Big data – Banana origin determination

M. Alewijn

WAGENINGEN
UNIVERSITY & RESEARCH

Big data – Banana origin determination

M. Alewijn

This research has been carried out by Wageningen Food Safety Research, institute within the legal entity Wageningen Research Foundation funded by the Dutch Ministry of Agriculture, Nature and Food Quality (project number KB-37-002-008.RIKILT).

Wageningen, April 2020

WFSR report 2020.006

WAGENINGEN
UNIVERSITY & RESEARCH

M. Alewijn, 2020. *Big data – Banana origin determination.* Wageningen, Wageningen Food Safety Research, WFSR report 2020.006. 30 pp.; 7 fig.; 4 tab.; 0 ref.

Project number: 1287363702
BAS-code: KB-37-002-008
Project title: KB big data
Project leader: S. van Ruth

This report can be downloaded for free at https://doi.org/10.18174/516096 or at www.wur.eu/food-safety-research (under WFSR publications).

P.O. Box 230, 6700 AE Wageningen, The Netherlands, T +31 (0)317 48 02 56, E info.wfsr@wur.nl, www.wur.eu/food-safety-research. WFSR is part of Wageningen University & Research.

WFSR report 2020.006

Contents

Summary

WFSR aims to develop a multivariate method based on instrumental analytical data that can support the origin claims of banana shipments. Such a method can reduce the possibilities for economically motivated fraud with these fruits. Regional variations like weather, soil characteristics and other influential factors, that will also change over time, are likely to influence the composition of these products subtly. If so, it could be possible to link product properties, i.e. banana composition to be analysed upon arrival, to searchable facts on the production location, such as soil profile and weather/ climate data. If these production location characteristics have a causal and predictive effect on the product properties, it will be possible to predict, and verify, the origin for all bananas being imported.

This report deals with the process of gathering, collating and processing the data from various sources, and the creation of predictive models related to origin of the product.

Soil and weather data were obtained from sources at WUR Soil Geography and Landscape, KNMI and SoilGrids.org. Analytical data were generated on bananas obtained from a set of 14 selected farms in Costa Rica and consisted of volatile profiles (proton transfer reaction mass spectroscopy), infrared spectroscopy, carotenoid profiles and element composition in peel and pulp.

These data were cleaned and collated and explored using univariate and multivariate techniques to correlate the results from each analytical technique to the available soil data. These correlations were present but were not considered strong enough for origin verification. Mid-level data fusion of all analytical parameters into several PLS models with leave-farm-out-cross validation to predict a total of 11 soil and weather parameters from SoilGrids.org provided reasonable fits. These fits were applied to actual soil profiles in Costa Rica and surrounding areas in Middle America. In most cases, the predicted soil profiles for each farm matched the actual profile with a confidence of >90%. This makes this approach promising, but (external) validation is needed before a solid judgement on the performance of this example model can be made.

The approach in this report, with R-codes provided, enables the collection and treatment of data from several analytical and non-analytical sources. The process of fusing (big) data into predictions that can be compared with a public database that covers the entire world offers the possibility of extrapolating predictions on origin on a global scale.

1 Introduction

Bananas are a huge commodity, are grown in tropical climates, and are largely exported to other parts in the world. Due to this long physical food chain it is difficult to ascertain the exact origin of a certain batch of bananas when they are being sold in, in our case, the Netherlands. Although the vast majority of the banana as we know it is one species – the 'Cavendish' banana, *Musa acuminata,* it is desired to be able to verify the origin of the banana we buy. Not every country in the world is equal in its production: environmental considerations in pesticide and fertilizer application, allocation of farming land, and social aspects around labour may vary across the different production locations. In addition to quality differences due to different natural growing conditions and different farming management, this leads to price differences due to origin. When price differences exist, there is an opportunity for fraudulent actions, and without means to verify the origin of bananas, it is likely that over time some of those fraudulent actions will succeed.

Therefore, analytical methods that can support the origin claims of banana shipments are needed. In previous projects, WFSR has gathered analytical data on a pilot set of bananas, aiming to develop a (multivariate/fingerprint classification) method directed at confirming a certain origin. However, next to differences between countries, also within a specific country, there might be significant regional variations like weather, soil characteristics and other influential factors that will change over time. Therefore, there is a need to link product properties, i.e. banana composition to be analysed upon arrival, to searchable facts on the production location, such as soil profile and weather/climate data. If these production location characteristics have a causal and predictive effect on the product properties, it will be possible to predict, and verify, the origin for all bananas being imported.

Aim

The aim of the overall project, therefore, is to develop a method that can link the composition of a bunch of bananas analysed in the Netherlands, to a certain origin. Preferably, this origin could be global, and thus the relevant data on the origin should be available for any of the possible production areas.

This report deals with the process of gathering, collating and processing the analytical and non-analytical data from various sources, and the creation of predictive models related to origin of the product.

2 Data description and sources

This report does not focus on the origin of the analytical data, but for its purpose it gives a very brief background, and data dimensions.

The core of the data set originates from a cooperation between WFSR and WUR Soil Geography and Landscape (SG). SG has had projects, contacts and knowledge on soils in Costa Rica and local banana farm management for a number of years, and was able to help WFSR with a selection of 15 farms that covered the variation of soil/climate conditions in that particular country in middle America. They also had access to analytical data on composition of soil and banana leaves (foliar) for each of these farms. In 2015, in a cooperation between SC and WFSR, 14[1] of those farms were visited, and banana samples were collected and transported to WFSR for further compositional analysis, performed between 2015 and 2018.

In 2018, soil and weather data accessible from external databases were extracted and linked to the existing data.
Details on the data are presented in Table 1.

Table 1 *Detailed data description*

Data	Dimensions (r x c)*	Nature, source, remarks
SG info	14 x 6	Classification of soil for each farm, categorical and ordinal variables. Provided by WUR Soil Geography and Landscape
SG soil	(2645 x 11) 14 x 11	pH, acidity and 9 elements in soil samples obtained between 2010-2014, analysed in Costa Rica. Several incomplete cases, various number of results per year per farm, sometimes missing year/farm. Averaged data for each farm, full data or per year
SG foliar	(2845 x 12) 14 x 12	12 elements in leaf samples obtained between 2010-2014, analysed in Costa Rica. Several incomplete cases, various number of results per year per farm, sometimes missing year/farm. Averaged data for each farm, full data or per year
SoilGrids	14 x 11	Continuous data for several chemical and physical soil data and climate data, retrieved from the 250m grid from SoilGrids.org (retrieved mid-2018) *(https://rest.soilgrids.org/query?lon=[LN]&lat=[LT), with coordinates LN and LT in decimal notation)*
KNMI	14 x 8	Continuous weather data, monthly (June 2015, date of collection and half-year, Jan-Jun 2015, growing period) averages. Obtained through KNMI from the CRU TS4.01 0.5°grid size database. Data per farm was assigned to the closest point (*source example for precipitation: https://climexp.knmi.nl/select.cgi?id=someone@somewhere&field=cru4_pre)*
PTR	187 x 155	Proton Transfer Reaction Mass Spectroscopy on the (freeze-dried) fruits, analysed at WFSR, results in concentrations (ppbv) for 155 nominal masses. Two or three replicates for each sample, 6 samples (bunches) per farm.
µNIR	(1215 x 125) 199 x 37	Micro-NIR reflectance spectra on the skin of the fresh, ripened banana, analysed at WFSR. Mostly 5 replicates per fruit, 3 fruits per bunch, 6 bunches per farm. Averaged per bunch and per position (top, bottom, middle of the hand). Some missing data. Edge removal, SNV and 1st derivative, then averaged over 3 consecutive columns.
Caro	(168 x 23) 84 x 23	HPLC-DA quantification of 23 carotenoid-like compounds in the banana peels, analysed at WFSR. (Averages of) 2 replicates per sample, 6 bunches per farm
ICPpeel	(172 x 25) 159 x 16	ICP-OES analyses of banana peel samples, analysed at Utrecht University. Two replicates per sample, 6 bunches per farm. Nine elements were considered too noisy and/or too low for reliable quantification, some outlying results were removed.
ICPpulp	(172 x 25) 163 x 15	See ICPpeel, but for pulp samples. Ten elements were discarded.

* Dimensions between brackets indicate the size of the available data, after removal of obvious errors and outliers and cases of incomplete data. The plain dimensions represent the data used in this report.

[1] The 15th farm could not be sampled due to dangerous road conditions at collection time. The data for this farm has been excluded from analysis.

2.1 Data (pre)treatment

All (analytical) data sets were initially individually scrutinised. In all cases, data were properly pre-treated. Refer to Table 1 for exact numbers of variables and samples retained for each data set. For elements and carotenoids, autoscaling (mean-center + column-wise standard deviation scaling) seemed appropriate, as it removes scale effects between elements that are not necessarily helpful. Autoscaling is dangerous and thus inappropriate if some responses have a large variance compared to their level. This happened for PTR data, where some nominal masses occurred near LOD, and ^{10}log scaling is preferred. Both ICP sets contained some elements that were very noisy, but it is harder to attribute this to meeting a certain LOD. Instead, we used expert opinion to identify and remove those elements that were not reliably quantified. The remainder of the sets were suitable to autoscale.

NIR data, acquired in reflection mode, suffers from baseline shifts. Therefore, signal normal variate (SNV) was applied, and a first derivative was taken to enhance the spectral features. The edges of the spectrum were noisy, which is normal for the technique used, and were removed. The number of points relative to the spectral features was high, and thus it was decided to reduce the number of points by averaging three consecutive points in each spectrum.

After these treatments, the individual data sets –before averaging, where applicable- were visually evaluated using PCA. The score plots were evaluated, first on obvious outliers, a few of which were found and removed, after which a new PCA was prepared. The subsequent PCA was evaluated for natural clustering. The appropriately scaled data did not appear to cluster into distinct groups but rather approached spherical structures in the first three dimensions, which is in line with the expectation of a multivariate normally distributed data set. Special care was taken to look for natural clustering or trends according to measurement series, and no obvious problems were observed. Then, a visual and again rather subjective, screening for repeatability was performed by checking the distance between sample replicates when plotted in several 2D-PCA plots (i.e. simultaneous display of plots for f1 vs f2, f1 vs f3, f1 vs f4, f2 vs f3, f2 vs f4 and f3 vs f4). For samples analysed in triplicate or more, the between-replicate distance relative to the spread of all samples in the PCA plot was used to identify deviating instances, which have led to a few omissions. If samples consisted of only duplicates, it is usually hard to identify the deviating instance of the two, if a large distance is observed. Nevertheless, in some cases one of the two sample replicates was sufficiently close to being an outlier that this instance was marked as a potential outlier, and averaging was not performed.

All remaining replicates were considered for averaging, although for some analyses (Table 1), the between-sample variance was considered important enough to keep them separate for further (multivariate) analysis.

2.2 Univariate analysis

Whenever a large number of data points is gathered, the most important question is whether a certain variable has a causal relationship with the parameter one wants to know. In some cases, this knowledge is available from experts or literature, or the causal relationship is strongly suspected. In many cases, however, they are just hypotheses. In the current case, the elements in bananas are considered to be possibly linked to soil composition, and NIR reflectance and PTR VOC concentrations are possibly linked to growing conditions (i.e. soil and weather) through their possible influence on the plant's metabolism. There is no hard evidence that this causal relationship exists, nor is there expert knowledge available to advice on which variables are the most important for our goal. However, we can perform simple univariate analysis of the chemical parameters measured to check their relationship with the parameters we want to predict. A straightforward way to do so is to draw correlation plots (Figure A1-A8). For 'medium-size' data sets, up to ~200 parameters, this allows to draw one plot that summarises the correlation of all variables with the intended parameters. A strong correlation (or anti-correlation) indicates that such a variable *might be* important for predicting a parameter, but it does not mean that this relationship is causal. Moreover, a weak direct correlation does not necessarily mean that a variable is not important. This is due to correlation effects and is

dealt with in multivariate analysis. Despite the uncertainties, correlations are a nice and easy way to give some direction towards easy to interpret results, and to have leads for further investigation of certain causal relationships.

Some of the correlation plots from the banana datasets are given in Annex I. The figures show that there are many correlations in the sense that any technique has one or more variables with a strong positive or negative correlation with any of the farm origin parameters. In very few cases, however, this effect is very strong or very close to 1. And many variables also have a very weak correlation. There are a few specific conclusions that can be drawn, but in this case, it proved difficult to reach hypothesis for causal insights from these plots. For the analyses on the bananas, there were hardly any variables without any correlation with any of the farm origin parameters to predict, so it was decided not to exclude any of the variables at this point.

2.3 Multivariate analysis

In any multivariate data set, correlation between variables is expected. This has at least two effects that are worth mentioning here: 1) Multiple variables with strong correlation may exist (even across analytical datasets – univariate correlation plots such as in Annex 1 confirm that this happens in these datasets too; data not shown) and these might cause undesired weight to the sample information they contain, especially if this information is not helpful in predicting the desired parameters. 2) Their correlation might change with any of the parameters to predict – or more concretely: two variables might overall be correlated and thus might both be low in certain samples and high in others. However, certain conditions in the soil might cause to influence the content of one of those variables, and cause that in certain samples from this condition the first variable could be low and the second could be relatively high. These effects are usually relatively subtle and are therefore almost impossible to spot in univariate views of the data. Multivariate analysis is perfectly suitable to deal with these effects, also if these correlations occur over more than two variables.

However, compared to univariate analysis, the results of multivariate analyses can be harder to interpret. Moreover, the more variables are included, the more chance that random correlations contribute to overly optimistic results, and therefore a proper validation of the model is required.

The number of options multivariate data analysis for this particular data set is huge (number of preprocessing options, algorithms, which parameters to predict), and only a part of the work performed on individual data sets is displayed in this report. Software used to perform the multivariate calculations were Pirouette 4.5 (Infometrix, Bothell, WA, USA) (mainly for visual data exploration) and R 3.5.0 (R Core Team, Austria). We only display regression predictions for the SoilGrid data, one parameter per individual dataset. That means that all classifications (prediction of categories) are omitted. Also the data from WUR Soil Geography is not displayed, as the latter is only available for a very limited number of production locations, and lacks extrapolation ability.

Table 2 presents the agreement (expressed as r^2) between the predicted (leave-one-out) and true value per individual data set. The goodness of these predictions are, very subjectively, considered as mediocre. There seems to be some capability of the models to predict the parameters, but it never gets 'really good'. Moreover, these results are obtained with leave-one-out validation (LOO), which is considered as the simplest and most optimistic form of validation. In LOO, as the name suggests, only one sample at a time is left out of the dataset, where all other samples are used to produce a model and predict the result for the left-out sample. Since there are many samples from one farm, all farms, with their own fixed parameters, are still represented by several samples in all sub-models. This makes this approach for this dataset over-optimistic. The next section (data fusion) will feature a more appropriate validation approach. If the individual analytical data sets are to be evaluated, software that features the option to leave-one-farm-out validation in this data is considered more suitable.

Table 2 *r^2 of validation (LOO) of PLS-predictions per individual sample set. Values in bold face indicate models that are graphically represented below.*

	Soil Organic Carbon stock *(tonnes/ha)*	Bulk density (fine earth fraction *(kg/m3)*	Soil texture fraction clay *(%)*	Soil texture fraction silt *(%)*	Soil texture fraction sand *(%)*	Cation exchange capacity (fine earth fraction *(cmolc/kg)*	Soil organic carbon content *(fine earth fraction) in permilles*	Soil pH in H_2O	Soil pH in KCl	Volumetric water content at wilting point pF 4.2	Average annual precipitation *(mm)*
PTR	.68	.59	.72	.36	.74	.45	.57	.64	.47	.54	**.73**
µNIR	.43	.40	.55	.33	**.59**	.19	.33	.43	.47	.42	.50
Caro	**.71**	.73	.76	.56	.75	.54	.44	.62	.33	.65	.78
ICPpeel	.65	.65	.68	.44	.72	.60	.62	.65	**.68**	.59	.74
ICPpulp	.64	**.78**	.74	.58	.70	.55	.60	.70	.80	.55	.76

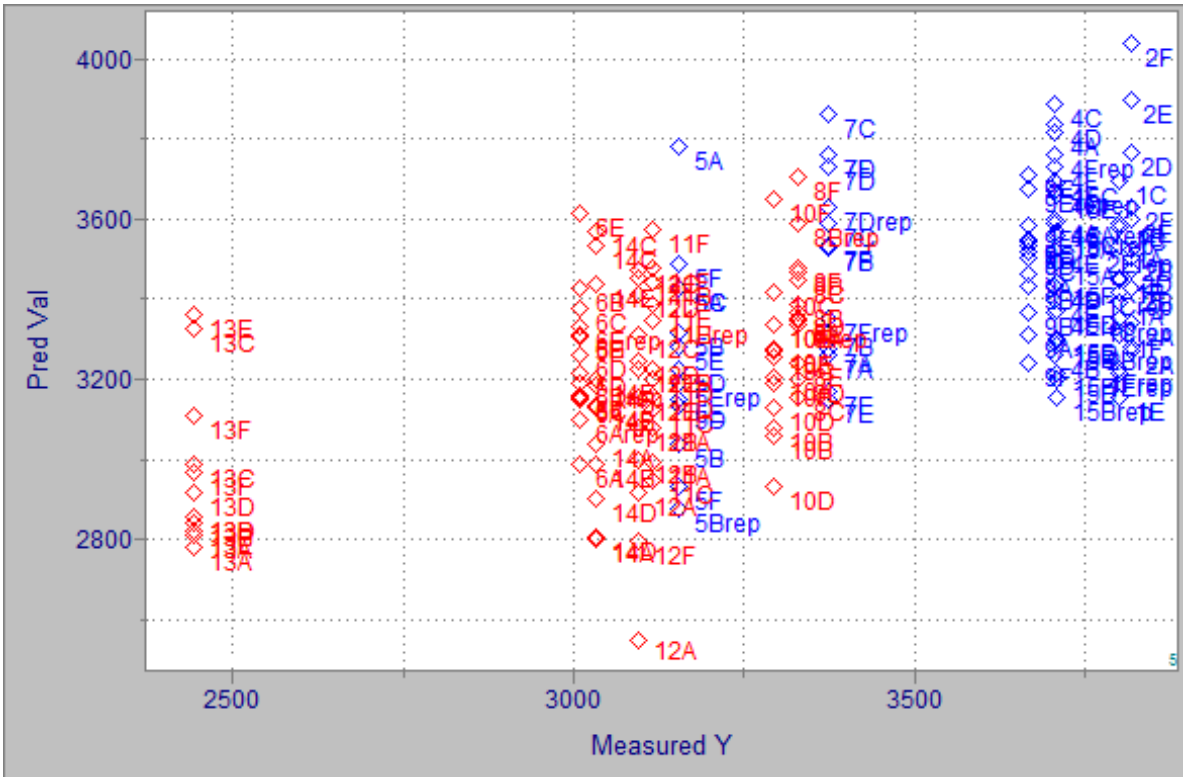

Figure 1 *PTR-MS prediction (^{10}log, mean-center) of average annual precipitation. Coloration by location north/south of Reventazon river.*

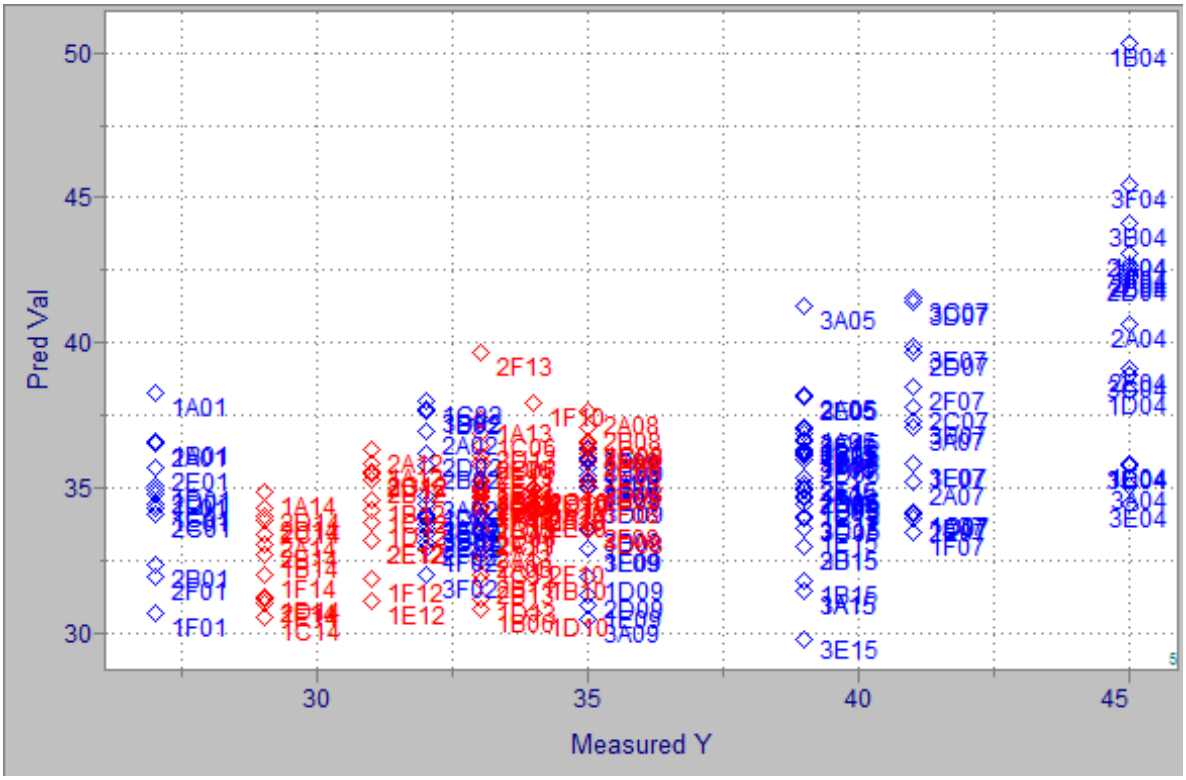

Figure 2 *Micro-NIR prediction (SNV, mean-center) of sand%. Coloration by location north/south of Reventazon river.*

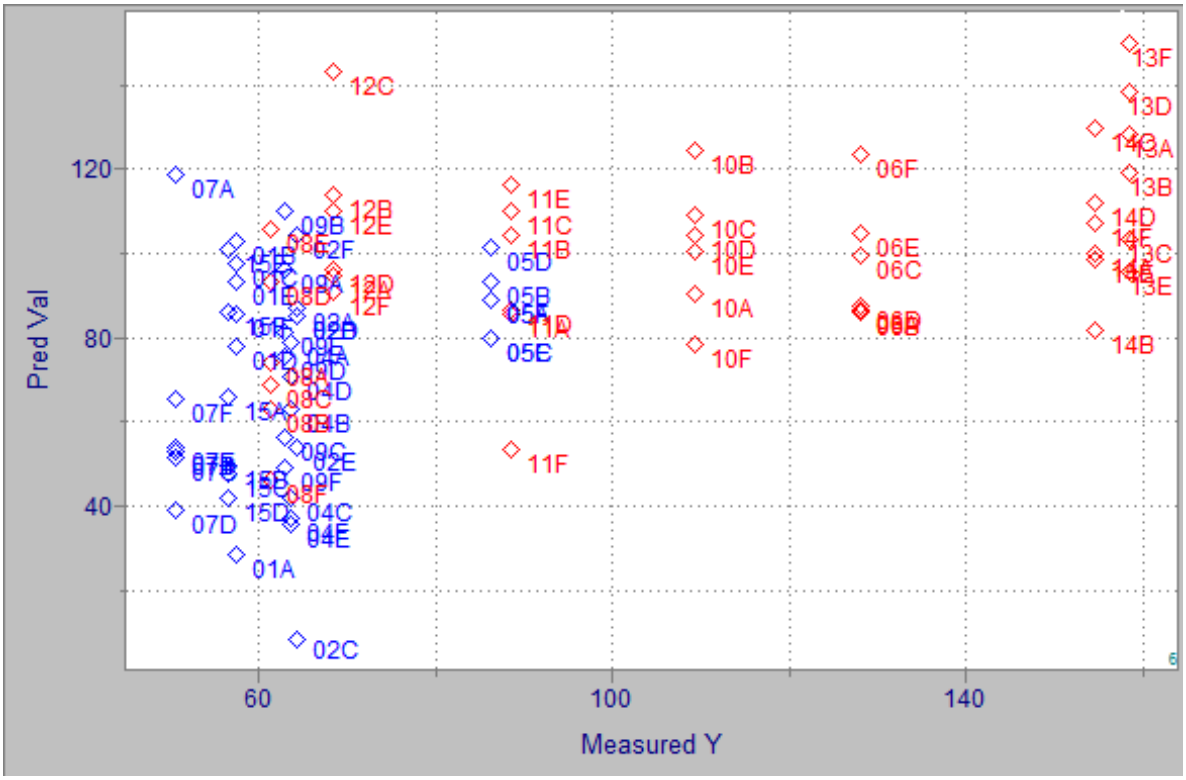

Figure 3 *Carotenoids prediction (autoscaled) of organic carbon stock. Coloration by location north/south of Reventazon river.*

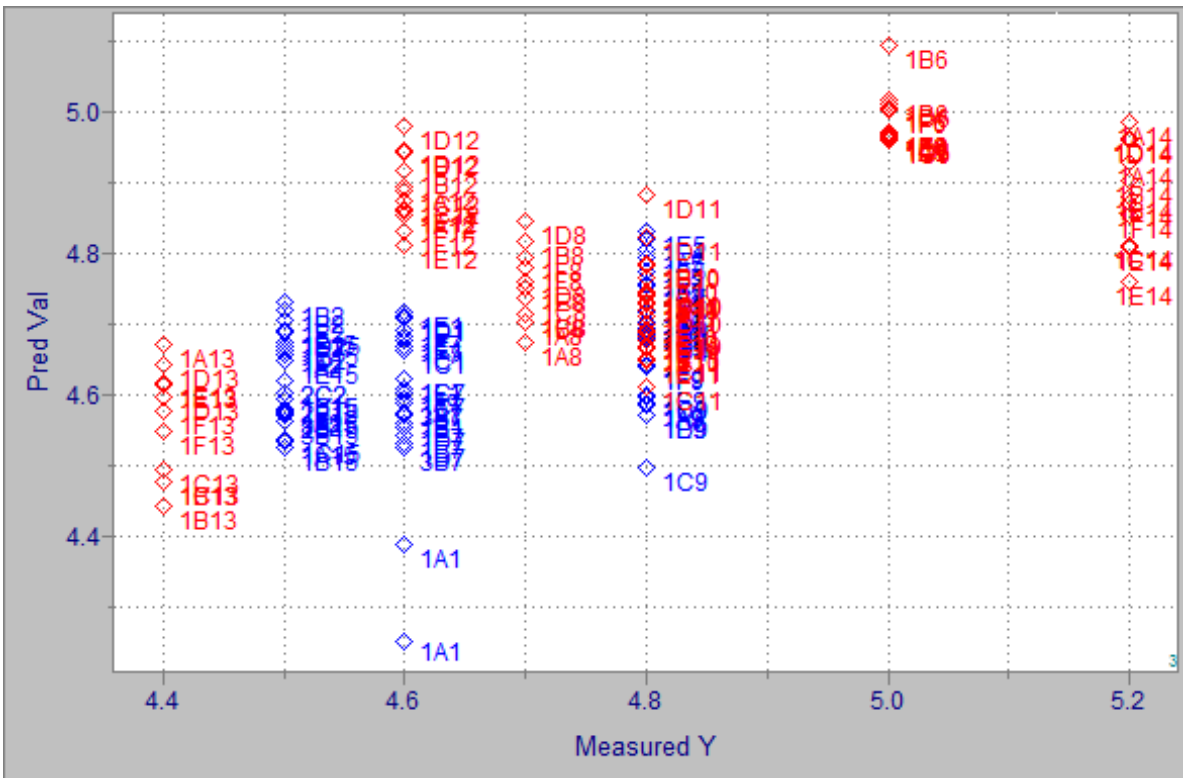

Figure 4 *ICP elements in peel prediction (autoscaled) of soil-pH in KCl. Coloration by location north/south of the Reventazon river.*

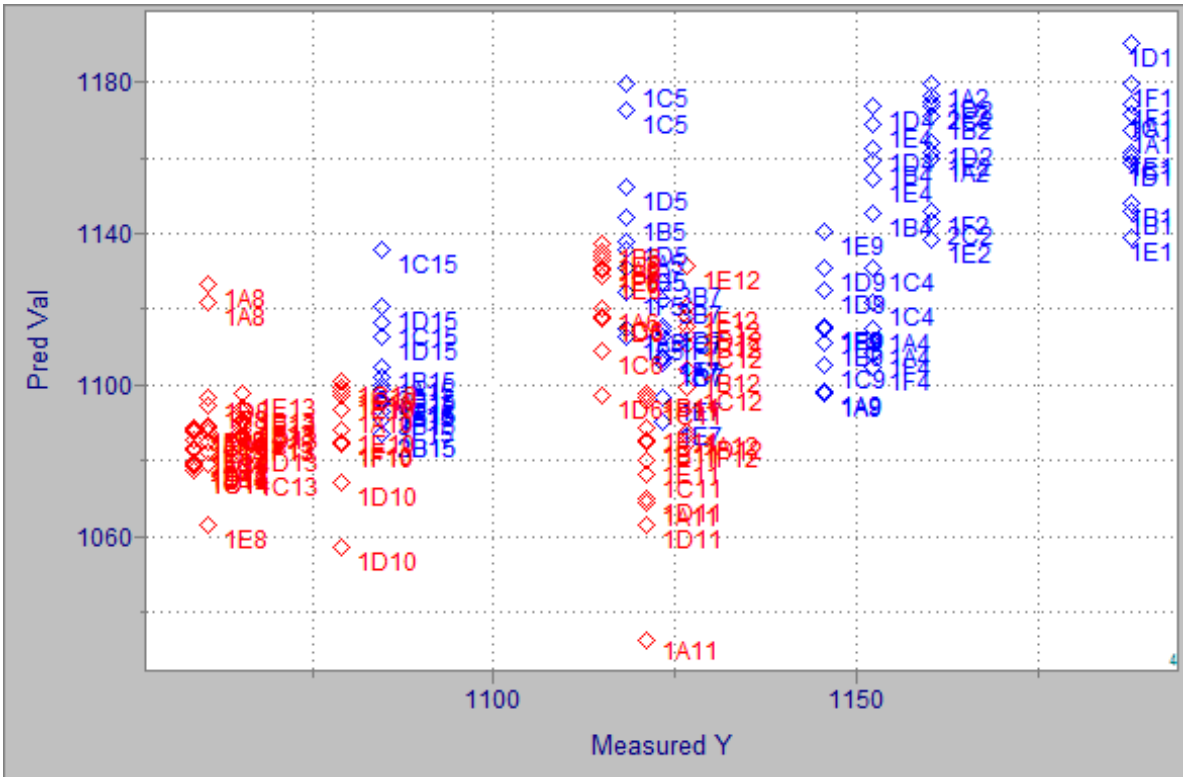

Figure 5 *ICP elements in pulp prediction (autoscaled) of bulk density. Coloration by location north/south of the Reventazon river.*

2.4 Data fusion

2.4.1 Introduction on data fusion

Data fusion is the process to bring data from different sources together in a sensible and useful way. It is a form of 'orthogonal analysis'; a phrase that is heard around WFSR the last months. There are three general modes of data fusion called low, mid, and high-level data fusion. In low-level data fusion, (analytical) data sets are aligned per sample, and put directly next to each other to form one data set. Mid-level data fusion first extracts features from each individual dataset and then merges these features into a data set. Features could be several variables with known predictive power for a certain parameter, or a number of factors from a PCA dimension reduction step, for example. High level data fusion indicates the process of making (multivariate) predictive models for individual data sets, and setting up rules how to use these multiple results into one final decision. In this report, low- and mid level approaches are implemented.

It is easy to argue that leaving an approach out makes this work incomplete. However, for any multivariate task there are simply too many valid possibilities to perform and compare. In most cases, it is impossible to predict upfront which (valid) approach is better. This is true for the form of data fusion, but also for pre-processing options, validation scheme, and model algorithm choice. Even within specific items, there are many degrees of freedom in the details. Examples of these details are the exact number of points to use for a spectrum first derivative, or the number of factors to employ for PLS predictions, which can be a fixed number, or determined using an algorithm, which in turn depends on the validation approach.

Every different approach and every change in detail will certainly lead to different outcomes, and there are always ways to further optimise the method. However, experience has shown that a conclusion whether a model is useable or not, i.e. the model's agreement with the data is poor – mediocre – reasonable – good – excellent, does not really depend on the details. What matters in the evaluation of multivariate methods, including those from multiple data set, in our opinion is:

- The data is of sufficient quality and size
- The data is appropriately pre-processed
- A robust algorithm is used (and only non-linear ones if the data needs that)
- A properly designed validation scheme is applied

We realise that this list is extremely subjective, and even among experts opinions will vary. It is therefore important for the credibility in a certain model that the steps are transparent and checkable for other experts.

2.4.2 Low level data fusion

In our current set, we have a number of analytical data sets, obtained on banana samples. We would like to use these to predict the conditions from their origin, which we know on farm level. That means that we have to fuse the analytical data on the banana samples, and to fuse the predictor variables.

The foliar analytical results, as we do not generally import banana leaf material with bananas, are not useful for our practical case and are disregarded. The analytical results on the farm's soils provided by WUR Soil Geography are included in the calculations, but have little practical value and are therefore not displayed in further sections of this report. The agreement between the predictions from the banana sample analyses and the WUR SG soil characteristics was comparable with the agreement with the SoilGrid's parameters.

Because of some missing analyses, some irregular additional analytical replicates, and some individual results classified as outliers, the analytical data sets on bananas were not equal in size (Table 1). The data sets contain both analytical replicates and replicates on bananas from the same farm. The variation between samples from the same farm come from analytical variation (repeatability and reproducibility effects) and natural variation between individual bananas (different bunches, and different bananas from the same bunch). Since the within-farm variation was rather large for all

techniques used, we decided that averaging all samples into one average per farm would solve the problem of unequal data dimensions for data fusion, but that would also reduce the analytical variation to create, in our opinion, a too optimistic model. Therefore, we created a matrix with exactly 14 (farms) x 6 (bunches) x 3 (replicates) = 252 samples. Data per farm/bunch are sampled from the original data if available, missing data are sampled from other bunches from the same farm *(R-code 1. Low-level data fusion script)* This is a practical approach that assumes equivalence of analytical replicates and samples from the same bunch. It creates a data set without any missing data, which is also perfectly balanced for farm/bunch. There is an element of chance in the sampling step, and this procedure should be and was repeated for a number of times without noticeable differences in the final results (data not shown).

R-code 1. Low-level data fusion script

This R-code snippet assumes data from 5 different techniques to be present in 5 different data frames, and fuses them into a farm-balanced flatted (i.e. matrix-like) list called 'FusedWorkData'.

```
OrgData <- list (PTRdata,
                 ICPPeeldata,
                 ICPPulpdata,
                 Carodata,
                 NIRdata)

SampleRep <- 3
FusedData <- list (PTR= matrix(data=NA, nrow=14*6*SampleRep, ncol=ncol(PTRdata)-2),
                   ICPPL= matrix(data=NA, nrow=14*6*SampleRep, ncol=ncol(ICPPeeldata)-2),
                   ICPPP= matrix(data=NA, nrow=14*6*SampleRep, ncol=ncol(ICPPulpdata)-2),
                   Caro= matrix(data=NA, nrow=14*6*SampleRep, ncol=ncol(Carodata)-2),
                   NIR= matrix(data=NA, nrow=14*6*SampleRep, ncol=ncol(NIRdata)-2))
#-2 is to remove the (non-numeric) Farm ID and Bunch IDs.
#14 for 14 farms, 6 for 6 bunches per farm

colnames(FusedData[["PTR"]]) <- colnames(PTRdata)[1:(ncol(PTRdata)-2)]
colnames(FusedData[["ICPPL"]]) <- colnames(ICPPeeldata)[1:(ncol(ICPPeeldata)-2)]
colnames(FusedData[["ICPPP"]]) <- colnames(ICPPulpdata)[1:(ncol(ICPPulpdata)-2)]
colnames(FusedData[["Caro"]]) <- colnames(Carodata)[1:(ncol(Carodata)-2)]
colnames(FusedData[["NIR"]]) <- colnames(NIRdata)[1:(ncol(NIRdata)-2)]

Counter <- 1; FusedID <- NULL
for (Farm in c(1,2,4,5,6,7,8,9,10,11,12,13,14,15)) {
  for (Bun in c("A", "B", "C", "D", "E", "F")) {
     for (Tech in 1:5) {  # 5 techniques
       if (length (which(OrgData[[Tech]]$Farm==Farm&OrgData[[Tech]]$Bunch==Bun))>0) {
           AddThese <-sample(which(OrgData[[Tech]]$Farm==Farm &
           OrgData[[Tech]]$Bunch==Bun), SampleRep,replace=TRUE) } else {
           AddThese <- sample(which(OrgData[[Tech]]$Farm==Farm), SampleRep,replace=TRUE)
           }
           FusedData[[Tech]][Counter:(Counter+SampleRep-1),] <-
                  as.matrix(OrgData[[Tech]][AddThese,1:(ncol(OrgData[[Tech]])-2)])}
       Counter <- Counter+SampleRep
    FusedID <- c(FusedID, rep(Farm, SampleRep))
  }  }

FusedWorkData <- cbind(FusedData[[1]], FusedData[[2]], FusedData[[3]], FusedData[[4]],
FusedData[[5]])
FusedTech <- NULL
for (i in 1:5) {FusedTech <- c(FusedTech, rep(names(FusedData)[i], ncol(FusedData[[i]])))
)}
```

2.4.3 Mid-level data fusion

As mid-level data fusion approach a PCA dimension reduction was used *(R-code 2. Mid-level data fusion).* The original data was replaced with the PCA scores. In a second step, the number of scores (per technique) was reduced, in this case to half of the original number of variables with a maximum of 20. This is a rather subjective choice, but we feel this fits the data well. An obvious alternative is to select the minimum number of samples to reach a certain percentage of explained variance. Although this sounds more objective, the cut-off point still needs to be selected, and the explained variance is strongly dependent on the nature of the data and its pre-processing. Moreover, the amount of total explained variance is not necessarily correlated with the predictive power of a latent variable. The resulting data is used for further processing.

```
R-code 2. Mid level data fusion
This script transforms the flat "low-level" fused data matrix into PCA scores (separately
per technique). Then, up to 20 latent variables with the highest explained variance are
retained, as they are stored in index-variable "TheseColumns".
```

```
FusedPCAData <- FusedWorkData
FusedPCAData[,which(FusedTech=="PTR")] <-
             princomp(scale(FusedWorkData[,which(FusedTech=="PTR")]))$scores
FusedPCAData[,which(FusedTech=="ICPPL")] <-
             princomp(scale(FusedWorkData[,which(FusedTech=="ICPPL")]))$scores
FusedPCAData[,which(FusedTech=="ICPPP")] <-
             princomp(scale(FusedWorkData[,which(FusedTech=="ICPPP")]))$scores
FusedPCAData[,which(FusedTech=="Caro")] <-
             princomp(scale(FusedWorkData[,which(FusedTech=="Caro")]))$scores
FusedPCAData[,which(FusedTech=="NIR")] <-
             princomp((FusedWorkData[,which(FusedTech=="NIR")]))$scores
FusedWorkData <- FusedPCAData

TheseColumns <- NULL;
  for (i in unique(FusedTech)) {
    nPCs <- round(length(which (FusedTech ==i))/2)
    if (nPCs>20) {nPCs <- 20} #half of number of original pcs, maximised at 20.
  TheseColumns <- c(TheseColumns, which (FusedTech ==i)[1:nPCs])  }
```

2.4.4 Outlier detection

Although the data was screened for outliers in the individual data sets, the fused data were not. The fused dataset appeared to suffer from a few outliers, which were influencing the predictions negatively. It is tricky to remove outliers twice, although it can be argued that by fusing data new insights and new situations might arise, which allows renewed outlier detection and removal. If the current models were to be used in practice, this step would need more careful consideration, but by applying the code below *(R-code 3. Outlier removal)*, 6 of the 252 samples were removed as outliers. This outlier detection routine is based on the Mahalanobis distance of autoscaled data, and checks whether there are extreme values that are not part of the (non-parametric) distribution.

```
R-code 3. Outlier removal (for fused data, but applicable for any multivariate dataset)
This script tests the Mahalanobis distance of each sample's (included, "TheseColumns")
variables in the dataset, and determines its distribution. Samples clearly separated from
this distribution are identified as outliers, and can be excluded in further analysis.
```

```
NonOL <- rep(TRUE, nrow(FusedWorkData))
FusedWorkData <- scale(FusedWorkData)
Maha <- mahalanobis(FusedWorkData[, TheseColumns], center=0,
           cov=cov(FusedWorkData[, TheseColumns]), inverted=TRUE)
```

```
MahaOL <- density(Maha)
MahaOL.index <- rle(MahaOL$y>(max(MahaOL$y)/20))
 for (i in 1:length(MahaOL.index$values)) {
  until.index <- sum(MahaOL.index$lengths[1:i])
  if (MahaOL.index$values[i] & (sum(MahaOL$y[1:until.index])/sum(MahaOL$y))>.5 )
           {break}  #that means that until.index contains the "TRUE"(density is higher
                    #than 5% of the max dens), AND the sum of the density is higher than
                    #50% of the total. The rest are probably outliers.
 }
NonOL <- Maha<= MahaOL$x[until.index]
```

2.4.5 Multivariate predictions

The predictions of the different parameters are performed by a simple PLS regression model *(R-code 4. Predictive modelling)*. This includes a proper validation step, in which all samples from one farm were removed from the dataset, and were predicted by the model based on the other farms. PLS is a robust and fairly standard algorithm for regression, and no other algorithms are compared.

The r2 of validation are given in Table 3, which can be compared to the individual data (calculated in Pirouette, Table 2).

```
R-code 4. Predictive modelling
This script performs the predictive modelling, on the (fused) data provided, without
outliers and possibly without the lower PCs. It automatically selects the number of PLS
components to minimise the PRESS (Predicted Residual Error Sum of Squares). Standard is
segmented (farm) cross validation, LOO validation is an option provided. The predictions
are stored in "predictions_value"
```

```
library (pls)
CVseg <- list() # update the segments for OLs
for (i in unique(FusedID)) {  CVseg <- c(CVseg, list(which(FusedID[NonOL]==i))) }
par (mfrow=c(6,7), mai=c(.2,.1,.2,.1)) #note: this scrip also visualises the output

predictions_r2 <- NULL
predictions_value <- NULL
for (i in 1:ncol(ThisPred)) {
  RegDF <- data.frame (DAT=FusedWorkData[NonOL,TheseColumns],
                       Y=ThisPred[Y_index[NonOL],i])
  PLSmodel <- plsr (Y~., data=RegDF, ncomp=10, validation="CV", segments=CVseg,
                    scale=TRUE, center=TRUE)
  #uncomment the next line for LOO validation:
  #PLSmodel <- plsr (Y~., data=RegDF, ncomp=10, validation="LOO", scale=TRUE,
                     center=TRUE)
  NumComp <- which.min(PLSmodel$validation$PRESS)
  LinPred <- lm(PLSmodel$validation$pred[,1,NumComp]~ThisPred[Y_index[NonOL],i])
  Corr <- cor(PLSmodel$validation$pred[,1,NumComp],ThisPred[Y_index[NonOL],i])
  predictions_r2 <- c(predictions_r2, Corr)
  predictions_value <- rbind (predictions_value, PLSmodel$validation$pred[,1,NumComp])
  MainCol <- "red"; if (abs(Corr)>.5) {MainCol <- "orange"};if (abs(Corr)>.7)
             {MainCol <- "darkgreen"}
  plot(PLSmodel$validation$pred[,1,NumComp]~ThisPred[Y_index[NonOL],i],
       main=list (paste0(colnames(ThisPred)[i], " r2=", round(Corr,2)," nC=", NumComp),
       col=MainCol), xlab=NA, ylab=NA, cex.main=.8, cex.axis=.8)
  abline(a=LinPred$coefficients[1], b=LinPred$coefficients[2], col=MainCol)
}
```

Table 3 *r^2 of validation of PLS-predictions for fused data.*

	Soil Organic Carbon stock *(tonnes/ha)*	Bulk density (fine earth fraction *(kg/m3)*	Soil texture fraction clay *(%)*	Soil texture fraction silt *(%)*	Soil texture fraction sand *(%)*	Cation exchange capacity (fine earth fraction *(cmolc/kg)*	Soil organic carbon content *(fine earth fraction) in permilles*	Soil pH in H_2O	Soil pH in KCl	Volumetric water content at wilting point pF 4.2	Average annual precipitation *(mm)*
Fused, LOO	.79	.87	.80	.70	.79	.75	.83	.78	.82	.81	.79
PCA, LOO	.69	.75	.71	.54	.70	.48	.65	.62	.66	.63	.77
Fused, Val	.09	.10	.40	.35	.12	.39	.04	.42	.35	.04	.27
PCA, Val	.08	.22	.05	.54	.17	.61	.11	.36	.09	.20	.31

Tables 2 and 3 show that the LOO-validated data for the fused data set is clearly better than the individual data sets. However, when a single farm is left out, the performance drops severely, and would in most cases be considered of insufficient quality. As discussed previously, there is room for optimisation of the performance of the prediction. The most obvious improvement would be to select those variables that are relevant for each of the desired parameters, and build separate models for each parameter. This is likely to yield some improvement in prediction accuracy. This variable selection needs to be performed carefully, and with some form of validation. This step is recommended if this data set is to be used in the future.

2.4.6 Retrieve external data

Initially, the training data from SoilGrid.org (and the KNMI database) were extracted manually from the respective websites. To extrapolate the results, data for all relevant areas are needed, and it proved very easy to automate the extraction *(R-code 5. Data retrieval)*. For convenience for future users, the code used for SoilGrid data retrieval and processing into useable format is given below.

```
R-code 5. Data retrieval from SoilGrid.org
This script performs data extraction and 'flattening' into a matrix where the desired 11
properties are stored per coordinate. Two sets of coordinates are given, one for Costa
Rica, one for a larger rectangular area covering Middle- and a a part of South America at
two different resolutions.
```

```
library(jsonlite)
Xcoord <- seq(-86,-82.5, by=.025)  #Costa Rica
Ycoord <- seq(8,11, by = .025)     #Costa Rica
Xcoord <- seq(-96,-67, by=.05)  #For a larger area middle and south america
Ycoord <- seq(-3,21, by = .05)  #For a larger area middle and south america

SOILGRID <- array(NA, dim=c(length(Xcoord), length(Ycoord), 11))
for (x in 1:length(Xcoord)) {
  for (y in 1: length(Ycoord)) {

dat <- read_json(paste0("https://rest.soilgrids.org/query?lon=", Xcoord[x], "&lat=",
Ycoord[y],"&attributes=OCSTHA,BLDFIE,CLYPPT,SLTPPT,SNDPPT,CECSOL,ORCDRC,PHIHOX,PHIKCL,WWP
,PREMRG"))  #this single line retrieves the desired data as javascript object

SOILGRID[x,y,1] <- tryCatch ({mean(as.numeric(dat$properties$OCSTHA$M[ which
(names(dat$properties$OCSTHA$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]))}, error=function(e){NA})
```

```
SOILGRID[x,y,2] <- tryCatch ({mean(as.numeric(dat$properties$BLDFIE$M[ which
(names(dat$properties$BLDFIE$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,3] <- tryCatch ({mean(as.numeric(dat$properties$CLYPPT$M[ which
(names(dat$properties$CLYPPT$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,4] <- tryCatch ({mean(as.numeric(dat$properties$SLTPPT$M[ which
(names(dat$properties$SLTPPT$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,5] <- tryCatch ({mean(as.numeric(dat$properties$SNDPPT$M[ which
(names(dat$properties$SNDPPT$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,6] <- tryCatch ({mean(as.numeric(dat$properties$CECSOL$M[ which
(names(dat$properties$CECSOL$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,7] <- tryCatch ({mean(as.numeric(dat$properties$ORCDRC$M[ which
(names(dat$properties$ORCDRC$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,8] <- tryCatch ({mean(as.numeric(dat$properties$PHIHOX$M[ which
(names(dat$properties$PHIHOX$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,9] <- tryCatch ({mean(as.numeric(dat$properties$PHIKCL$M[ which
(names(dat$properties$PHIKCL$M) %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,10] <-tryCatch ({mean(as.numeric(dat$properties$WWP$M   [ which
(names(dat$properties$WWP$M)    %in% c("sl1", "sl2", "sl3", "sl4","sd1", "sd2", "sd3",
"sd4")) ]) )}, error=function(e){NA})
SOILGRID[x,y,11] <-tryCatch ({mean(as.numeric(dat$properties$PREMRG$M))},
error=function(e){NA})
  } }
```

2.4.7 Model performance

Although the (validated) predictions from the model are not great, they are not totally random either. Moreover, they need to be put in perspective. That is, we are not necessarily interested in accurate prediction in the soil properties, but in a location where a banana sample originated from. One of the many possible ways to integrate the predictions into useable information is to correlate the set of (11) predictions with the known values for each geographical point. A script *(R-code 6. Create location probabilities)* to do so is provided, followed by a graphical representation of the results (correlation of the soil type predicted from the banana composition with the actual soil type for a large (Figure 6) area around Costa Rica, and in more detail for Costa Rica only (Figure 7).

R-code 6. Create location probabilities

```
This script plots the possible locations based on the predicted SoilGrid parameters. A
matrix is filled with correlations between predicted (validated!) and known values. These
points are ordered, and interpolated between 0 and 1, so that 0 represents the point with
the worst (numeric) correlation with the predicted data, 1 the best. This is plotted as a
map, with some custom color setting.
```

```
for (i in unique(FusedID)) {  #this will loop over the farms
thisFarm <- which(FusedID[NonOL]==i)
thisFarmAvgPred <- rowMeans(predicted_values[,thisFarm])
thisDistMap.c <- matrix (NA, ncol=dim(SOILGRID)[2], nrow=dim(SOILGRID)[1])
for (x in 1:length(Xcoord)) {
   for (y in 1: length(Ycoord)) {
     thisDistMap.c[x,y] <- cor(thisFarmAvgPred, SOILGRID[x,y,]) }}
```

```
C_prob <- ecdf(thisDistMap.c)
thisDistMap <- matrix(C_prob(thisDistMap.c)^2,nrow=nrow(thisDistMap.c))

plot(NA, xlim=c(min(Xcoord),max(Xcoord)), ylim=c(min(Ycoord),max(Ycoord)),
     main=paste0("Farm ", i), xlab="Longitude", ylab="Latitude")
rect(min(Xcoord),min(Ycoord),max(Xcoord),max(Ycoord), col="blue") #color the sea blue
image(x=Xcoord, y=Ycoord, z=thisDistMap,add=TRUE, breaks=c(0, 0.5, 0.8, 0.85, 0.9, 0.92,
0.94, 0.96, 0.98, 0.99,1),col=heat.colors(10))
points (FarmLocs$X.Coordinates[i], FarmLocs$Y.Coordinates[i], pch=21, col="green",
cex=1.2, lwd=2) #FarmLocs contains the coordinates from the farms
readline ("Enter to continue")  #waits for Enter until the next farm is plotted
}
```

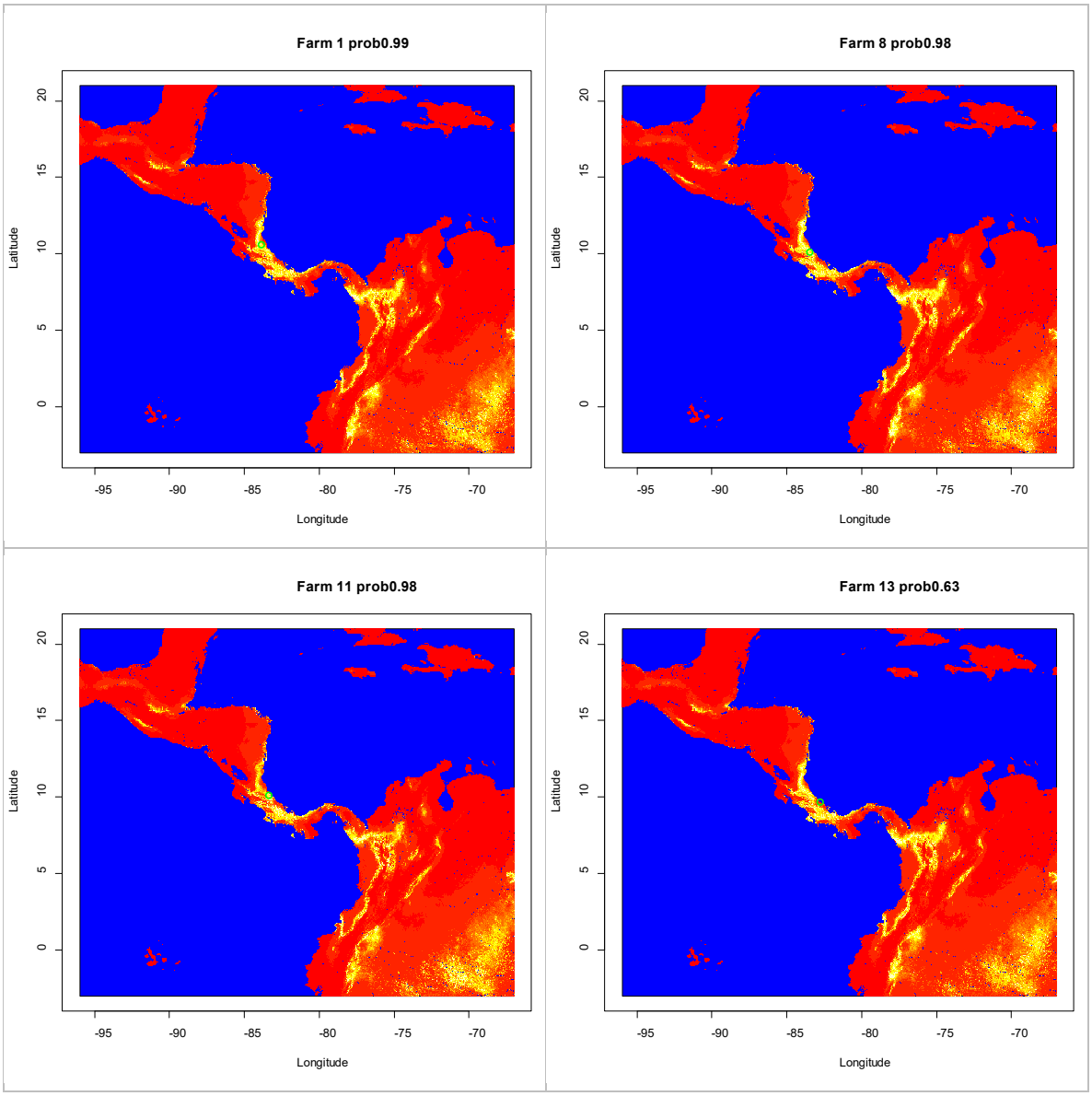

Figure 6 *Location probability based on predicted soil properties. Prediction is based on PCA-preprocessed data with leave-farm-out validation. Colour relates to increasing correlation of the predicted profile with the actual profile, intense yellow are probably locations of origin, red is unlikely. The actual farm location is marked with a green circle.*

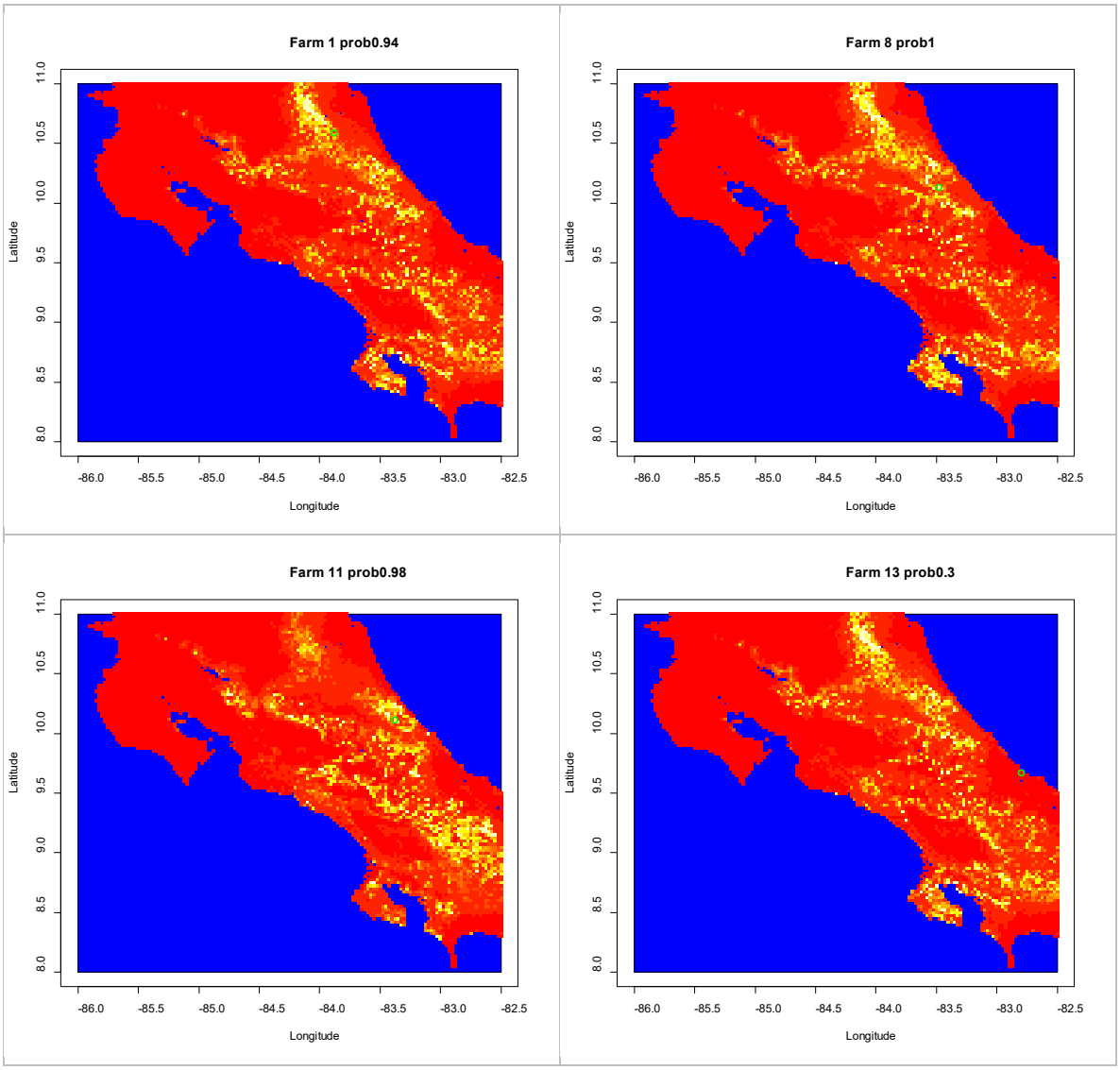

Figure 7 *Location probability based on predicted soil properties, only for Costa Rica and direct surrounding area. Prediction is based on PCA-pre-processed data with leave-farm-out validation. Colour relates to increasing correlation of the predicted profile with the actual profile, intense yellow are probably locations of origin, red is unlikely. The actual farm location is marked with a green circle.*

Table 4 *Probabilities of the actual farm location as predicted by the fused model (leave-farm-out validated, PCA processed results). Values of 1.000 would indicate the best correlating pixel (location) on the map.*

	Costa Rica	America		Costa Rica	America
Farm 1	.936	.990	Farm 9	.928	.992
Farm 2	.850	.997	Farm 10	.933	.989
Farm 4	.827	.997	Farm 11	.982	.983
Farm 5	.982	.942	Farm 12	.796	.997
Farm 6	.643	.898	Farm 13	.301	.635
Farm 7	.906	.965	Farm 14	.968	.980
Farm 8	.998	.978	Farm 15	.673	.936

The results (Table 4) indicate that for some farms (scores >90%) the overall prediction comes quite close, and perhaps close enough, to what is needed to verify banana origin. But there are also a few farms that show that this model is most probably not yet suited for practical use.

3 Conclusions and recommendations

This document aims to give an example of the combination of different types of analytical data for certain products, and the combination of different external data sources. It shows some practical ways and concrete scripts that can be used in similar cases. Some considerations on modelling and data fusion are given too, with the idea of reusability of the scripts in mind.

As the degrees of freedom in concrete approaches to model the data is almost infinite, the most important issue is the validation of the model. In the current set, two forms of validation are used, which show that a 'proper' validation gives dramatically different results than an optimistic (and common) validation approach. And even the 'proper' validation approach is a form of cross validation, before any attempt to use the current data set in practice, external validation should be performed.

The results for the current data set, in terms of prediction of individual soil parameters, are considered to be mediocre at best. There are certainly ways to optimise the models, most probably through high-level data fusion and careful variable selection. This is quite labour intensive, and given the current results, the improvement is expected to be only minor.

However, when the individually predicted parameters are combined and compared with the soil data that is available, the results are reasonably good: there appear to be relatively few places which correlate well enough with the predicted values. This means that many locations in Costa Rica and certainly in the wider Middle America region can be excluded as possible origins, as shown in the maps in this report.

Based on that finding, the combination with publicly available data sources and analytical data on products with a known (or claimed) origin can still be considered promising.

In order to bring the data and approaches to such a level that they can be used to effectively evaluate the origin of bananas it is recommended to:

- Develop and apply (multivariate) QC approaches for each of the analyses to ensure analytical stability over years.
- Establish the long-term stability and suitability of the model by checking the performance with samples from new harvest dates from the 14 farms used in this study, and
- validate the model with new samples from known origin, from Costa Rica and other relevant production areas.
- Carefully define the model's mode of application and performance. For instance, if new samples with a claimed origin need to be verified, what is the threshold for authenticating this claim, and what is its confidence?

More generally, the following research is necessary in order to gain more insight in the applicability of combining general (big) data with specific analytical results in order to evaluate the product claims

- Develop and apply (multivariate) QC approaches from the beginning of the study. A starting point might be the approach formulated in SOP F-0094.
- Develop infrastructure and protocols to facilitate merging and using data from different sources, starting with data from analytical instruments and their meta data. FAIR principles are a good starting point.
- Develop, share and maintain code, according to good coding practices, as code is needed to perform the classification as described in this study. This type of work cannot be realistically be performed in 'standard' office programmes (excel), and expertise to use, maintain and improve this code should be present in more than one or two individuals.

Acknowledgements

Thanks go to Saskia van Ruth (project leader), Yamine Bouzembrak, Cheng Liu and Leen van Ginkel for their review and valuable discussions, Naomi Jonker, Marijke Reijnen, Paula Fernandez, René van der Molen for the analyses, Marie Wesselink for the soil measurements and information, Eric Cuijpers for getting the banana samples.

Annex 1 Correlation plots

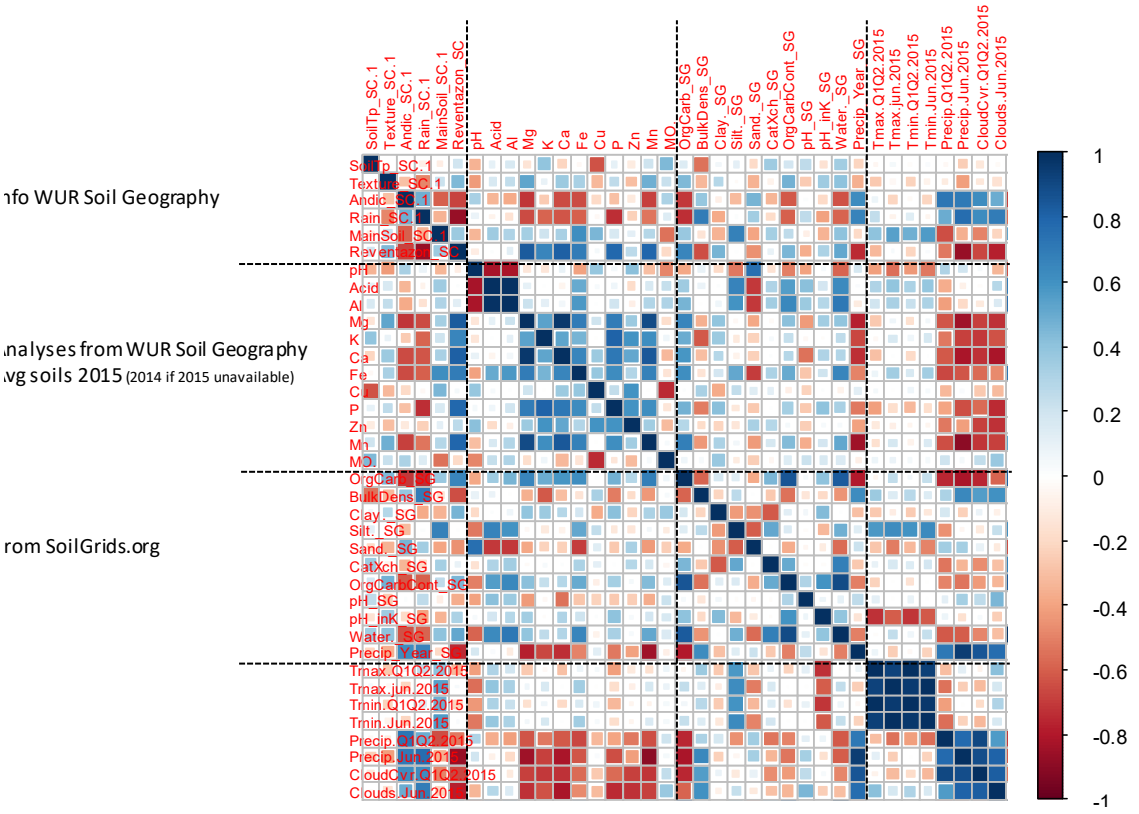

Figure A1 *Correlation plot between different 'external' data sources: data on the 14 banana farms in Costa Rica.*

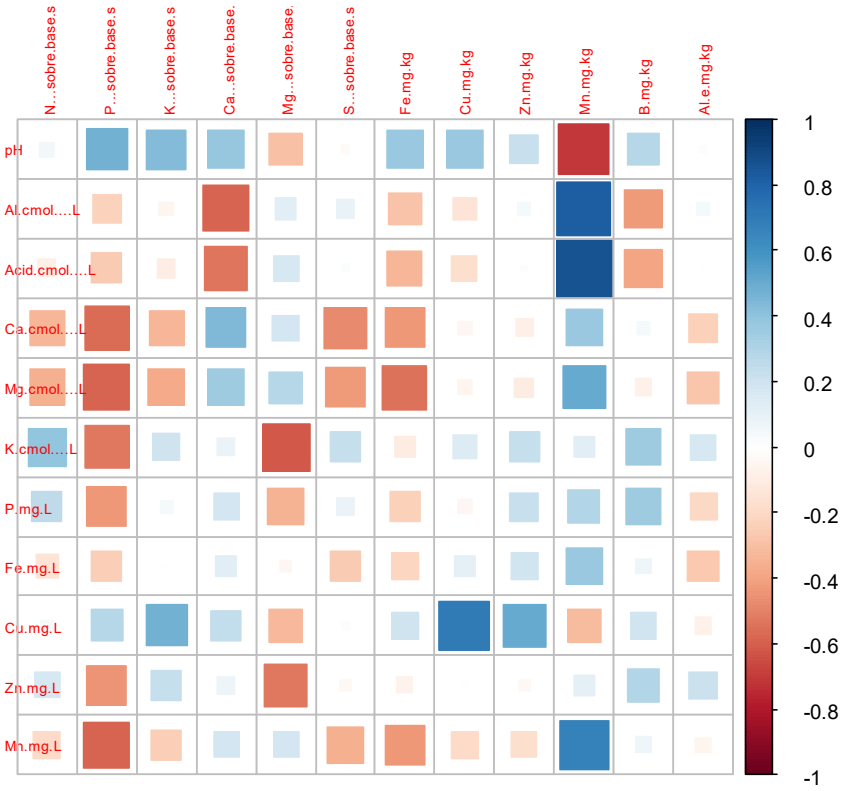

Figure A2 *Correlation plot between soil and foliar composition data on the 14 banana farms in Costa Rica. Analyses in Costa Rica in cooperation with WUR Soil Geography. Averages over the available data 2008-2015.*

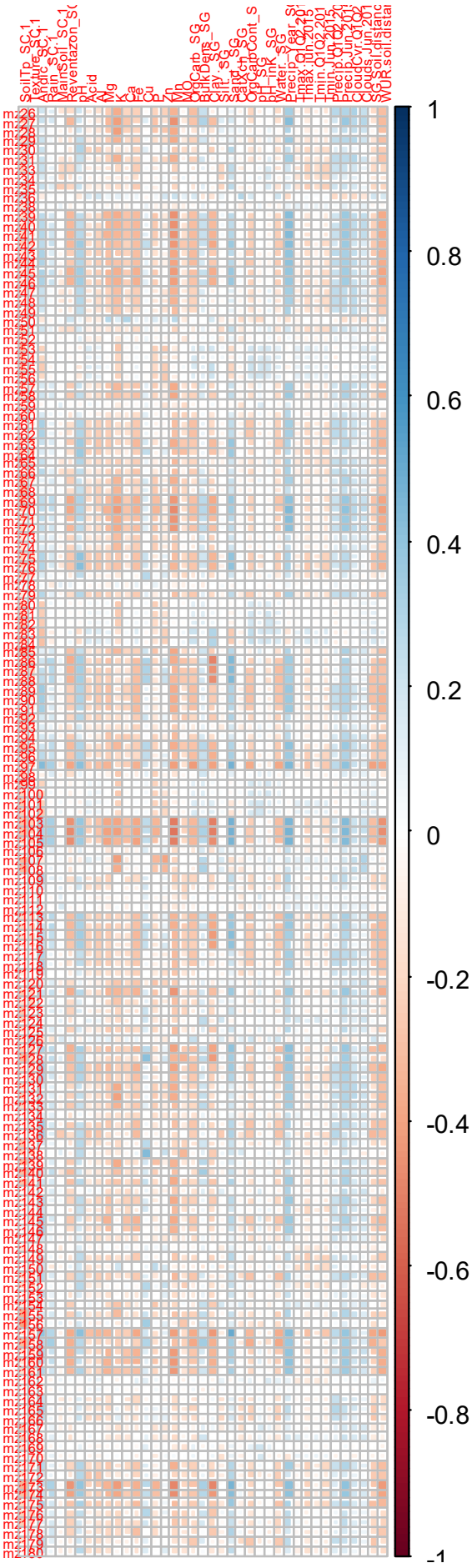

Figure A3 *Correlation plot between PTR-MS variables (VOC masses measured in bananas, at WFSR) and known data – analytical and public database-extracts- on the farms of origin.*

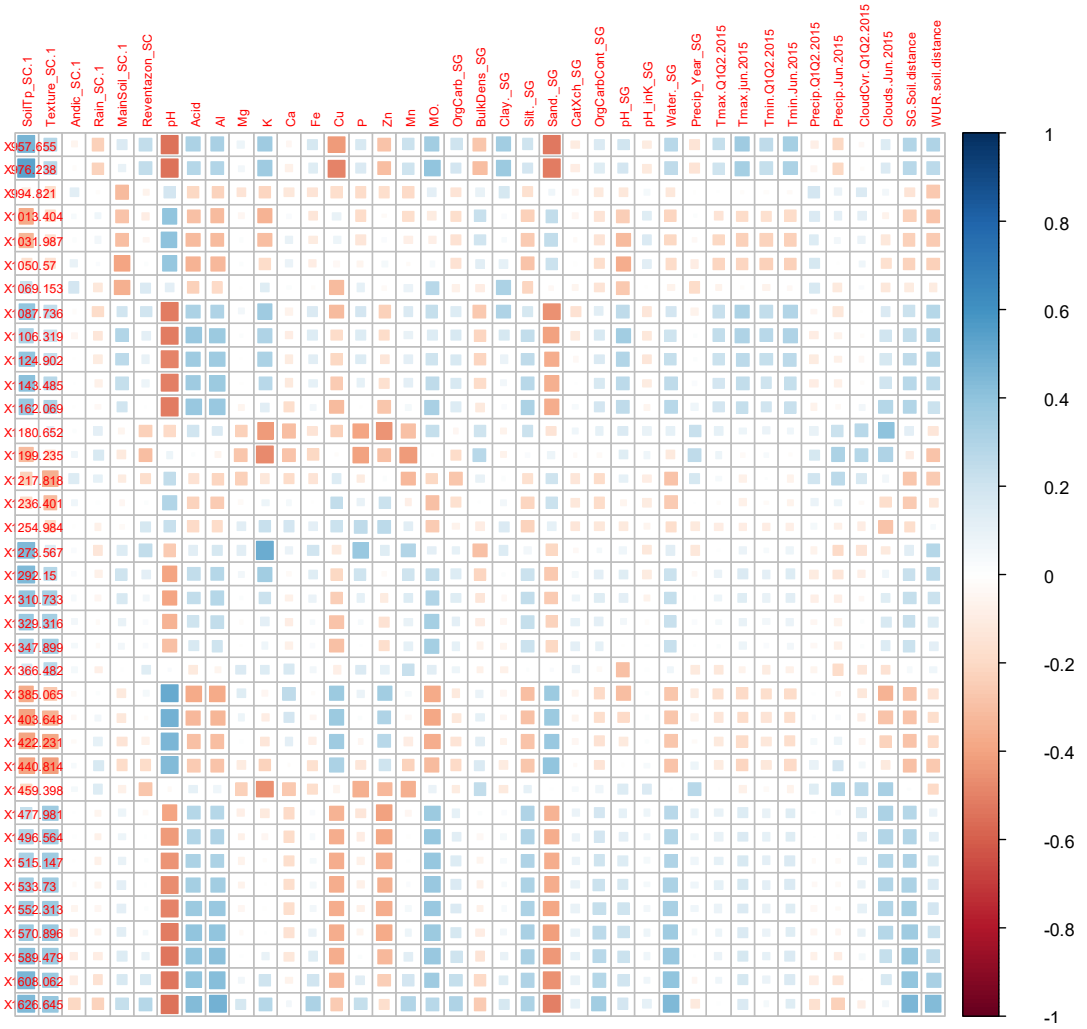

Figure A4 *Correlation plot between micro-NIR variables (wavelengths, after SNV and derivative, measured at bananas skins, at WFSR) and known data – analytical and public database-extracts- on the farms of origin.*

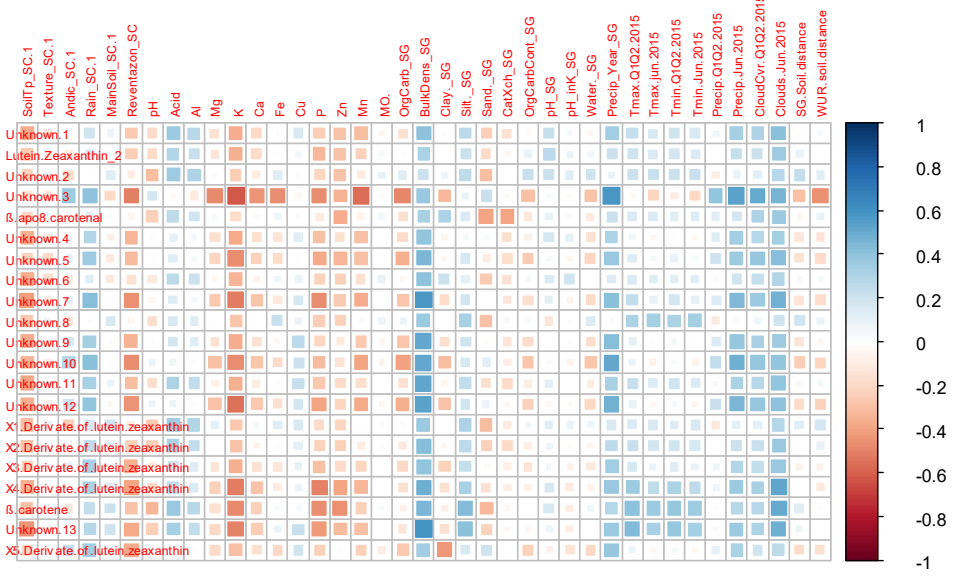

Figure A5 *Correlation plot between HPLC-DAD variables (carotenoids or lookalikes measured in banana peel samples at WFSR) and known data – analytical and public database-extracts- on the farms of origin.*

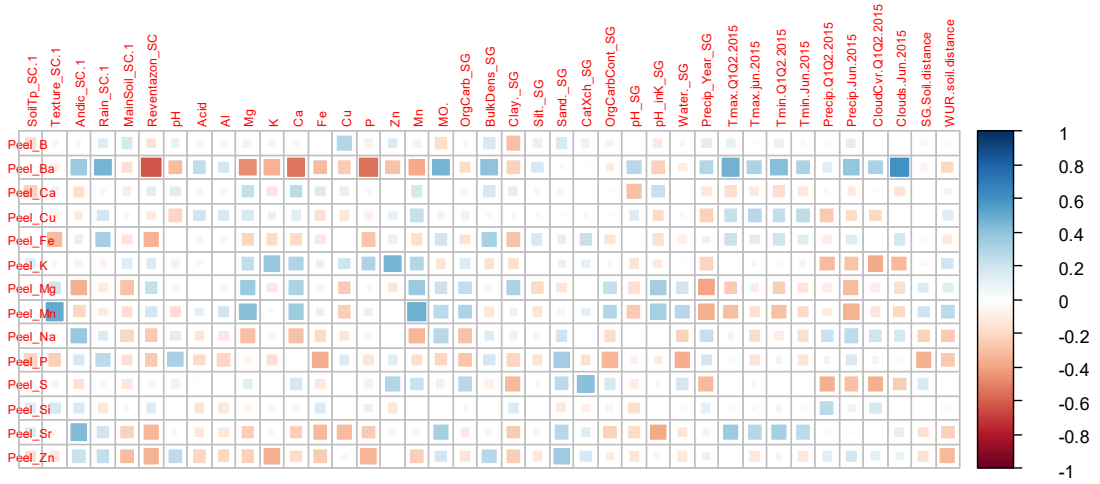

Figure A6 *Correlation plot between ICP elements in banana peel samples and known data – analytical and public database-extracts- on the farms of origin.*

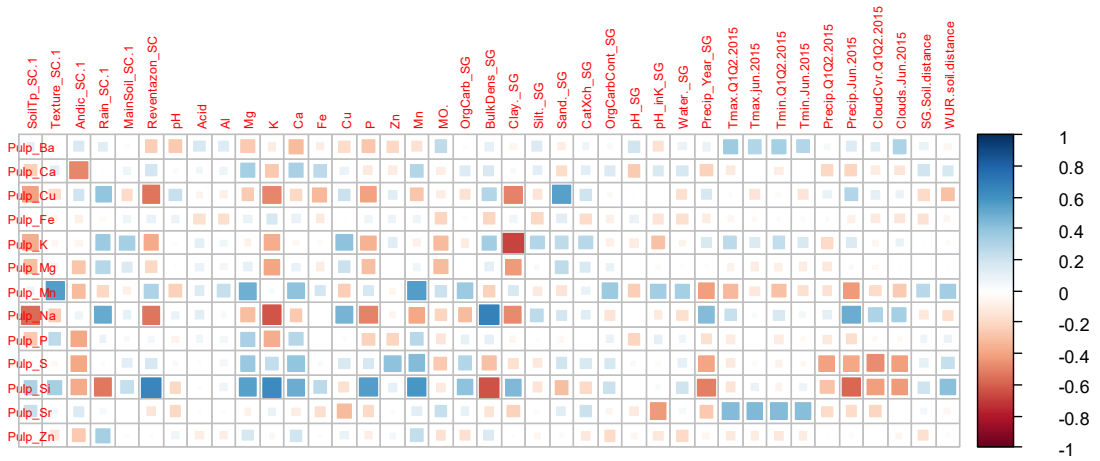

Figure A7 *Correlation plot between ICP elements in banana pulp samples and known data – analytical and public database-extracts- on the farms of origin.*

Wageningen Food Safety Research
P.O. Box 230
6700 AE Wageningen
The Netherlands
T +31 (0)317 48 02 56
www.wur.eu/food-safety-research

WFSR report 2020.006

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 5,000 employees and 12,000 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.

To explore the potential of nature to improve the quality of life

Wageningen Food Safety Research
P.O. Box 230
6700 AE Wageningen
The Netherlands
T +31 (0)317 48 02 56
www.wur.eu/food-safety-research

WFSR report 2020.006

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 5,000 employees and 12,000 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.