
# Reverse modelling of yield-influencing soil variables in case of few soil data

István SISÁK[1*], András BENŐ[1], Kornél SZABÓ[2], Mihály KOCSIS[1], János ABONYI[3]

[1] University of Pannonia, Georgikon Faculty, 16 Deák F. st. H-8360 Keszthely, Hungary
[2] Dr. Szabó Agrochemical Co. Ltd., 36 Batthyany st., H-8790 Zalaszentgrót, Hungary
[3] University of Pannonia, Faculty of Engineering, 10 Egyetem st. H-8200 Veszprém, Hungary
[*] Corresponding author's e-mail address: sisak@georgikon.hu, Phone: +36 30 288 7775

## Abstract

Our hypothesis was that simple models can be applied to predict yield by using only those yield data which spatially coincide with the soil data and the remaining yield data and the models can be used to test different sampling and interpolation approaches commonly applied in precision agriculture and to better predict soil variables at not observed locations. Three strategies for composite sample collection were compared in our study. Point samples were taken 1.) along lines within homogenous NDVI areas, 2.) along lines within homogenous electric conductivity scan zones and 3.) in circles around predefined regular grid points. Multiple regression models were developed to predict yields. Digital elevation, five to eight soil variables and in one case three agrotechnical variables (variable rate fertilizer use and seeding) were retained in the prediction equations with $R^2$ values of 0.557, 0.248 and 0.191 for circular, soil EC based and NDVI based sampling, respectively. Spline interpolation method proved to be the best in two cases and IDW method was the best in the third case. The attempt to predict soil variables with fine spatial detail brought mixed results.

## Keywords

few composite soil samples, detailed yield map, simple local models, test of competing methods

## Introduction

Precision agriculture has been slowly gaining ground in Hungary for the last decade. However, different practices are not adopted at the same speed. Not independently from the strong promotion activity of machinery distributors, technical solutions coupled with specialized and expensive machines (sprayers, harvesters etc.) are more widespread. Yield monitors for example have been sold well even though they are not always used properly.

There is an established soil sampling system in Hungary which must be followed by the farmers who apply for agri-environmental subsidies from EU sources. One composite sample must be collected and get analyzed in an accredited laboratory from each 5 hectare units of land in every 5 year. However, it is not detailed enough for precision nutrient management neither in space nor in time. Further, only users of some 4-500 thousand hectares receive such subsidies and that is approximately 10 % of the agricultural land. Farmers who are not part of that scheme, are not obliged to adhere to the established soil sampling methodology and very often wet laboratory soil tests fall victim of cost saving despite declared goal of the farmers to practice precision agriculture. Sensor based methods have been suggested as cost saving alternatives for wet laboratory tests but available nutrient content in soil cannot be accurately derived from sensor data not to mention that these methods are still not widespread in our country.

There is a new development, too: a small but active group of agricultural service providers offer different precision farming methods mainly soil sampling and traditional wet laboratory soil tests based on different approaches for management unit delineation: 1) NDVI derived from aerial photographs or space images or 2) soil electric conductivity scanning with contact or electromagnetic equipment. Strange enough, in some cases the average size of the suggested management units is larger than 5 hectares.

As a result of the divergent processes, more and more yield maps are produced in Hungary but reliable soil data are missing to explain causal relationships between soil factors and yield or fine scale patterns within the land parcels.
Farmers will hopefully recognize that there is no way to achieve trustable information on profitability of the precision technology without reliable soil data. But in short term the situation is given and the discrepancy between few soil data and abundant yield data cannot be solved. New scientific methods should be developed to make use of the available information. Our hypothesis was that simple models can be applied to predict yield by using only those yield data which spatially coincide with the soil sampling points and the remaining yield data and the derived simple models should be used to test different sampling and interpolation approaches commonly applied in precision agriculture and to better predict soil variables at not observed locations.

**Material and methods**

Three fields under precision farming at different locations in Hungary were selected for the study a 33 hectare field at Zimány (site 1: 46.4391° 17.9079°) a 22 hectare field at Solt (site 2: 46.8109° 19.0142°) and a 153 hectare field at Békéssámson (site 3: 46.4551° 20.6664°). The approximate northern latitudes and eastern longitudes of the center points are given in the brackets (Figure 1). Site 1 is located in South-Transdanubia on hilly surface between elevations 148.4-171.0 meter above sea levels (m.a.s.l.), while site 2 and 3 are located on the Great Plain with elevation differences between 94.0-95.8 and 88.0-90.9. m.a.s.l., respectively. Soils of site 1 are mainly Typic Hapludalfs and their eroded varieties. Soils of site 2 and 3 are Mollic Natraqualfs, Typic Natraquolls and Aquic Calciudolls associations. Average annual precipitation and temperature for the three sites are 670-540-560 mm, and 10.1-10.5-10.6 °C, respectively with usually wet spring and autumn, with warm and dry summer and relatively dry and cold winter, thus it is a typical continental climate with small variations.

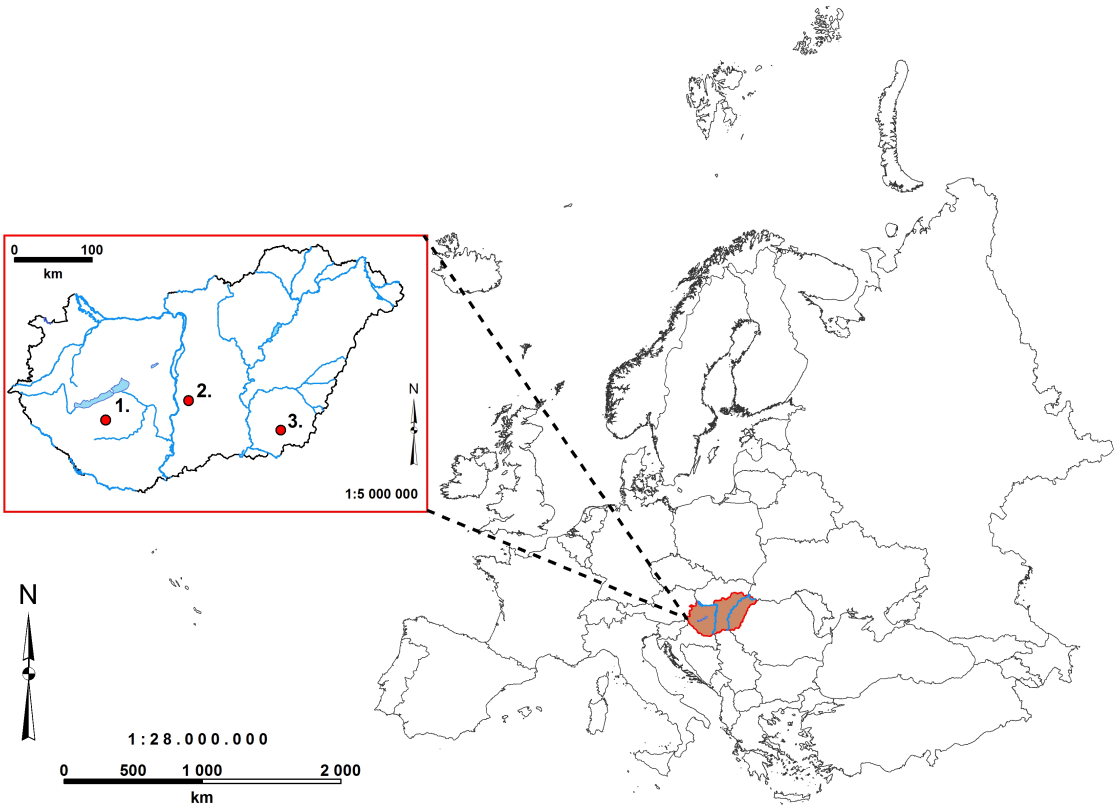


**Figure 1 The location of experimental sites 1 to 3 in Hungary**

Maize was grown at site 1 in 2012 when severe drought decimated the yield and also wild pigs damaged some patches. Also maize was grown at site 2 in 2016 which was a regular year and hard grain wheat was grown at site 3 in 2015. Variable rate fertilizer use and seeding rate were applied at site 2 but uniform soil management and cropping practices were applied at site 1 and 3. Yield maps were available for the three fields.

Three strategies for composite sample collection were applied in the study areas. Point samples were taken in circles with 30 m radius around predefined regular grid points at site 1, along lines within homogenous NDVI zones at site 2 and along lines within homogenous electric conductivity zones at site 3. The exact locations of sampling points to collect composite samples were known in all cases. Soil sampling was done 3 years before the year of investigation at site 1, in the previous year at site 2 and two years later at site 3. The soil data from 2017 can be considered valid for the yield in 2015 at site 3 because the farmer wanted to covert the field into biological farming and as a first step zero input soil management was introduced between 2015 and 2017. Uniform soil and nutrient management was applied at site 1 between soil sampling and the year of investigation (2009 and 2012). Homogenous NDVI zones were established upon aerial photographs of sunflower canopy in August 2013 at site 2. Variable rate fertilizer and seeding applications for maize in 2015 were based on yield variances of maize in 2014.

The above description well represents the real world situation of data analysis in precision farming. There are many variables and altering circumstances. It is difficult to find similar fields.

The results of the wet laboratory soil tests were assigned to the sampling points from where the composite samples were collected. Four different methods were used to interpolate soil data from observation points. The first one was simple pairing of soil data with the homogenous zones they represented. This method could not be used for site 1 because no zones were defined only 12 grid-like center points were set. The second method was simple kriging which also could not be applied for site 1 since we had only 12 grids.

This small number of points with uniform distribution is not recommended for kriging. The third method was IDW and the forth was spline interpolation which were used for all data.

Soil data were considered to coincide with yield data if yield monitor points were within a 30 m circles of grid center points at site 1, within 10 m circles at site 2 where too many sampling points were recorded and within 20 m circles at site 3. These points were called model development area.

Multiple linear regression models were fitted to predict yield by wet laboratory soil data and digital elevation data which were available from yield monitors and also from alternative sources. Variable fertilizer and seeding rates were also used as independent variables at site 2. Stepwise variable selection method was applied. The derived equations were used with interpolated soil data to predict yield for those points (test area) which were not included in the development of the model equation. $R^2$ values (variance explained by the model) was used to compare different interpolation methods to the original model and the interpolation methods to each other.

Yield residuals of test area were analyzed by kriging whether they still show spatial pattern. Existing spatial pattern would be proof for existence of unknown variables or for not explained variability of known variables for example because of insufficient representation of their spatial pattern. Rearranged model equations were used in the test area to predict fine scale spatial pattern of the most important soil variable from yield assuming that the other soil variables follow the distribution predicted by the best interpolation method.

**Results and discussion**

The average yield for maize was 4.27 $Mg \cdot ha^{-1}$ in 2012 at site 1 and 9.04 $Mg \cdot ha^{-1}$ in 2016 at site 2 and for wheat it was 5.40 $Mg \cdot ha^{-1}$ in 2015 at site 3. Multiple damages at site 1 took serious toll on yield. The linear predictors of yield in the regression equations in decreasing order of effect size for site 1: elevation, liquid limit according to Arany (LLA a common simple physical soil test in Hungary), nitrate content, plant available potassium (paK), sodium content, plant available phosphorus (paP), pH, magnesium and humus content; for site 2: humus content, ammonium nitrate fertilizer use, seeding rate, elevation, sulfate content, nitrate content, sodium content, manganese content and monoammonium-phosphate fertilizer use; for site 3 LLA, paK, paP, nitrate content, soluble salt content and elevation.

**Table 1 Model performances in model development and test areas with various interpolation methods**

| | site 1 | site 1 | site 2 | site 2 | site 3 | site 3 |
|---|---|---|---|---|---|---|
| | | % of model dev. area | | % of model dev. area | | % of model dev. area |
| N (model dev. area) | 1330 | | 3557 | | 5163 | |
| N (test area) | 11761 | | 17270 | | 44200 | |
| $R^2$ (model dev. area) | 0.557 | 100 | 0.191 | 100 | 0.248 | 100 |
| $R^2$ (test area - IDW) | 0.3624 | 65.1 | 0.1149 | 60.2 | 0.1827 | 73.7 |
| $R^2$ (test area - kriging) | | | 0.0858 | 44.9 | 0.1630 | 65.7 |
| $R^2$ (test area - spline) | 0.4292 | 77.1 | 0.1391 | 72.8 | 0.1600 | 64.5 |
| $R^2$ (test area - pairing) | | | 0.1163 | 60.9 | 0.1534 | 61.9 |

Model performances for the model development and test areas are shown in Table 1. The best $R^2$ value was produced at site 1 ($R^2$=0.557) where elevation had the strongest effect, which represents water availability in drought situation. Model performance at site 3 for wheat was weaker but here the soil data without fertilizer use still had medium strong effect ($R^2$=0.248). Model performance was the weakest at site 2 ($R^2$=0.191) where relatively high yield was achieved. This example was clearly at the plateou stage of the

yield curve where most of the influencing variables are at or near optimum level. Elevation was a medium strong factor here, which can be explained by its relationship with the depth of sodium rich subsurface layers that reduces yield at deep laying areas.

Spatial representation of composite samples were satisfactory for site 1 and 2 (approximately one composite sample for 3 hectare area, 12 and 7 composite samples respectively) but it was rather rough for site 3 (one sample for 14 hectare on the average, 11 composite samples). The best interpolation method for site 1 was spline function (77.1 % of the variance of the model development area) and this was the overall best, too. Also spline method performed best at site 2 (72.8 %) and IDW method was the best at site 3 (73.7 %). As expected, with relative small number of individual measurements (composite samples) simple methods perform better even if these values are distributed in several points. Despite unaccounted variables in the equation at site 1 (wild pig damage) spline function performed surprisingly well. This might be the consequence of the sampling scheme. Composite samples represented relatively small, compact areas (circles) with yield variances that are not due to soil factors. In contrast with that at site 2, composite samples were taken from multiple field polygons which had same NDVI values. That might be the reason why kriging so badly underperformed (only 44.9 % of the variance of the model development area). Simple pairing average values with the source field polygon cannot be recommended since performance of those models was weak. The relatively good performance of IDW interpolation for site 3 may be explained by the sparse distribution of points thus by low spatial representation.

The residual yield data in the test area (not explained by the best interpolation method) shown explicit spatial pattern. This pattern for site 2 can be seen on Figure 2. There is a pattern which follow the rows of harvesting and another pattern depicting elevation differences which was, however, not captured by the model despite inclusion of elevation among the significant independent variables.

The humus was the strongest soil variable in the equation for site 2, thus we rearranged the equation and used it to estimate humus content with help of estimated yield data and the other variables. The kriged result is shown on Figure 3. For comparison, Figure 4 shows the original estimated humus content from soil sampling points with spline interpolation. The range of data is wider for the original estimate, but the overall pattern is very similar. The narrower range is the methodological consequence of kriging. The second estimate of the humus content did not brought significant gain in explaining the spatial pattern of the organic matter in soil.

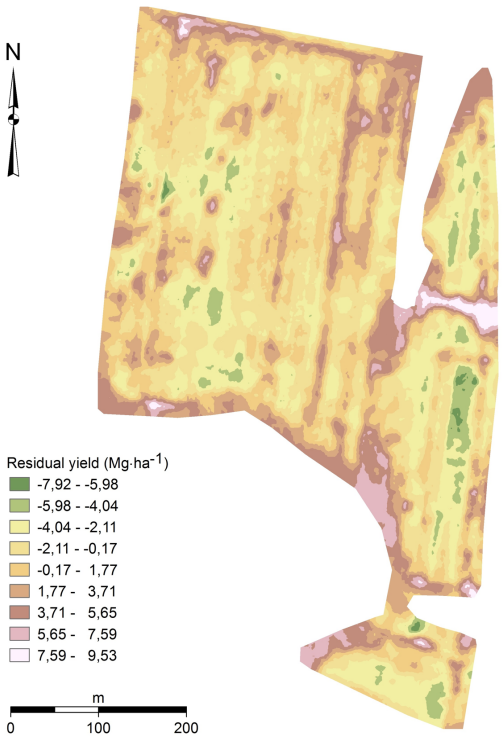


**Figure 2 Kriged residual yield data not explained by the regression equation for site 2**

In the case of the other two sites, similar results were obtained for the residual yield data but worse results were obtained for the strongest significant soil variable, the liquid limit according to Arany (the data are not shown).

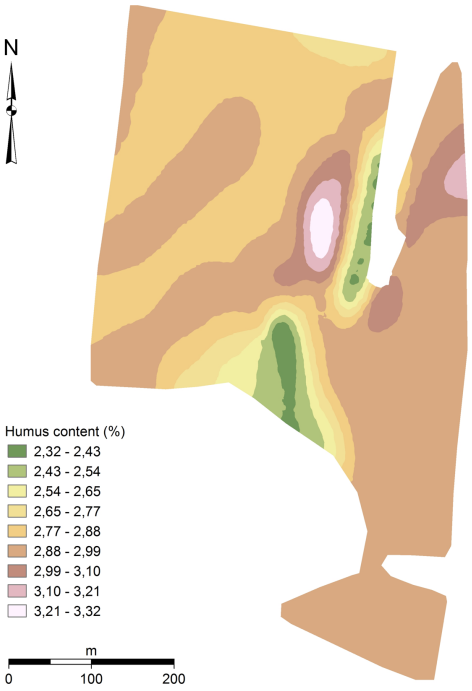


**Figure 3 Humus content estimate with the yield-soil data relationship**

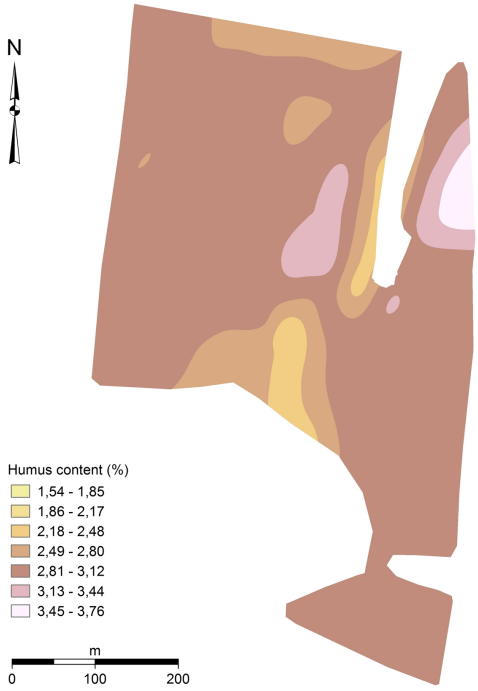


**Figure 4 Humus content estimate from the original sampling point data with spline interpolation**

## Conclusions

We have found that small local models perform well if the yield variance within a model development area is small and the yield variance between average soil samples is large which requirement was best satisfied with circular placement of point samples at site 1. Spline interpolation seems to be the best method in case of relatively few composite samples.

Further soil sampling strategies can be formulated as a conclusion of our study. The representative samples should be placed within the field by using soil related sensor measurements (such as EC). Future soil sample locations should partially coincide with the previous ones but other, previously not investigated locations should also be selected to test the performance of the model. This step by step knowledge acquiring approach may lead to a thorough understanding of local interactions of yield-influencing environmental variables which is the core of the precision farming.