424 degrees of freedom
Multiple R-squared:  0.584,   Adjusted R-squared:  0.5575
F-statistic: 22.05 on 27 and 424 DF,  p-value: < 2.2e-16
```

CENTRE.2008

```
Call:
lm(formula = log10no3 ~ nhx + gt06 + geom, data = centre.2008)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97948 -0.25679 -0.07953  0.20878  1.32694

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.017e-01  2.753e-01   1.822  0.06921 .
nhx         -5.593e-05  4.478e-05  -1.249  0.21245
gt061        3.452e-02  2.666e-01   0.130  0.89703
gt062        1.016e-01  2.452e-01   0.414  0.67879
gt063       -2.887e-01  4.843e-01  -0.596  0.55145
gt064        3.818e-01  2.452e-01   1.557  0.12031
gt065        1.134e-01  2.573e-01   0.441  0.65969
gt066        5.286e-01  2.646e-01   1.997  0.04654 *
gt067        6.330e-01  3.428e-01   1.847  0.06560 .
gt068        4.851e-01  3.487e-01   1.391  0.16501
gt069        6.609e-01  2.510e-01   2.634  0.00881 **
gt0610       8.010e-01  2.538e-01   3.156  0.00173 **
gt0611      -6.836e-01  2.885e-01  -2.370  0.01832 *
geom10       3.702e-01  1.756e-01   2.109  0.03566 *
geom12       5.429e-01  1.341e-01   4.048 6.31e-05 ***
geom13       5.207e-01  2.019e-01   2.579  0.01029 *
geom14       2.645e-01  1.311e-01   2.018  0.04430 *
geom16       3.423e-01  1.571e-01   2.179  0.02999 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4162 on 361 degrees of freedom
Multiple R-squared:  0.4515,  Adjusted R-squared:  0.4257
F-statistic: 17.48 on 17 and 361 DF,  p-value: < 2.2e-16
```

CENTRE.2009

```
Call:
lm(formula = log10no3 ~ om10 + stone5 + stone6 + stone8 + gt06 +
    bbg06 + laf + vds, data = centre.2009)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76840 -0.18852 -0.04166  0.14811  1.31326

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.603880   0.227213   2.658 0.008261 **
om10        -0.012154   0.005883  -2.066 0.039655 *
stone5       0.002947   0.001387   2.124 0.034443 *
stone6      -0.009424   0.002066  -4.561 7.27e-06 ***
stone8       0.006884   0.001621   4.247 2.85e-05 ***
gt061        0.221024   0.258227   0.856 0.392677
gt062        0.313407   0.216952   1.445 0.149552
gt063       -0.019667   0.423542  -0.046 0.962992
gt064        0.416132   0.212445   1.959 0.051007 .
gt065        0.021447   0.219891   0.098 0.922362
gt066        0.154510   0.221491   0.698 0.485940
gt067        0.323689   0.330120   0.981 0.327571
gt068        1.160130   0.407156   2.849 0.004665 **
gt069        0.615659   0.212147   2.902 0.003965 **
gt0610       0.844237   0.212324   3.976 8.66e-05 ***
gt0611       0.790462   0.242497   3.260 0.001235 **
bbg0660     -0.782863   0.148029  -5.289 2.29e-07 ***
laf3        -0.085087   0.054281  -1.568 0.117980
laf4        -0.090790   0.077314  -1.174 0.241152
laf5        -0.168015   0.087046  -1.930 0.054467 .
laf6         0.495077   0.126921   3.901 0.000117 ***
laf7         0.362032   0.105376   3.436 0.000669 ***
laf8         0.078562   0.085157   0.923 0.356936
laf9         0.130694   0.227149   0.575 0.565448
vds8        -0.317481   0.112320  -2.827 0.005001 **
vds10       -0.118286   0.123613  -0.957 0.339339
vds13        0.107419   0.097660   1.100 0.272186
vds14        0.067884   0.136364   0.498 0.618958
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3507 on 320 degrees of freedom
Multiple R-squared:  0.5292,  Adjusted R-squared:  0.4894
F-statistic: 13.32 on 27 and 320 DF,  p-value: < 2.2e-16
```

SOUTH.2007

```
Call:
  lm(formula = log10no3 ~ om05 + om40 + om60 + nhx + gt06 + ahn +
       bbg06 + kwel2 + lgn6 + geom, data = south.2007)

Residuals:
  Min       1Q   Median       3Q      Max
-1.62829 -0.33387  0.03042  0.34728  1.94725

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.300e+00  2.206e-01   5.892 4.65e-09 ***
om05         2.158e-02  6.583e-03   3.278 0.001068 **
om40         5.631e-03  2.561e-03   2.199 0.028042 *
om60        -7.143e-03  2.655e-03  -2.690 0.007217 **
nhx          1.123e-04  2.700e-05   4.161 3.34e-05 ***
gt062       -5.609e-01  1.807e-01  -3.104 0.001945 **
gt064       -4.051e-01  1.497e-01  -2.706 0.006876 **
gt065       -5.317e-01  1.539e-01  -3.456 0.000564 ***
gt066       -1.871e-01  1.617e-01  -1.157 0.247308
gt067       -1.389e-01  1.466e-01  -0.947 0.343541
gt068       -2.108e-01  1.498e-01  -1.407 0.159507
gt069        6.071e-04  1.469e-01   0.004 0.996703
gt0610       1.131e-01  1.503e-01   0.753 0.451821
gt0611      -9.241e-02  1.672e-01  -0.553 0.580504
ahn          5.301e-05  1.666e-05   3.183 0.001489 **
bbg0660     -9.204e-01  2.094e-01  -4.396 1.18e-05 ***
kwel2        2.819e-02  6.949e-03   4.056 5.23e-05 ***
lgn62        1.238e-01  3.549e-02   3.487 0.000501 ***
lgn63        5.797e-03  5.269e-02   0.110 0.912405
lgn64        2.565e-01  6.893e-02   3.721 0.000205 ***
lgn65        1.112e-01  4.991e-02   2.229 0.025987 *
lgn66        1.950e-01  4.744e-02   4.110 4.17e-05 ***
lgn610       4.223e-01  1.266e-01   3.336 0.000870 ***
lgn611      -5.440e-01  2.061e-01  -2.640 0.008382 **
lgn612      -5.598e-01  2.158e-01  -2.594 0.009585 **
lgn626      -3.242e-01  3.537e-01  -0.917 0.359382
lgn645      -3.003e-01  1.472e-01  -2.040 0.041547 *
lgn661       4.575e-01  1.888e-01   2.423 0.015511 *
geom7        3.879e-01  1.884e-01   2.059 0.039695 *
geom8        1.273e-01  1.841e-01   0.692 0.489288
geom9       -4.268e-01  5.228e-01  -0.816 0.414397
geom10       2.129e-01  1.976e-01   1.078 0.281333
geom12       3.065e-01  1.826e-01   1.678 0.093484 .
geom13       2.144e-01  1.795e-01   1.194 0.232500
geom14       1.477e-01  1.784e-01   0.828 0.407649
geom15       1.684e-01  1.966e-01   0.857 0.391817
geom16       8.403e-03  1.799e-01   0.047 0.962759
geom22       2.666e-01  5.207e-01   0.512 0.608683
---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4899 on 1565 degrees of freedom
Multiple R-squared:  0.4933,  Adjusted R-squared:  0.4814
F-statistic: 41.19 on 37 and 1565 DF,  p-value: < 2.2e-16
```

SOUTH.2008

```
Call:
lm(formula = log10no3 ~ om05 + om10 + om40 + om60 + stone5 +
    stone6 + stone7 + nhx + gt06 + bbg06 + kwel2 + lgn6 + geom +
    draf + slaf, data = south.2008)
Residuals:
     Min       1Q   Median       3Q      Max
-1.55062 -0.29139  0.05603  0.32940  1.29702
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.768e-01  5.560e-01   1.757 0.079083 .
om05        -4.875e-02  1.248e-02  -3.907 9.59e-05 ***
om10         5.518e-02  1.110e-02   4.970 7.14e-07 ***
om40         5.372e-03  2.665e-03   2.016 0.043930 *
om60        -5.501e-03  2.498e-03  -2.202 0.027731 *
stone5       1.963e-03  7.607e-04   2.581 0.009906 **
stone6       1.636e-03  8.163e-04   2.005 0.045110 *
stone7      -3.899e-03  1.143e-03  -3.411 0.000657 ***
nhx          4.167e-05  1.733e-05   2.405 0.016251 *
gt062       -5.104e-01  1.709e-01  -2.987 0.002847 **
gt064       -3.262e-01  1.472e-01  -2.216 0.026790 *
gt065       -4.228e-01  1.521e-01  -2.779 0.005487 **
gt066       -1.084e-01  1.610e-01  -0.673 0.500715
gt067       -1.186e-01  1.467e-01  -0.809 0.418778
gt068       -1.112e-01  1.475e-01  -0.754 0.450957
gt069        1.487e-02  1.460e-01   0.102 0.918914
gt0610       2.243e-01  1.473e-01   1.523 0.127911
gt0611      -7.432e-02  1.603e-01  -0.464 0.642973
bbg0651      1.298e-01  1.490e-01   0.871 0.383902
bbg0660     -1.139e+00  1.869e-01  -6.096 1.26e-09 ***
bbg0661     -1.092e+00  3.645e-01  -2.996 0.002759 **
bbg0662     -7.729e-01  4.087e-01  -1.891 0.058743 .
kwel2        2.995e-02  6.066e-03   4.938 8.41e-07 ***
lgn62        3.247e-01  2.911e-02  11.153  < 2e-16 ***
lgn63        2.356e-01  4.791e-02   4.919 9.28e-07 ***
lgn64        1.929e-01  5.971e-02   3.231 0.001249 **
lgn65        2.928e-01  4.223e-02   6.935 5.18e-12 ***
lgn66        4.442e-01  3.035e-02  14.636  < 2e-16 ***
lgn610       6.815e-01  8.452e-02   8.063 1.15e-15 ***
lgn611       1.558e-01  1.403e-01   1.110 0.267090
lgn612       7.661e-02  1.469e-01   0.521 0.602088
lgn625      -5.450e-01  5.064e-01  -1.076 0.281943
lgn626      -5.155e-01  4.888e-01  -1.055 0.291689
lgn635       1.800e-01  5.919e-01   0.304 0.761096
lgn636      -1.751e-01  3.696e-01  -0.474 0.635733
lgn637      -5.894e-01  4.131e-01  -1.427 0.153698
lgn638      -5.570e-01  3.545e-01  -1.571 0.116277
lgn639      -5.075e-01  3.610e-01  -1.406 0.159924
lgn640      -5.191e-01  4.002e-01  -1.297 0.194747
lgn645       1.053e-01  1.443e-01   0.730 0.465681
lgn661       7.344e-01  1.856e-01   3.957 7.79e-05 ***
geom7        9.797e-01  2.070e-01   4.733 2.33e-06 ***
geom8        6.534e-01  2.014e-01   3.244 0.001193 **
geom10       9.110e-01  2.105e-01   4.328 1.57e-05 ***
geom12       7.509e-01  2.015e-01   3.727 0.000198 ***
geom13       7.622e-01  1.998e-01   3.815 0.000140 ***
geom14       6.692e-01  1.999e-01   3.348 0.000825 ***
geom15       3.630e-01  2.173e-01   1.670 0.095011 .
geom16       5.778e-01  2.005e-01   2.882 0.003990 **
geom22       6.777e-01  2.699e-01   2.511 0.012108 *
draf4       -1.066e-01  1.404e-01  -0.759 0.447870
draf5       -2.370e-01  1.521e-01  -1.558 0.119300
draf6       -1.235e-01  1.565e-01  -0.789 0.429966
draf7       -3.678e-01  1.820e-01  -2.021 0.043434 *
slaf3        1.672e-01  5.101e-01   0.328 0.743075
slaf4        4.213e-02  4.993e-01   0.084 0.932763
slaf5       -3.853e-01  4.965e-01  -0.776 0.437765
```

```
slaf6       -7.374e-02  5.028e-01  -0.147 0.883430
slaf7       -1.422e-02  5.046e-01  -0.028 0.977522
slaf8        6.869e-02  5.056e-01   0.136 0.891948
slaf9       -1.414e-01  5.070e-01  -0.279 0.780266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4803 on 2462 degrees of freedom
Multiple R-squared:  0.5935,  Adjusted R-squared:  0.5835
F-statistic:  59.9 on 60 and 2462 DF,  p-value: < 2.2e-16
```

SOUTH.2009

```
Call:
lm(formula = log10no3 ~ om05 + om10 + om25 + om40 + om60 + nhx +
    gt06 + kwel2 + pawn + lgn6 + geom + slaf + slont, data = south.2009)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81319 -0.37211  0.06602  0.37878  1.65725

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.960e+00  8.531e-01   2.297 0.021707 *
om05        -1.594e-01  3.381e-02  -4.713 2.60e-06 ***
om10         2.057e-01  3.400e-02   6.051 1.71e-09 ***
om25        -1.299e-02  5.749e-03  -2.260 0.023950 *
om40         1.312e-02  5.166e-03   2.539 0.011180 *
om60        -9.259e-03  3.501e-03  -2.644 0.008243 **
nhx          5.702e-05  2.114e-05   2.696 0.007066 **
gt062       -5.591e-01  2.282e-01  -2.450 0.014363 *
gt064       -5.675e-01  1.914e-01  -2.964 0.003070 **
gt065       -6.578e-01  1.982e-01  -3.319 0.000918 ***
gt066       -5.034e-01  2.028e-01  -2.483 0.013110 *
gt067       -3.815e-01  1.895e-01  -2.014 0.044144 *
gt068       -3.895e-01  1.910e-01  -2.040 0.041514 *
gt069       -2.112e-01  1.882e-01  -1.123 0.261770
gt0610       6.647e-02  1.911e-01   0.348 0.727979
gt0611      -4.158e-01  2.106e-01  -1.975 0.048454 *
kwel2        4.385e-02  7.742e-03   5.664 1.69e-08 ***
pawn3        1.349e+00  6.537e-01   2.063 0.039201 *
pawn4       -8.400e-01  4.872e-01  -1.724 0.084838 .
pawn5        1.917e-01  3.431e-01   0.559 0.576399
pawn7        3.041e-01  3.499e-01   0.869 0.384822
pawn8        2.902e-01  3.241e-01   0.895 0.370634
pawn9       -1.700e-01  2.961e-01  -0.574 0.566019
pawn10      -1.512e-01  2.992e-01  -0.505 0.613354
pawn11      -2.152e-01  3.022e-01  -0.712 0.476353
pawn12      -2.492e-02  2.986e-01  -0.083 0.933513
pawn13      -1.247e-01  3.005e-01  -0.415 0.678172
pawn14      -3.219e-01  3.176e-01  -1.014 0.310891
pawn15       2.374e-02  3.309e-01   0.072 0.942810
pawn16      -6.325e-01  3.388e-01  -1.867 0.062055 .
pawn18      -5.456e-01  3.461e-01  -1.577 0.115015
pawn19      -4.888e-01  3.095e-01  -1.579 0.114412
pawn20      -4.657e-01  3.455e-01  -1.348 0.177900
lgn62        2.844e-01  3.301e-02   8.614  < 2e-16 ***
lgn63        1.945e-01  5.611e-02   3.466 0.000538 ***
lgn64        3.049e-01  6.275e-02   4.858 1.27e-06 ***
lgn65        2.550e-01  4.884e-02   5.222 1.95e-07 ***
lgn66        3.518e-01  3.661e-02   9.610  < 2e-16 ***
lgn610       5.218e-01  1.080e-01   4.833 1.44e-06 ***
lgn611      -5.298e-01  1.466e-01  -3.615 0.000307 ***
lgn612      -8.339e-01  1.728e-01  -4.825 1.50e-06 ***
lgn626       5.883e-02  2.679e-01   0.220 0.826235
lgn645      -8.793e-02  2.035e-01  -0.432 0.665796
geom6        1.118e-02  5.976e-01   0.019 0.985079
geom7        3.227e-01  5.365e-01   0.601 0.547603
```

```
geom8        4.071e-01  5.343e-01   0.762 0.446256
geom10       3.598e-01  5.393e-01   0.667 0.504726
geom11       3.049e-01  7.535e-01   0.405 0.685810
geom12       3.353e-01  5.337e-01   0.628 0.529883
geom13       3.184e-01  5.333e-01   0.597 0.550601
geom14       2.519e-01  5.343e-01   0.472 0.637310
geom15      -3.295e-02  5.424e-01  -0.061 0.951571
geom16       1.058e-01  5.346e-01   0.198 0.843166
geom22       1.248e-01  5.733e-01   0.218 0.827741
slaf3       -9.481e-02  5.942e-01  -0.160 0.873252
slaf4        1.981e-02  5.510e-01   0.036 0.971321
slaf5       -6.497e-01  5.405e-01  -1.202 0.229487
slaf6       -4.352e-01  5.371e-01  -0.810 0.417871
slaf7       -4.305e-01  5.364e-01  -0.802 0.422384
slaf8       -3.233e-01  5.366e-01  -0.602 0.546912
slaf9       -4.891e-01  5.369e-01  -0.911 0.362465
slont3      -1.237e-01  8.354e-02  -1.481 0.138832
slont4      -2.580e-01  7.793e-02  -3.310 0.000948 ***
slont5      -1.479e-01  7.846e-02  -1.885 0.059557 .
slont6      -8.793e-02  7.798e-02  -1.128 0.259576
slont7      -9.049e-02  8.229e-02  -1.100 0.271617
slont8      -3.143e-02  9.076e-02  -0.346 0.729144
slont9      -1.317e-01  9.283e-02  -1.418 0.156203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5304 on 2069 degrees of freedom
Multiple R-squared:  0.4171,  Adjusted R-squared:  0.3983
F-statistic:  22.1 on 67 and 2069 DF,  p-value: < 2.2e-16
```

ALL.REGIONS.2007

```
Call:
lm(formula = log10no3 ~ om120 + stone5 + stone6 + nhx + gt06 +
    ahn + bbg06 + gronds + kwel2 + pawn + lgn6 + geom + dront +
    laf, data = all.regions.2007)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56249 -0.33239  0.00063  0.32595  1.75183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.372e-01  5.532e-01   1.333 0.182682
om120        1.823e-03  8.363e-04   2.180 0.029272 *
stone5       6.272e-04  3.134e-04   2.001 0.045447 *
stone6      -6.623e-04  3.290e-04  -2.013 0.044128 *
nhx          6.848e-05  1.801e-05   3.803 0.000144 ***
gt061       -1.878e-01  1.857e-01  -1.011 0.312011
gt062       -2.969e-01  9.876e-02  -3.006 0.002657 **
gt063       -1.436e-01  1.378e-01  -1.042 0.297626
gt064       -2.235e-01  9.364e-02  -2.386 0.017047 *
gt065       -2.709e-01  9.392e-02  -2.884 0.003939 **
gt066       -3.946e-02  9.568e-02  -0.412 0.680040
gt067       -4.527e-02  9.329e-02  -0.485 0.627482
gt068        6.865e-03  9.380e-02   0.073 0.941658
gt069        8.253e-02  9.220e-02   0.895 0.370738
gt0610       1.051e-01  9.374e-02   1.121 0.262250
gt0611       7.966e-02  1.018e-01   0.783 0.433922
ahn          3.025e-05  9.772e-06   3.096 0.001975 **
bbg0611      3.725e-01  5.001e-01   0.745 0.456377
bbg0612     -6.008e-01  5.881e-01  -1.022 0.307055
bbg0634     -5.642e-01  7.566e-01  -0.746 0.455873
bbg0651      1.353e-01  4.820e-01   0.281 0.778901
bbg0660     -1.056e+00  4.928e-01  -2.143 0.032165 *
bbg0661     -1.350e+00  5.465e-01  -2.471 0.013510 *
bbg0662     -1.086e+00  5.611e-01  -1.936 0.052924 .
bbg0678     -1.487e-01  6.240e-01  -0.238 0.811628
gronds20     2.342e-01  4.302e-02   5.445 5.42e-08 ***
gronds21     1.851e-01  5.501e-02   3.365 0.000770 ***
gronds30     2.391e-01  1.081e-01   2.213 0.026974 *
gronds40     1.385e-01  1.211e-01   1.144 0.252830
gronds50     1.245e-01  1.282e-01   0.971 0.331724
gronds60    -1.070e-02  1.326e-01  -0.081 0.935684
gronds70    -3.704e-01  4.939e-01  -0.750 0.453279
kwel2        2.014e-02  3.222e-03   6.250 4.44e-10 ***
pawn2        2.488e-01  1.090e-01   2.282 0.022537 *
pawn3        1.402e-01  1.103e-01   1.272 0.203586
pawn4        8.340e-02  1.515e-01   0.551 0.581929
pawn5        1.293e-01  1.031e-01   1.254 0.210060
pawn6       -8.700e-02  2.910e-01  -0.299 0.765018
pawn7        1.518e-01  1.285e-01   1.181 0.237518
pawn8        1.108e-01  1.190e-01   0.932 0.351580
pawn9        2.781e-01  1.052e-01   2.643 0.008247 **
pawn10       2.408e-01  1.149e-01   2.096 0.036110 *
pawn11       8.933e-02  1.073e-01   0.832 0.405177
pawn12       3.921e-01  1.077e-01   3.639 0.000276 ***
pawn13       2.551e-01  1.055e-01   2.417 0.015664 *
pawn14       2.997e-01  1.233e-01   2.431 0.015099 *
pawn15       1.696e-02  1.656e-01   0.102 0.918404
pawn16      -8.607e-02  1.614e-01  -0.533 0.593950
pawn17       1.361e-01  1.686e-01   0.807 0.419478
pawn18      -1.968e-02  1.581e-01  -0.125 0.900911
pawn19       2.450e-01  1.477e-01   1.659 0.097241 .
pawn20       3.283e-01  1.589e-01   2.067 0.038822 *
pawn21       6.841e-01  4.872e-01   1.404 0.160311
```

```
lgn62        1.599e-01  2.036e-02   7.853 4.87e-15 ***
lgn63        2.699e-01  2.687e-02  10.045  < 2e-16 ***
lgn64        3.206e-01  3.986e-02   8.042 1.08e-15 ***
lgn65        2.408e-01  2.770e-02   8.694  < 2e-16 ***
lgn66        2.524e-01  3.800e-02   6.642 3.41e-11 ***
lgn610       3.797e-01  1.034e-01   3.672 0.000243 ***
lgn611       1.121e-01  9.970e-02   1.125 0.260777
lgn612      -5.974e-02  1.086e-01  -0.550 0.582349
lgn616       1.608e-01  1.968e-01   0.817 0.413757
lgn623       2.031e-01  3.382e-01   0.601 0.548192
lgn625       2.715e-01  3.370e-01   0.806 0.420533
lgn626       1.307e-01  1.164e-01   1.123 0.261488
lgn636       1.770e-01  2.534e-01   0.699 0.484852
lgn637       1.688e-01  3.117e-01   0.541 0.588203
lgn638       2.651e-04  3.959e-01   0.001 0.999466
lgn641       5.409e-02  3.421e-01   0.158 0.874369
lgn645      -1.375e-01  6.557e-02  -2.096 0.036095 *
lgn661       3.220e-01  1.699e-01   1.895 0.058117 .
geom2       -5.294e-01  3.496e-01  -1.514 0.130001
geom3       -7.964e-02  1.470e-01  -0.542 0.588023
geom6       -1.704e-01  1.501e-01  -1.136 0.256111
geom8       -1.296e-01  9.791e-02  -1.324 0.185530
geom9       -6.710e-03  1.664e-01  -0.040 0.967832
geom10      -1.721e-01  1.045e-01  -1.647 0.099694 .
geom11       1.606e-02  4.864e-01   0.033 0.973658
geom12       6.553e-03  9.223e-02   0.071 0.943358
geom13      -7.863e-02  8.973e-02  -0.876 0.380924
geom14      -1.341e-01  9.074e-02  -1.478 0.139479
geom15      -1.148e-01  1.069e-01  -1.074 0.282892
geom16      -2.145e-01  9.177e-02  -2.337 0.019466 *
geom22      -7.972e-02  2.897e-01  -0.275 0.783216
geom4       -1.569e-01  2.922e-01  -0.537 0.591408
geom7        8.968e-02  1.068e-01   0.840 0.401045
dront3       1.413e-01  4.292e-02   3.293 0.000998 ***
dront4       2.342e-01  4.437e-02   5.279 1.35e-07 ***
dront5       2.241e-01  4.730e-02   4.739 2.21e-06 ***
dront6       2.900e-01  4.993e-02   5.810 6.63e-09 ***
dront7       2.592e-01  8.792e-02   2.948 0.003213 **
laf2         1.540e-02  2.140e-01   0.072 0.942659
laf3         5.307e-02  2.139e-01   0.248 0.804029
laf4         3.024e-02  2.143e-01   0.141 0.887763
laf5         2.100e-03  2.152e-01   0.010 0.992214
laf6        -1.417e-01  2.167e-01  -0.654 0.513332
laf7        -1.466e-01  2.161e-01  -0.678 0.497702
laf8        -2.095e-01  2.176e-01  -0.963 0.335823
laf9        -2.432e-01  2.288e-01  -1.063 0.287942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4749 on 5284 degrees of freedom
Multiple R-squared:  0.4718,  Adjusted R-squared:  0.462
F-statistic: 48.16 on 98 and 5284 DF,  p-value: < 2.2e-16
```

ALL.REGIONS.2008

```
Call:
   lm(formula = log10no3 ~ om60 + om80 + om100 + stone5 + stone6 +
        stone7 + stone8 + nhx + gt06 + ahn + bbg06 + kwel2 + pawn +
        lgn6 + geom + slaf + slont, data = all.regions.2008)

 Residuals:
   Min       1Q   Median       3Q      Max
 -1.58346 -0.32860 -0.00323  0.32693  1.81543

 Coefficients:
   Estimate Std. Error t value Pr(>|t|)
 (Intercept)  1.496e+00  4.192e-01   3.568 0.000362 ***
 om60        -1.485e-03  4.613e-04  -3.220 0.001289 **
 om80         2.175e-03  9.668e-04   2.250 0.024487 *
 om100       -2.487e-03  1.162e-03  -2.140 0.032420 *
 stone5       7.671e-04  3.118e-04   2.461 0.013896 *
 stone6      -1.423e-03  4.771e-04  -2.982 0.002875 **
 stone7       2.051e-03  7.866e-04   2.607 0.009160 **
 stone8      -1.409e-03  6.730e-04  -2.094 0.036287 *
 nhx          3.990e-05  1.430e-05   2.789 0.005303 **
 gt061       -1.711e-01  1.518e-01  -1.127 0.259729
 gt062       -1.701e-01  7.179e-02  -2.369 0.017867 *
 gt063       -2.402e-01  1.587e-01  -1.513 0.130295
 gt064       -1.251e-01  6.952e-02  -1.800 0.071920 .
 gt065       -2.433e-01  6.980e-02  -3.485 0.000495 ***
 gt066       -9.843e-02  7.327e-02  -1.343 0.179191
 gt067       -3.111e-02  6.955e-02  -0.447 0.654657
 gt068        5.808e-02  7.013e-02   0.828 0.407616
 gt069        8.323e-02  6.825e-02   1.219 0.222708
 gt0610       1.283e-01  6.974e-02   1.840 0.065885 .
 gt0611      -5.320e-02  7.989e-02  -0.666 0.505508
 ahn          3.425e-05  8.898e-06   3.849 0.000120 ***
 bbg0611     -3.497e-01  3.680e-01  -0.950 0.341999
 bbg0651     -3.555e-01  3.666e-01  -0.970 0.332185
 bbg0660     -1.536e+00  3.775e-01  -4.067 4.82e-05 ***
 bbg0661     -1.610e+00  4.350e-01  -3.701 0.000217 ***
 bbg0662     -1.492e+00  4.496e-01  -3.318 0.000912 ***
 kwel2        1.572e-02  2.924e-03   5.378 7.79e-08 ***
 pawn2       -1.354e-01  9.133e-02  -1.483 0.138228
 pawn3       -3.128e-01  9.595e-02  -3.260 0.001119 **
 pawn4       -2.224e-01  1.199e-01  -1.855 0.063578 .
 pawn5       -1.314e-01  8.527e-02  -1.540 0.123501
 pawn6       -3.907e-01  1.841e-01  -2.122 0.033846 *
 pawn7       -8.391e-02  1.115e-01  -0.752 0.451883
 pawn8        4.661e-02  9.591e-02   0.486 0.627019
 pawn9        9.013e-03  8.786e-02   0.103 0.918293
 pawn10      -6.640e-02  9.471e-02  -0.701 0.483231
 pawn11      -2.310e-01  8.985e-02  -2.571 0.010177 *
 pawn12       1.738e-01  8.868e-02   1.960 0.050046 .
 pawn13      -1.304e-02  8.841e-02  -0.147 0.882756
 pawn14      -6.954e-02  1.062e-01  -0.655 0.512788
 pawn15      -1.304e-01  1.217e-01  -1.071 0.284144
 pawn16      -3.139e-01  1.003e-01  -3.130 0.001757 **
 pawn17      -4.867e-01  1.017e-01  -4.784 1.76e-06 ***
 pawn18      -2.292e-01  8.891e-02  -2.578 0.009952 **
 pawn19      -1.056e-01  9.233e-02  -1.144 0.252642
 pawn20      -1.738e-01  1.136e-01  -1.530 0.126057
 lgn62        3.281e-01  1.917e-02  17.117  < 2e-16 ***
 lgn63        3.823e-01  2.644e-02  14.458  < 2e-16 ***
 lgn64        2.845e-01  3.832e-02   7.422 1.31e-13 ***
 lgn65        3.259e-01  2.557e-02  12.746  < 2e-16 ***
 lgn66        4.527e-01  2.419e-02  18.715  < 2e-16 ***
 lgn610       6.913e-01  7.647e-02   9.041  < 2e-16 ***
 lgn611       1.787e-01  9.465e-02   1.888 0.059067 .
```

```
lgn612       8.897e-02  1.021e-01   0.871 0.383787
lgn625      -5.876e-02  2.682e-01  -0.219 0.826591
lgn626       4.415e-01  1.962e-01   2.250 0.024458 *
lgn635       8.435e-02  5.361e-01   0.157 0.875000
lgn636      -2.226e-02  2.403e-01  -0.093 0.926177
lgn637      -1.264e-02  2.694e-01  -0.047 0.962583
lgn638       2.310e-02  2.494e-01   0.093 0.926197
lgn639      -3.756e-01  2.622e-01  -1.433 0.151965
lgn640      -5.588e-01  3.692e-01  -1.514 0.130191
lgn641      -1.927e-01  3.460e-01  -0.557 0.577546
lgn642       5.613e-01  4.794e-01   1.171 0.241712
lgn645      -5.721e-03  6.369e-02  -0.090 0.928429
lgn661       6.083e-01  1.706e-01   3.565 0.000367 ***
geom2       -2.374e-01  2.822e-01  -0.841 0.400229
geom6       -6.063e-01  2.338e-01  -2.593 0.009540 **
geom8       -1.930e-01  1.492e-01  -1.293 0.196055
geom9       -5.066e-01  2.081e-01  -2.434 0.014970 *
geom10      -1.416e-01  1.542e-01  -0.918 0.358560
geom11      -6.809e-02  5.021e-01  -0.136 0.892143
geom12      -6.896e-02  1.478e-01  -0.466 0.640893
geom13      -1.040e-01  1.470e-01  -0.707 0.479392
geom14      -2.064e-01  1.472e-01  -1.402 0.160906
geom15      -2.370e-01  1.597e-01  -1.484 0.137785
geom16      -2.294e-01  1.479e-01  -1.551 0.120954
geom22      -1.861e-01  2.348e-01  -0.792 0.428206
geom7        1.367e-01  1.562e-01   0.875 0.381420
slaf2        5.858e-02  1.034e-01   0.567 0.570963
slaf3        4.660e-02  1.002e-01   0.465 0.642043
slaf4        1.732e-01  9.529e-02   1.818 0.069104 .
slaf5        1.820e-01  9.513e-02   1.913 0.055738 .
slaf6        2.459e-01  9.350e-02   2.630 0.008560 **
slaf7        2.766e-01  9.370e-02   2.952 0.003168 **
slaf8        3.209e-01  9.411e-02   3.410 0.000654 ***
slaf9        2.389e-01  9.488e-02   2.518 0.011816 *
slont2      -3.557e-02  6.807e-02  -0.523 0.601301
slont3       5.570e-02  6.708e-02   0.830 0.406367
slont4       6.044e-02  6.673e-02   0.906 0.365090
slont5       1.485e-01  6.753e-02   2.200 0.027874 *
slont6       1.147e-01  6.762e-02   1.696 0.089946 .
slont7       1.349e-01  6.847e-02   1.971 0.048770 *
slont8       2.146e-01  7.017e-02   3.058 0.002240 **
slont9       1.217e-01  7.237e-02   1.682 0.092611 .
---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4778 on 6148 degrees of freedom
Multiple R-squared:  0.554,  Adjusted R-squared:  0.5472
F-statistic: 81.24 on 94 and 6148 DF,  p-value: < 2.2e-16
```

ALL.REGIONS.2009

```
Call:
  lm(formula = log10no3 ~ stone5 + stone6 + stone7 + stone8 + nhx +
        gt06 + ahn + kwel2 + pawn + lgn6 + geom + dront + slaf +
        vds, data = all.regions.2009)
Residuals:
   Min       1Q   Median       3Q      Max
-1.73807 -0.34174 -0.01295  0.34677  1.97089

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.335e-01  1.910e-01   4.888 1.05e-06 ***
stone5       7.556e-04  3.311e-04   2.282 0.022516 *
stone6      -1.070e-03  5.095e-04  -2.100 0.035785 *
stone7       1.975e-03  8.397e-04   2.352 0.018692 *
stone8      -1.605e-03  7.100e-04  -2.261 0.023822 *
nhx          5.657e-05  1.597e-05   3.541 0.000401 ***
gt061       -9.983e-03  1.862e-01  -0.054 0.957248
gt062       -1.449e-01  8.336e-02  -1.738 0.082180 .
gt063       -6.262e-02  1.376e-01  -0.455 0.648932
gt064       -1.545e-01  8.074e-02  -1.914 0.055705 .
gt065       -1.786e-01  8.033e-02  -2.223 0.026249 *
gt066       -1.093e-01  8.360e-02  -1.307 0.191284
gt067       -4.184e-02  8.065e-02  -0.519 0.603920
gt068        5.603e-03  8.104e-02   0.069 0.944880
gt069        9.295e-02  7.916e-02   1.174 0.240359
gt0610       1.500e-01  8.096e-02   1.853 0.063924 .
gt0611      -1.162e-01  9.484e-02  -1.225 0.220526
ahn          4.502e-05  9.869e-06   4.562 5.17e-06 ***
kwel2        1.902e-02  3.345e-03   5.688 1.35e-08 ***
pawn2       -6.526e-02  7.513e-02  -0.869 0.385097
pawn3       -1.377e-01  9.885e-02  -1.394 0.163518
pawn4       -2.162e-01  1.137e-01  -1.902 0.057280 .
pawn5       -5.768e-02  7.206e-02  -0.800 0.423505
pawn6       -4.344e-01  2.676e-01  -1.623 0.104614
pawn7        2.722e-01  1.492e-01   1.824 0.068168 .
pawn8        1.761e-01  1.076e-01   1.636 0.101822
pawn9        8.944e-02  9.358e-02   0.956 0.339268
pawn10       8.960e-02  1.027e-01   0.872 0.383233
pawn11      -5.871e-02  9.662e-02  -0.608 0.543453
pawn12       2.896e-01  9.581e-02   3.023 0.002515 **
pawn13       6.455e-02  9.433e-02   0.684 0.493775
pawn14      -1.973e-01  1.247e-01  -1.583 0.113549
pawn15      -1.159e-01  1.375e-01  -0.843 0.399131
pawn16      -3.411e-01  1.208e-01  -2.824 0.004754 **
pawn17      -2.406e-01  1.237e-01  -1.945 0.051875 .
pawn18      -1.130e-01  1.311e-01  -0.862 0.388749
pawn19       2.476e-02  1.007e-01   0.246 0.805700
pawn20      -1.744e-01  1.441e-01  -1.210 0.226227
lgn62        3.173e-01  2.003e-02  15.842  < 2e-16 ***
lgn63        3.123e-01  2.781e-02  11.229  < 2e-16 ***
lgn64        3.622e-01  3.857e-02   9.390  < 2e-16 ***
lgn65        2.829e-01  2.691e-02  10.513  < 2e-16 ***
lgn66        4.572e-01  2.759e-02  16.569  < 2e-16 ***
lgn610       6.615e-01  9.044e-02   7.314 2.96e-13 ***
lgn611      -3.919e-01  7.021e-02  -5.581 2.50e-08 ***
lgn612      -9.533e-01  9.314e-02 -10.235  < 2e-16 ***
lgn616       1.226e-02  2.051e-01   0.060 0.952325
lgn623       5.404e-01  4.992e-01   1.083 0.279012
lgn625       4.406e-02  2.891e-01   0.152 0.878874
lgn626       2.728e-01  1.449e-01   1.883 0.059798 .
lgn636      -1.182e+00  2.129e-01  -5.552 2.96e-08 ***
lgn637      -1.245e+00  3.600e-01  -3.459 0.000546 ***
lgn641       2.992e-02  2.912e-01   0.103 0.918164
lgn642      -1.666e-01  3.542e-01  -0.470 0.638028
```

```
lgn645      -3.256e-02  8.285e-02  -0.393 0.694306
lgn661       4.471e-01  1.596e-01   2.802 0.005094 **
geom2       -3.721e-01  2.857e-01  -1.302 0.192872
geom6       -2.143e-01  2.339e-01  -0.916 0.359584
geom8        5.814e-02  1.399e-01   0.416 0.677787
geom9       -2.101e-01  2.034e-01  -1.033 0.301547
geom10      -7.969e-02  1.464e-01  -0.544 0.586281
geom11       2.213e-01  3.784e-01   0.585 0.558649
geom12      -6.189e-03  1.382e-01  -0.045 0.964284
geom13      -4.092e-02  1.371e-01  -0.298 0.765438
geom14      -1.300e-01  1.374e-01  -0.946 0.344264
geom15      -2.279e-01  1.539e-01  -1.481 0.138683
geom16      -2.269e-01  1.383e-01  -1.641 0.100942
geom22      -2.259e-01  2.175e-01  -1.039 0.299050
geom7        6.599e-02  1.462e-01   0.451 0.651779
dront3      -1.146e-02  4.698e-02  -0.244 0.807309
dront4       5.964e-02  4.887e-02   1.220 0.222364
dront5       6.087e-02  5.158e-02   1.180 0.237973
dront6       1.239e-01  5.428e-02   2.283 0.022447 *
dront7      -7.144e-03  1.189e-01  -0.060 0.952092
slaf2        1.253e-01  1.056e-01   1.187 0.235321
slaf3        1.243e-01  1.057e-01   1.176 0.239541
slaf4        1.340e-01  1.001e-01   1.339 0.180689
slaf5        1.868e-01  1.002e-01   1.864 0.062405 .
slaf6        2.137e-01  9.859e-02   2.168 0.030235 *
slaf7        2.403e-01  9.900e-02   2.427 0.015257 *
slaf8        3.288e-01  9.926e-02   3.313 0.000929 ***
slaf9        2.381e-01  1.009e-01   2.360 0.018294 *
vds7        -6.745e-02  1.234e-01  -0.546 0.584790
vds8         4.527e-02  8.566e-02   0.528 0.597219
vds10       -1.044e-01  3.441e-02  -3.035 0.002414 **
vds11        5.907e-01  1.640e-01   3.602 0.000318 ***
vds13       -3.041e-02  5.873e-02  -0.518 0.604630
vds14       -2.143e-01  1.011e-01  -2.120 0.034050 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.498 on 5633 degrees of freedom
Multiple R-squared:  0.4349,  Adjusted R-squared:  0.4262
F-statistic: 49.84 on 87 and 5633 DF,  p-value: < 2.2e-16
```