*Manuscript
Click here to download Manuscript: Manuscript.docx Click here to view linked References

Identifying the origin of groundwater Samples in a Multi-Layer Aquifer System with random Forest Classification

Paul Baudron[1-4-5*], Francisco Alonso-Sarría[2], José Luís García-Aróstegui[3], Fulgencio Cánovas-García[2], David Martínez-Vicente[1,2], Jesús Moreno-Brotóns[2].

[1] Fundación Instituto Euromediterráneo del Agua, Complejo Campus de Espinardo, Ctra. N301, 30100 Espinardo (Murcia), Spain

[2] University of Murcia, Institute for Water and Environment (INUAMA), Campus de Espinardo, 30100 Murcia (Murcia), Spain

[3] Geological Survey of Spain (IGME), Avda. Miguel de Cervantes, 45 – 5º A, 30009 Murcia (Murcia), Spain

[4] Institut pour la Recherche et le Développement, UMR G-EAU, Cemagref, 361 rue Jean-François Breton, BP 5095, 34196 Montpellier, Cedex 5, France

[5] Université de Paris Sud, UMR CNRS-UPS 8148 IDES, Avenue du Belvédère, Bâtiment 504, Orsay, France

This is an Accepted Manuscript of an article published by ELSEVIER in Journal of Hydrology on 15 Jul 2013, available at:
https://doi.org/10.1016/j.jhydrol.2013.07.009

©2013. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
https://creativecommons.org/licenses/by-nc-nd/4.0/(opens in new tab/window)

*** Corresponding author:** Paul Baudron

- E-mail address: paul.baudron@baudron.com
- Tel.: +33 628 040 391
- Postal address: Institut pour la Recherche et le Développement, UMR G-EAU, Cemagref, 361 rue Jean-François Breton, BP 5095, 34196 Montpellier, Cedex 5, France

*** Email addresses of co-authors:**

- Francisco Alonso-Sarría : alonsarp@um.es
- José Luís García-Aróstegui : j.arostegui@igme.es
- Fulgencio Cánovas-García : fulgencio.canovas@um.es
- David Martínez-Vicente : davidmv@um.es
- Jesús Moreno-Brotóns : jesusmorenobrotons@gmail.com,

Abstract

Accurate identification of the origin of groundwater samples is not always possible in complex multilayered aquifers. This poses a major difficulty for a reliable interpretation of geochemical results. The problem is especially severe when the information on the tubewells design is hard to obtain. This paper shows a supervised classification method based on the Random Forest (RF) machine learning technique to identify the layer from where groundwater samples were extracted. The classification rules were based on the major ion composition of the samples. We applied this method to the Campo de Cartagena multi-layer aquifer system, in southeastern Spain. A large amount of hydrogeochemical data was available, but only a limited fraction of the sampled tubewells included a reliable determination of the borehole design and, consequently, of the aquifer layer being exploited. Added difficulty was the very similar compositions of water samples extracted from different aquifer layers. Moreover, not all groundwater samples included the same geochemical variables. Despite of the difficulty of such a background, the Random Forest classification reached accuracies over 90%. These results were much better than the Linear Discriminant Analysis (LDA) and Decision Trees (CART) supervised classification methods.

From a total of 1,592 samples, 805 proceeded from one unique identified aquifer, 403 proceeded from a possible blend of waters from several aquifers and 279 were of unknown origin. Only 468 of the 805 unique-aquifer samples included all the chemical variables needed to calibrate and validate the models. Finally, 107 of the groundwater samples of unknown origin could be classified. The uncertainty on the identification of training samples was taken in account to enhance the model. Most of the samples that could not be identified had an incomplete dataset.

Keywords: Multi-layer aquifer, Longscreen boreholes, Machine Learning, Random Forest, Hydrogeochemistry, Hydrogeology.

1. Introduction

In complex multi-layer groundwater systems, the correct determination of the origin of a sample is the basic driving condition for a reliable interpretation of geochemical and hydrodynamic results. However, if there is no available information on the tubewell design, this driving condition can be hard to validate (Mayo, 2010). As a consequence, despite the large quantities of geochemical and piezometric data available, only those corresponding to fully documented tubewells should be used for investigation. Hence, there is a need for a tool that could provide an automatic and accurate estimation of the aquifer layer from which a water sample has been extracted. A possibility is to base this tool on geochemical criteria. Such a method must deal with additional difficulties as similar water types, temporal changes in the origin of groundwater, or having different ions analyzed in different samples. Moreover, it should be applicable with common major ion geochemistry. Such a tool could be helpful to all applications of geochemical data in Hydrogeology, as identifying anthropogenic transformation (e.g. Celle-Jeanton et al., 2009), understanding paleoclimates (e.g. Jiráková et al., 2009), determining mineralization processes (e.g. Gillon et al., 2012; Lorenzren et al., 2012), assessing groundwater flow patterns (e.g. Cronin et al., 2005) or calibrating groundwater flow models (e.g. Dahan et al., 2004), among other uses.

Statistical methods have been widely used in hydrology and hydrogeology (e.g.; Adams et al., 2001 ; Lambrakis et al., 2004 ; Cloutier et al., 2008; Daughney et al., 2012). Generally, as a tool to subdivide and classify large hydrogeochemical datasets to facilitate interpretation. They might also be used to estimate mixing proportions (e.g. Valder et al., 2012). The techniques most applied are principal components analysis (PCA) and hierarchical cluster analysis (HCA). These techniques highlight tendencies inside groups of samples, allowing an easier representation of the

80 results. However, these methods show several limitations, like the subjectivity of the criteria
81 defining the classes, or its unsupervised nature. That is, they can be used to create a set of classes
82 out of the whole dataset but they cannot assign samples to a set of a priori classes.

83 In contrast, in the supervised classification approach, the prediction of the output class of any
84 new sample is enabled by a set of decision rules (classification model) defined out of a set of
85 labeled training samples. This approach enables the prediction of the correct output class for any
86 new input case including the same predictor variables. Linear Discriminant Analysis (LDA) is a
87 classical multivariate technique for supervised classification (Vaselli et al., 1997).

88 However, traditional statistical methods have been proven inadequate to identify complex
89 patterns and relationships that could be revealed by more sophisticated procedures (De'ath and
90 Fabricius, 2000). These new procedures include computer intensive machine learning techniques
91 based on recursion, sampling and randomizations (Babovic, 2005, Prasad et al., 2006).

92 Approaches based on decision trees (Breiman et al., 1984) are among the most applied supervised
93 classification methodologies. Random Forest (Breiman, 2001), is the one that have recently
94 received most interest. It combines a large numbers of decision trees (usually 500 to 2000) to
95 obtain a more accurate classification without overfitting the model to a specific dataset.

96 Studies using Decision Trees can be found in Remote Sensing (e.g. Guhimre et al., 2010),
97 Medicine (e.g. Lempitsky et al., 2009), Genetics (e.g. Cutler and Stevens, 2006), Chemistry (e.g.
98 Svetnik et al., 2004), Ecology (e.g. Cutler et al., 2007) or Soil Science (e.g. Schmidt et al., 2008).
99 Only a few studies use supervised classification methods in Hydrogeology. Use of decision trees
100 as a supervised classification method has been limited to the studies by Loos and Elsenber (2011)
101 on the links between overland flow generation and topography, and by Peters et al. (2008) on

102 groundwater-dependent vegetation patterns. LDA has been applied to classify groundwater
103 samples only in rare occasions (e.g. Lambrakis et al., 2004). Other machine learning methods as
104 Neural Networks can be found in Hydrogeology (e.g. Kurtulus and Razack, 2007), but they are
105 more difficult to calibrate and were not used in the present study. Except the recent studies by
106 Smith et al. (2010) on bacterial source tracking in lakes and Olson and Hawking (2012) on
107 stream base-flow water chemistry, we have not been able to find any studies using Random
108 Forest neither in Hydrogeology nor for the analysis of hydro-geochemical datasets.

109 Our main goal was to test the Random Forest classification method to determine the origin of
110 groundwater samples based on their geochemistry. The study was conducted in an intensively
111 irrigated region with hundreds, many of them undocumented, tubewells. These tubewells provide
112 a large geochemical dataset whose interpretation is hazardous due to the lack of design-
113 information. Linear discriminant analysis and a simple classification tree were also used to
114 compare results.

115 **2. Study site**

116 The Campo de Cartagena, in southeastern Spain (Figure 1), is a 1,440 km^2 coastal plain whose
117 elevation ranges between 0 and 200 m a.m.s.l. The climate is semiarid with an average
118 temperature of 18 °C and an average rainfall of about 300 mm per year. High variability is
119 another characteristic of precipitation. Several years have registered values lower than 200 mm
120 and, at the same time, more than 150 mm can be registered during a few days, mainly in spring
121 and autumn. The main consequence is the lack of permanent watercourse, though several
122 ephemeral streams drain the area. Groundwater and the Tagus-Segura water transfer, initiated in
123 1980 to derive water from the Tagus basin to the Segura basin, are the main sources of water

supply (Baudron et al., 2013).

The economy of the area relies on the agro-industrial sector with crops covering 1/3 of the total surface. Due to the low precipitation rate and a lack of permanent surface water, intensive irrigated agriculture has historically been mainly supported by groundwater extraction from the regional multi-layer aquifer system.

Figure 1 : Map of the Study Area, with the location of all registered wells and the geological cross-section of Figure 2.

2.1. Hydrogeological settings

The Campo de Cartagena area corresponds to a Neogene-Quaternary sedimentary basin located on the eastern part of the Betic Cordillera. The permeable sedimentary deposits, with a maximum thickness of 2,000 m, created one of the most important aquifers of the Mediterranean basin (Margat and Vallée, 2000). The main geological and hydrodynamic characteristics of the area, detailed by Jiménez-Martínez et al. (2012), are summarized hereafter. From the Tortonian to the Quaternary, several layers of high-permeability rocks (limestones, sands and conglomerates) were deposited (Figure 2), interlayered with detrital, low-permeability marls. Sands and conglomerates of Tortonian age, organic limestones of Messinian and sandstones deposited during the Pliocene form the three confined layers of the aquifer. The detrital Quaternary sediments form the upper unconfined aquifer. A fifth aquifer, corresponding to slightly evolved Triassic limestone from the substratum, appears locally. A small compartment of the Pliocene aquifer in the Northeastern part, is isolated of the rest of the system by a normal fault. It

All layers are intensively exploited by agriculture, with a maximum estimated extraction of more than 200 hm^3 per year with high temporal variability. Natural recharge is scarce and depends on the extent of the layer's respective outcrop areas. Underlying the crops, the Quaternary aquifer is mainly recharged by the irrigation return flow.

More than 40 years of groundwater survey by the Geological Survey of Spain (Instituto Geológico y Minero de España, IGME) provide a large quantity of geochemical and piezometric data, covering a large spatio-temporal range. Nevertheless, due to the lack of design information for most tubewells, the origin of groundwater samples is usually unclear. Identifying representative samples from each aquifer layer, a basic step for any hydrogeological study, is a difficult task.

Figure 2 : A-A' Geological cross-section of the study area.

3. Dataset and Methods

3.1. Geochemical dataset

The first step in building a supervised classification model is to collect and prepare a "learning" or "training" dataset to be analyzed. It is used to learn how the value of a qualitative variable, or « target variable » (here, the aquifer layer) is related to the values of a set of « predictor » variables (here, the geochemical ions).

3.1.1. Collecting data

The dataset was obtained by collecting geochemical data from a wide variety of sources. More than 80% of the data came from the official groundwater quality surveys performed between the early 1970s and early 2000s by IGME. Complementary data was provided by the River Segura Basin Authority (Confederación Hidrográfica del Segura, CHS), in several sampling campaigns from 2005 to 2008 and from 2010 to present. Additional geochemical results came from research projects conducted by the Universities of Granada (2009, unpublished data), the University of Murcia (2009 and 2011, unpublished data) and the IDES laboratory of the Paris Sud University (2011, unpublished data). Data from unpublished IGME reports and groundwater analysis kindly eased in the field by wells owners was also included. Finally, the dataset was composed by 1,592 groundwater samples (Table 1) collected over a wide range of years, sampling conditions and analytical means and corresponding to different aquifer layers. Most boreholes of the study area are undocumented and were constructed by private owners on their own initiative. Therefore, determination of the borehole design is only available for 300 (15%) of the boreholes. In order to check the descriptions, these 300 boreholes were reviewed in the field.

3.1.2. Review of borehole-design information

In order to determine the original aquifer of each sample, we collected and reviewed all available borehole-design information corresponding to more than 1200 tubewells in the area. Most data came from the last inventory of wells by IGME, started in 1973 and partly updated at the beginning of the 1980s. Complementary partial inventories (e.g. Conesa-García, 1990) were added, as well as technical reports provided by well owners and drilling companies. The review of the data was based on the following criteria: depth, localization of the screen, presence of a cement ring, age and state of conservation of the tubewells, and the water table evolution (when

188 available). No geochemical criterion was considered. Finally, we could establish that 805 out of a
189 total of 1,592 groundwater samples came from a single identified aquifer layer (Table 1).

190 Based on the criteria of the above-described data review, we established an Aquifer Reliability
191 (AR) to weight the reliability of the aquifer assessment for each tubewell. Three levels were
192 defined: A (high reliability), B (medium reliability) and C (low reliability). This index did not
193 take into account any chemical data, but only construction criteria. Indeed, this review of the
194 inventory of wells highlighted that some previous hydrogeological studies could have relied on
195 partially inappropriate aquifer assessment, leading to hazardous interpretations.

196 ### 3.1.3. Water types

197 Based on the 805 samples assigned to just one aquifer layer, the geochemical water-type of each
198 aquifer can be assessed. The Piper diagram for samples with high and medium degree of
199 reliability (AR=A and AR=B) is presented in Figure 3. The Tortonian and Triassic aquifers are
200 well differentiated, with Mg-Na-HCO3 and Ca-SO4 water types, respectively. Nonetheless, the
201 Quaternary, Pliocene and Messinian aquifer are all included in the same Na-Ca-Cl to mixed-Cl
202 water type.

203 This similarity between water types is a strong limitation to the identification of the characteristic
204 geochemical signature of the three upper aquifer layers. Hypotheses to explain this situation are
205 based on: i) the quite similar geological composition of the aquifer compartments, responsible for
206 a similar mineralization of groundwater (confirmed by similar saturation rates); ii) the irrigation
207 return-flow to the Quaternary aquifer, mixing water coming from the lower layers into the upper
208 ones; and iii) the inside-borehole mixings between water masses (Jiménez-Martínez et al., 2011),
209 which could even have reached a regional scale.

The 805 samples from a single identified aquifer were used to calibrate the model. The model was then used to estimate the aquifer of origin of the 279 samples (Table 1) for which no design information was available and had a complete dataset. The geochemical ions (thereafter called variables) selected to perform the classification are the concentrations (expressed in mg/l) of Cl-, SO42-, HCO3-, NO3-, Ca2+, Na+, Mg2+, K+ and SiO2. One of the problems encountered was that the 8 concentrations were not measured in all the samples. Only 468 of the 805 labeled samples included all the variables needed to calibrate and validate the models (Table 1). Minor and trace elements were not taken in consideration due to the very limited number of samples featuring their determination.

Figure 3 : Piper diagram for the training samples from single identified aquifer tubewells with AR=A. 1=Quaternary, 2=Pliocene, 3= Messinian, 4=Tortonian, 5=Triassic.

3.2. Models used

3.2.1. Linear Discriminant Analysis (LDA)

LDA (Vaselli et al.,1997) is one of the most simple methods for supervised classification. It is used to classify samples into mutually exclusive groups on the base of independent variables. This objective is attained by maximizing the between-group variance and minimizing the within-group variance. It is closely related to the unsupervised principal component analysis (PCA) in that they both look for linear combinations of variables that best explain the data. An important assumption of LDA is that the independent variables are normally distributed. If only two variables are available, the separators between the groups will become lines. If three variables are

available, the separator is a plane. When the number of variables is higher than three, the separators become a hyper-plane.

3.2.2. Decision Trees

Decision trees are used to build a model by a recursive binary partition of a labeled dataset into increasingly homogeneous nodes. Homogeneity is measured with the Gini index (Breiman et al., 1984), defined as $G = \Sigma k\ p_k \cdot (1 - p_k)$, where p_k is the proportion of observations in the k^{th} class. This index is minimized when all observations belong to the same class. At each step the node with the highest G value is split; a an optimization is done to select the predictor variable and the numeric threshold, or group of values is the variable is categorical, that would produce the lowest G value in the subsequent nodes The splitting process continues until no further subdivision can reduce the Gini index (Cutler et al., 2007).

The final result should be a fully-grown classification tree whose lower nodes include cases belonging to just one class. However, the lower nodes are seldom, if ever, completely homogeneous. In this case, the predominant class is used to label the node, being the other cases classification errors. On the basis of these errors it is possible to prune the tree to allow a higher generalisation capacity. A typical pruned classification tree has 3 to 12 terminal nodes. This trained decision tree can then be used to classify an unlabeled dataset. Interpretation of classification trees increases in complexity as the number of terminal nodes increases (Cutler et al., 2007).

3.2.3. Ensemble Learning

The main problem of classifying with a unique tree is its high sensitivity to the input data, small modifications in the dataset can produce completely different models. Ensemble Learning

techniques have recently received much interest as a tool to overcome this limitation of decision trees, to obtain better predictive performance.

Bagging (Breiman, 1994) is one of the most used ensemble learning methods. It generates independent trees by re-sampling the same dataset by bootstrapping. That is generating new datasets of the same size as the initial one by random sampling with replacement. Around 67% of the original observations occur at least once in each new generated dataset. Observations not included in any of the new datasets are called "out-of-bag" observations. The trees obtained are not pruned and are used to classify the out-of-bag observations. As each initial observations is included inside the out-of-bag of several trees, its class is estimated several times. The final estimation assigns each observation to the most "voted" class (Liaw and Wiener, 2002).

3.2.4. Random Forest (RF)

Random Forest (RF) is a bagging based method proposed by Breiman (2001). It generates several trees (500 to 2,000) using bootstrapping; each tree is then trained using a randomized subset of the predictors. This somewhat anti-intuitive modification adds randomness to bagging and decreases the correlation between trees. Uncorrelation is a desiderable property in ensemble learning classifiers to guarantee that different results give sense to the voting system. Random Forest produces very good results compared to other machine learning based classification systems (Support Vector Machines or Neural Networks) or to other decision tree algorithms (Breiman, 2001; Liaw and Wiener, 2002).

Random Forests do not overfit the model to the dataset since the classification error of one permutation can be overcome by the ensemble of permutations (Ghimire et al., 2010). This way the large number of trees reduces generalization error (Breiman, 2001; Pal, 2005 ; Prasad et al.,

276 2006). Since the out-of-bag observations are not used in the fitting of the trees, the out-of-bag
277 estimates can be used to perform a cross-validation accuracy estimation (Cutler et al., 2007).

278 One of the parameters that can be set up by the user is the number of variables included in each
279 classification tree. Nevertheless, the method does not seem to be very sensitive to this value,
280 which is by default the square root of the total number of variable used (Gislason et al., 2006).
281 Another user configurable parameter is the number of generated trees, although a higher number
282 does not seem to provide a substantial increase in the classification accuracy (Liaw and Wiener,
283 2002). In general, random forests do remarkably well and require very little tuning (Hastie et al.,
284 2003).

285 A disadvantage of Random Forest compared to the simple classification tree approach is that
286 individual trees cannot be examined separately, thus becoming a "black box" approach (Prasad et
287 al., 2006). However, it does provide several metrics that help in interpretation. Variable
288 importance is evaluated based on how much worse the prediction would be if the data for that
289 predictor were permuted randomly. The resulting values can be used to compare relative
290 importance among predictor variables. In this way, the procedure is much more interpretable than
291 methods such as Neural Networks, and it has been called a "grey box" approach (Prasad et al.,
292 2006).

293 #### 3.2.5. Validation

294 Random Forest includes its own cross validation procedure (out-of-bag cross validation). While
295 some authors consider it unnecessary to perform a separate cross-validation (Efron and
296 Tibshirani, 1997; Breiman, 2001; Svetnik et al., 2004), others like Mitchell (2011) affirm that this
297 internal cross validation can generate biases in the classification. Although it is computationally

more intensive, we preferred to perform a separate leave-one-out cross validation to compare Random Forest results with other methods' (LDA and decision trees) with the same validation tool.

The results of a cross-validation are organized in a confusion matrix (Table 2) where columns (j) correspond to real classes, and lines (i) show the model results. Each element of the n_{ij} matrix represents the number of observations corresponding to class j that were classified as class i. Several indices measuring the accuracy of the classification can be generated from the confusion matrix (Congalton and Green, 2008). The overall accuracy is the proportion of cases in the principal diagonal. The omission error of class i is the proportion of cases from class i not classified as such. The commission error of class i is the proportion of cases incorrectly classified as class i. Finally, the kappa index corrects the overall accuracy for random chance agreement as detailed in Congalton and Green (2008):

$$K = \frac{n\sum_{i=1}^{J} n_{ii} - \sum_{i=1}^{J} n_{i+}n_{+i}}{n^2 - \sum_{i=1}^{J} n_{i+}n_{+i}}$$

3.3. Formulation of tested models

A general schema of the methodology used in this study is shown in Figure 4. Five models were tested:

1. Linear Discriminant Analysis (LDA)
2. Classification Tree using the CART algorithm (CART)
3. Random Forest (RF0)
4. Random Forest eliminating unreliable samples (RF1)
5. Random Forest eliminating variables to increase accuracy (RF2)

Several algorithms to apply classification tress have been proposed. In this study we have used the Classification and Regression Trees (CART) proposed by Breiman et al. (1984).

Model RF1 was an attempt to increase the accuracy of the results by the detection and elimination of unreliable samples (Figure 4). Two strategies were applied. First, the ionic balance for each water sample was calculated to determine if errors in classification could be related with errors in the balance. Secondly, we considered the qualitative evaluation of the reliability of the initial aquifer assessment (AR) for each borehole. As for the previous case, the aim was to assess whether unreliable ground water samples decreased the accuracy of the classifications.

6. The purpose of model RF2 was to deal with the decrease in accuracy observed when the number of variables reaches a certain threshold. This phenomenon is known as Hugues effect, or Curse of Dimensionality (Hugues, 1968). It can be attributed to a significant reduction of the sample density in the space of variables as the increase in the number of variables is not compensated by an increase in the sample size. Several models were built to analyze this phenomenon and check its effects. We started with the simplest model, with only the most important variable, using Random Forest variable ordination. Then, we added the variable that most increased the accuracy of the model. Because of the random behavior of the Random Forest, the selection of this variable was not based on only one classification but on 50 different classifications of each new generated model. Thus, we obtained the corresponding distribution of accuracy parameters, in this case calculated using out-of-bag cross-validation instead of leave-one-out cross validation to save computing resources and because in this case different Random Forest results are being compared. Using a similar procedure, the other

variables were progressively added. The expected result was a fast increase in accuracy when adding the first variables, followed by a stabilization or even a decrease in accuracy, due to the Hugues effect with the incorporation of the less important variables.

3.3.1.

The work was carried out with the R programming language (R Development Core Team, 2010) using the R packages *rpart* (Therneau et al., 2011) and *randomforest* (Liaw and Wiener, 2002) that implement the CART and Random Forest algorithms, respectively.

Figure 4 : Methodological scheme

4. Results and discussion

4.1. Linear Discriminant Analysis (LDA)

Table 3 shows the results of the LDA classification. Overall accuracy reaches 84.8% with a kappa index of 0.764. Despite of these significantly high values, omission and commission errors reach 100% for the Pliocene aquifer. This means that no sample from the Pliocene was classified as such and that all samples classified as Pliocene were incorrectly classified.

4.2. Classification and Regression Trees (CART)

According to the confusion matrix (Table 4), the results are quite good, with an overall accuracy of 88%. As for LDA, the omission error reaches 100% in the case of the Pliocene aquifer;

however, the commission error was 0%, meaning that no sample was classified as Pliocene.

In the decision tree produced (Figure 5), each one of the 8 internal nodes is defined by a condition. The sample continues on the left branch if this condition is fulfilled and on the right branch if not. The 9 final nodes correspond to the 5 layers, except Pliocene which, as has been said, did not receive any observation. Figures 6 to 8 illustrate and explain the main geochemical nodes of the classification tree obtained. They show how the first decision rules split the space of the variables into a set of different subregions corresponding to different aquifers. Another way to display the nodes is to use a binary axis. In the space defined by NO_3^- and Ca^{2+} (Figure 6), a high number of Quaternary samples were correctly classified because of NO_3^- concentrations above 44 mg/l. The number of badly classified samples was 6 from the Pliocene aquifer, 6 from the Messinian and 2 from the Triassic aquifer. One possible explanation could be a mixing with Quaternary water with high contents in NO_3^-. As well, all samples with less than 44.0 mg/l of NO_3^- and less than 55.5 mg/l of Ca^{2+} are directly classified as Tortonian. These include 92% of the Tortonian samples and 3 samples coming from other aquifers, therefore badly classified. The samples with less than 44.0 mg/l of NO_3^- and more than 55.5 mg/l of Ca^{2+} generate two sub-trees that are analyzed in Figure 7 and Figure 8. The first sub-tree involves samples with less than 44.0 mg/l of NO_3^- and Ca^{2+} between 55.5 mg/l and 277.5 mg/l (Figure 7). The definitive assignation (Messinian or Quaternary) of the samples is based on the Mg^{2+} and Cl^- contents. The second sub-tree (Figure 8) includes three variables: Cl^- on the abscissa, HCO_3^- on the ordinate and the threshold in Cl^-, highlighted by the size of the points. Figure 8 also shows how successful the classification in this part of the tree is, with only one error for the Triassic aquifer and three for the Pliocene, probably partly explained by the mixing process cited above.

The confusion between Pliocene and Quaternary can be explained by Quaternary nitrate-rich

383 (NO_3^- <44.0 mg/l) water entering the Pliocene through long-screen boreholes, and in some cases
384 with high Cl^- as well. The confusion with the Messinian seems to be linked to the same problem,
385 but it also has to be taken in account that both sample types are located in the same regions of the
386 space of the variables.

387

388 **Figure 5 : Classification tree generated by CART**

389 **Figure 6 : Plot of the NO_3^- < 44, Ca^{2+} >= 55.5 and Ca^{2+} < 277.5 nodes of the decision tree**
390 **obtained by the CART model. Note: values in mg/l**

391 **Figure 7 : Plot of the Mg^{2+} < 179.5 and Cl^- < 1024 nodes of the decision tree (CART model).**
392 **Note: values in mg/l**

393 **Figure 8 : Plot of the Cl^- >= 716, HCO_3^- < 542.5 and K^+ >= 20.5 nodes of the decision tree**
394 **(CART model). Note: values in mg/l**

395

396 **4.3. Random Forest (RF0)**

397 The confusion matrix after applying Random Forest to the whole dataset and all the available
398 variables is shown in Table 5 together with its analysis. Compared to the CART model, overall
399 accuracy increased from 88.0% to 90.6%, i.e. 21.7% of the total scope of improvement. The
400 omission error for the Pliocene aquifer decreased from 100% to 70.0%, showing a clear
401 enhancement, although this value remains high. The commission error for the Pliocene is also
402 high (40.0%). Table 6 shows the importance of each variable according to one of the Random

403 Forest importance criteria: the increase in accuracy provided by this variable to the classification.

404 **4.4. Random Forest after elimination of unreliable samples (RF1)**

405 Figure 9 shows the distribution of the ionic balance error (absolute values) for each result of the
406 model. The actual and the estimated classes appear separated by a dot on the horizontal axis.
407 Well-classified samples (1.1, 2.2, 3.3, 4.4, 5.5) seem to have a lower ionic balance error than
408 badly classified samples. Nevertheless, overlapping areas between categories are very large, so
409 no clear threshold can be assessed. As a general rule, we decided to eliminate all samples above
410 5% of absolute ionic balance error (2 samples) for the calibration of the model. Anyway, there is
411 a slight tendency to obtain better classifications in samples with a low error in the ionic balance.
412 We think this supports the use of this classification method.

413

414 **Figure 9 : Absolute Ionic Balance Error**

415

416 Depending on the borehole, the initial aquifer assigned to the samples is more or less reliable.
417 The Aquifer Reliability (AR) expresses three levels of confidence: A (high), B (medium) and C
418 (low). Figure 10 shows the distribution of AR for each one of the 25 possible classification cases
419 (correct and incorrect). It is organised as a confusion matrix in which the elements are pie charts
420 displaying, for each combination of real and estimated classes, AR distribution. The colors are
421 chosen as follows: green (AR=”A”), yellow (AR=”B”) and red (AR=”C”) while the number in
422 brackets indicates the number of cases. Nine combinations never occur (white circles); for
423 example, samples from the Quaternary aquifer wrongly classified as Tortonian. In most cases of

bad classification, a predominance of low reliability initial aquifer assignment (AR="C") is found. Specifically for Quaternary and Pliocene samples incorrectly classified as Messinian, most samples feature a highly reliable initial aquifer assignment (AR="A"). In some cases (Pliocene, Messinian and Trias), well-classified samples present relatively high percentages of medium and low AR. In view of these results, it was decided to eliminate the samples featuring low AR.

After eliminating all samples with AR="C", together with those with an ionic balance error higher than 5, the classification accuracy increases (Table 7). Especially relevant is the decrease from 70% to 48% in the omission error of the Pliocene aquifer. The commission error for the same layer reaches a reasonable value of 13.3%. Overall accuracy, increases from 90.6% to 93%.

Figure 10 : Distribution of reliability index (ARI) for the different combinations of actual and classified aquifers

4.5. Random Forest after elimination of variables (RF2)

To assess if any of the variables was producing a decrease in accuracy, different models were generated by adding and eliminating variables. We started with a model containing only NO_3^-, the variable that had obtained the higher importance in the RF0 model. This one-variable model reached an accuracy of 71%. Then we check which variable produced the highest increase in accuracy. It, turned out to be Ca^{2+}. This two-variable model attained an accuracy of 81%. Continuing step by step with the same procedure, a model including all the variables was obtained (Figure 11). Due to the random behavior of Random Forest, the results can vary from

one run to another. Therefore, the protocol was repeated 50 times for each model. The accuracy results were obtained by out-of-bag cross-validation.

Figure 11 : Accuracy of different models generated by adding and eliminating variables

A decrease in accuracy was observed after adding the last variable (Cl^-). Eliminating chloride therefore improved the model (Table 8): a reliability of 94.3% was reached with a decrease of the commission error for all classes. Specially important is the decrease in the omission error of the Pliocene aquifer, from 48% to 40%. Removing Cl^- also seemed to reduce the confusion between Pliocene and Messinian aquifers.

It is interesting to compare the evaluation of variables given by Random Forest with the decision tree generated by the CART model (Figure 5), In the former, NO_3^- remains as the most important variable, Ca^{2+} and Mg^{2+} maintain a fairly high importance. The main difference is the importance that Random Forest gives to Na^+ and the rejection of Cl^- that seems to reduce the confusion between Pliocene and Messinian.

4.6. Statistical models comparison

The similarity between water types was initially expected to be a strong limitation to the identification of the characteristic geochemical signatures of the three upper layers. Another problem is that groundwater mixings between aquifer layers are accumulative over time, producing temporal variation in the geochemistry of groundwater samples. Despite such limitations, RF2 model reaches high accuracy and low omission error for the Pliocene compared

466 to the other methods (Figure 12), Therefore, the RF2 model is selected as the best model. Out of
467 the 171 tubewells of unknown design featuring geochemical data, and based on the training set of
468 73 tubewells, it succeeded to identify the aquifer corresponding to 66 tubewells (Figure 13).

469

470 **Figure 12 : Accuracy indicators for the different models: LDA, CART, RF, RF1 and RF2.**

471 **Figure 13 : Map of the RF2 results**

472

473 **4.7. Predictive capacities of the model**

474 The results of the RF2 model for unknown samples are represented in a Piper diagram (Figure
475 14). Piper diagrams display geochemical water-types, i.e. the relative proportion of several
476 geochemical species in the total mineralization of a sample. They use slightly different data than
477 the used by Random Forest. First, data appear as percentages whereas the model is built on
478 concentration values; secondly, some of the ions appear added. So, we think that a Piper diagram
479 of the classified samples can be used as a second validation approach and to check the predictive
480 capacity of the model.

481 The water types displayed on Figure 14 (classified samples) are similar to those showed on
482 Figure 3 (training samples), confirming the reliability of the method to identify the origin of
483 groundwater samples. Two problems, not directly attributable to the model, still appear. Some of
484 the samples could not be identified because not all the ion concentrations had been measured
485 when the samples were collected, making the classification impossible. Secondly, some of the

classified samples could actually represent a mixing between different aquifers layers; this already mentioned phenomenon is characteristic of the study area. Both problems, incomplete datasets and mixing samples, are to be dealt in future works.

Figure 14 : Piper diagram for samples from unknown origin featuring all variables and identified with the RF2 model. 1=Quaternary, 2=Pliocene, 3= Messinian, 4=Tortonian, 5=Triassic.

5. Conclusions and perspectives

Based on training samples featuring all variables, the first two models (LDA and CART) showed overall accuracies of 84.8 and 88.0% (respectively). A high disparity was found between geochemically easy distinguishable aquifer layers (Tortonian, Triassic) and others that present higher geochemical similarity (Quaternary, Pliocene and Messinian). Although these values seem quite acceptable, these models did not succeed to correctly classify any of the Pliocene training samples. With the same dataset, the first Random Forest model (RF) reached slightly higher overall accuracies (90.6%) and succeeded to classify part of the Pliocene samples. The elimination of less-reliable samples, based on both geochemical and tube-well design criteria, provided a stronger Random Forest model (RF1) with exactitude of 93.0%. After eliminating the less useful variables, the final Random Forest model (RF2) achieved an overall accuracy of 94.3% and the best classification.

These good results prove that Random Forest allows to identify the aquifer of origin of groundwater samples based on commonly available major ions geochemistry, even when the different aquifer layers have similar geochemical water types. Random Forest also provide more

accurate classification than LDA or CART. The identification of the aquifer of origin of unknown samples optimizes the hydrogeochemical dataset, enhancing the possibilities of geochemical interpretations. The results of this study present a wide interest limited neither to this kind of problem nor to the Campo de Cartagena aquifer system. Indeed, many multi-layer aquifer systems feature long-screen boreholes, and could benefit from this methodology to increase the geochemical knowledge. More generally, the Random Forest methodology does show potential for a wide range of hydrological, hydrogeological and geochemical applications, and offers novel prospects in this field.

Still, developing several aspects could enhance the present classification model. Firstly, a strategy to identify water samples produced by the mix of groundwater from different layers inside longscreen boreholes would improve the results. Secondly, several samples were not used to calibrate the model because not all the 8 predictor variables had been measured. It would be necessary to check the accuracy of the method with samples with much less information. Thirdly, temporal variability is an accumulative factor that can introduce temporal variation in the geochemistry of samples and, consequently, noise in the models. Finally, the spatial variability of the agricultural activity, and the introduction of NO_3^- in the aquifers, is not the same in the whole area. This spatial variability could be also affecting the models. In forthcoming works, these tracks will be investigated.

Acknowledgments

We thank Dr Christian Leduc (IRD, UMR G-EAU) for his constructive comments on the manuscript. This work was developed within the scope of the Project "Modelación Hidrológica en Zonas Semi Aridas" financed by the Regional Ministry of Universities, Business and Research (Region of Murcia, Spain). The authors acknowledge the Fundación Instituto Euromediterráneo del Agua (Murcia, Spain) for its fundamental financial support. Additional supports came through the "CARTAG-EAU" project financed by the French "SICMED-MISTRALS" initiative and the 08225/PI/08 research project financed by "Programa de Generación del Conocimiento Científico de Excelencia" of the Fundación Seneca, Región de Murcia (II PCTRM 2007-10). Some co-authors thank the SWAM Project ("Increasing Regional Competiveness through RTD&I on Sustainable Water Resources Management", 7FP, Grant Agreement nº 245427), for the international visibility of groundwater research at the Campo de Cartagena case study.

References

Adams, S., Titus, R., Pietersen, K., Tredoux, G., Harris, C., 2001. Hydrochemical characteristics of aquifers near Sutherland in the Western Karoo, South Africa. Journal of Hydrology. 241, 91–103.

Babovic, V., 2005. Data Mining in Hydrology. Hydrological Processes. 19, 1511-1515.

Baudron P, Barbecot F, García-Aróstegui JL, Leduc C, Travi Y, Martinez-Vicente D. 2013. Impacts of human activities on recharge in a multilayered semiarid aquifer (Camp de Cartagena, SE Spain). Hydrological Processes, in press. DOI: 10.1002/hyp.9771

Breiman, L., 1994. Bagging predictors. Technical Report No. 421.

550 Breiman, L., 2001. Random Forests. Machine Learning. 45(1), 5–32.
551 Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees.
552 Wadsworth and Brooks/Cole, Monterey, California, USA.
553 Celle-Jeanton, H., Huneau, F., Travi, Y., Edmunds, W.M., 2009. Twenty years of groundwater
554 evolution in the Triassic sandstone aquifer of Lorraine: Impacts on baseline water quality.
555 Applied Geochemistry. 24, 1198–1213.
556 Cloutier, V., Lefebvre, R., Therrien, R., Savard, M.M., 2008. Multivariate statistical analysis of
557 geochemical data as indicative of the hydrogeochemical evolution of groundwater in a
558 sedimentary rock aquifer system. Journal of Hydrology. 353, 294–313.
559 Conesa-García, C., 1990. El Campo de Cartagena. Clima e hidrología de un medio semiáárido.
560 Universidad de Murcia, Ayuntamiento de Cartagena. Comunidad de Regantes del Campo de
561 Cartagena.
562 Congalton, R.G., Green, K., 2008. Assessing the Accuracy of Remotely Sensed Data. Principles
563 and Practices. CRC Press.
564 Cronin, A.A., Barth, J.A.C., Elliot, T., Kalin, R.M., 2005. Recharge velocity and geochemical
565 evolution for the Permo-Triassic Sherwood Sandstone, Northern Ireland. Journal of Hydrology.
566 315, 308–324.
567 Cutler, A., Stevens, R.J., 2006. Random forests for microarrays. Methods in Enzymology. 411,
568 422–432.
569 Cutler, D.., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007.
570 Random Forest for Classification in Ecology. Ecology. 88(11), 2783–2792.
571 Dahan, O., McGraw, D., Adar, E., Pohll, G., Bohm, B., Thomas, J., 2004. Multi-variable mixing
572 cell model as a calibration and validation tool for hydrogeologic groundwater modeling. Journal
573 of Hydrology. 293, 115–136.

574 Daughney, C., Raiber, M., Moreau-Fournier, M., Morgenstern, U., van der Raaij, R., 2012. Use
575 of hierarchical cluster analysis to assess the representativeness of a baseline groundwater quality
576 monitoring network: comparison of New Zealand's national and regional groundwater
577 monitoring programs. Hydrogeology Journal. 20, 185–200.
578 De'ath, G., Fabricius, K., 2000. Classification and regression trees: a powerful yet simple techn
579 ique for ecological data analysis. Ecology 81, 3178-3192.
580 Efron, B., Tibshirani, R., 1997. Improvements on Cross-Validation: The .632+ Bootstrap
581 Method. Journal of the American Statistical Association. 92, pp. 548–560.
582 Ghimire, B., Rogan, J., Miller, J., 2010. Contextual land-cover classification: incorporating
583 spatial dependence in land-cover lassification models using random forests and the Getis statistic.
584 Remote Sensing Letters. 1:1, 45–54.
585 Gillon, M., Renard, F., Crancon, P., Aupiais, J., 2012. Kinetics of incongruent dissolution of
586 carbonates in a Chalk aquifer using reverse flow modelling. Journal of Hydrology. 420, 329–
587 339.
588 Gislason, P.O., Benediktsson, L.A., Sveinsson, J.R., 2006. Random Forests for land cover
589 classication. Pattern Recognition Letters. 27, 294–300.
590 Guhimre, B., Rogan, J., Miller, J., 2010. Contextual land-cover classification: incorporating
591 spatial dependence in land-cover classification models using random forests and the Getis
592 statistic. Remote Sensing Letters. 1:1, 45–54.
593 Hastie, T., Tibshirani, R., Friedman, J., 2003. The Elements of Statistical Learning: Data Mining,
594 Inference, and Prediction. Springer.
595 Hughes, G.F., 1968. On The Mean Accuracy Of Statistical Pattern Recognizers. IEEE Trans. on
596 Information Theory 14-1, 55–63.

597 Jiménez-Martínez, J., Aravena, R., Candela, L., 2011. The Role of Leaky Boreholes in the
598 Contamination of a Regional Confined Aquifer. A Case Study: The Campo de Cartagena
599 Region, Spain. Water Air Soil Pollut. 215, 311–327.
600 Jiménez-Martínez, J., Candela, L., García-Aróstegui, J.L., Aragón, R., 2012. A quasi 3D
601 geological model of the Campo de Cartagena, SE Spain: Hydrogeological implications.
602 Geologica Acta. 10(2), 1–13.
603 Jiráková, H., Huneau, F., Celle-Jeanton, H., Hrkal, Z., Coustumer, P.L., 2009. Palaeorecharge
604 conditions of the deep aquifers of the Northern Aquitaine region (France). Journal of Hydrology.
605 368, 1–16.
606 Kurtulus, B., Razack, M., 2007. Evaluation of the ability of an artificial neural network model to
607 simulate the input-output responses of a large karstic aquifer: the La Rochefoucauld aquifer
608 (Charente, France). Hydrogeol. J. 15, 241–254.
609 Lambrakis, N., Antonakos, A., Panagopoulos, G., 2004. The use of multicomponent statistical
610 analysis in hydrogeological environmental research. Water Research. 38, 1862–1872.
611 Lempitsky, V., Verhoek, M., Noble, J.A., Blake, A., 2009. Functional Imaging and Modeling of
612 the Hear.
613
614 Liaw, A., Wiener, M., 2002. Classication and Regression by randomForest. R News. 2(3), 18–22.
615 Loos, M., Elsenbeer, H., 2011. Topographic controls on overland flow generation in a forest – An
616 ensemble tree approach. Journal of Hydrology. 409, 94–103.
617 Lorenzen, G., Sprenger, C., Baudron, P., Gupta, D., Pekdeger, A., 2012. Origin and dynamics of
618 groundwater salinity in the alluvial plains of western Delhi and adjacent territories of Haryana
619 State, India. Hydrological Processes 26, 2333–2345.

620 Margat, J., and Vallée, D. 2000. Water Resources and Uses in the Mediterranean Countries:
621 Figures and Facts. Blue Plan for the Mediterranean. Regional Activity Centre, Sophia-
622 Antipolis, France, 224 pp.
623 Mayo, A., 2010. Ambient well-bore mixing, aquifer cross-contamination, pumping stress, and
624 water quality from long-screened wells: What is sampled and what is not? Hydrogeology Journal.
625 18, 823–837.
626 Mitchell, M., 2011. Bias of Random Forest Out-of-Bag (OOB) Error for Certain Input
627 Parameters. Open Journal of Statistics. 1, 205–211.
628 Olson, J.R., Hawkins, C.P., 2012. Predicting natural base-flow stream water chemistry in the
629 western United States. Water Resources Research. 48.
630 Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of
631 Remote Sensing. 26, 217–222.
632 Peters, J., Baets, B.D., Samson, R., Verhoest, N.E.C., 2008. Modelling groundwater-dependent
633 vegetation patterns using ensemble learning. Hydrol. Earth Syst. Sci. 12, 603–613.
634 Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques:
635 bagging and random forests for ecological prediction. Ecosystems. 9, 181–199.
636 R Development Core Team, 2010. R: A language and environment for statistical computing. R
637 Foundation for Statistical Computing, Vienna, Austria.
638 Smith, A., Sterba-Boatwright, B., Mott, J., 2010. Novel application of a statistical technique,
639 Random Forests, in a bacterial source tracking study. Water Research. 44, 4067–4076.
640 Svetnik, V., Liaw, A., Tong, C., Wang, T., 2004. Application of Breiman's Random Forest to
641 modeling structure-activity relationships of pharmaceutical molecules, in: MCS. Springer-Verlag,
642 pp. 334–343.

643 Therneau, T.M., Atkinson, B. 2011. rpart: recursive partitioning. R package version 3.1-41. R
644 port by Brian Ripley.
645 Valder, J.F., Long, A.J., Davis, A.D., Kenner, S.J., 2012. Multivariate statistical approach to
646 estimate mixing proportions for unknown end members. Journal of Hydrology 460–461, 65–76.
647 Vaselli, O., Buccianti, A., Siena, C.D., Bini, C., Coradossi,N., Angelone,M, 1997. Geochemical
648 characterization of ophiolitic soils in a temperate climate: a multivariate statistical approach.
649 Geoderma 75, 117-133.
650

651 **Table 1: Summary of the groundwater samples included in the dataset. Most samples**
652 **marked with an asterisk (*) did not feature a full set of variables.**

653 **Table 2: Mathematical illustration of a confusion matrix. Adapted from Congalton and**
654 **Green (2008).**

655 **Table 3 : Confusion matrix of discriminant analysis**

656 **Table 4 : Confusion matrix of CART classification tree**

657 **Table 5 : Confusion matrix of random forest**

658 **Table 6 : Importance of variables in relation to the corresponding accuracy increase**

659 **Table 7 : Confusion matrix of random forest eliminating doubtful samples (RF1)**

660 **Table 8 : Confusion matrix of random forest eliminating Cl^-**

Highlights

- Identification of the origin of groundwater samples based on their geochemistry.
- Enhancement of a geochemical dataset featuring doubtful samples.
- Novel application of Random Forest (RF) machine learning technique in hydrogeology.
- High discrimination capacity, beyond similar water types and heterogeneous dataset.
- Optimization of the classification model by assessing the most useful variables.

Figure 1 : Map of the Study Area, with the location of the registered wells and the geological cross-section of Figure 2.

Figure 2 : A-A' Geological cross-section of the study area.

Figure 3 : Piper diagram for the labelled training samples from single identified aquifer tubewells with AR=A. 1=Quaternary, 2=Pliocene, 3= Messinian, 4=Tortonian, 5=Triassic.

Figure 4 : Methodological scheme

Figure 5 : Classification tree generated by CART

Figure 6 : Plot of the NO_3^- < 44, Ca^{2+} >= 55.5 and Ca^{2+} < 277.5 nodes of the decision tree obtained by the CART model. Note: values in mg/l

Figure 7 : Plot of the Mg^{2+} < 179.5 and Cl^- < 1024 nodes of the decision tree (CART model). Note: values in mg/l

Figure 8 : Plot of the Cl^- >= 716, HCO_3^- < 542.5 and K^+ >= 20.5 nodes of the decision tree (CART model). Note: values in mg/l

Figure 9 : Absolute Ionic Balance Error

Figure 10 : Distribution of reliability index (AR) for the different combinations of actual and classified aquifers

Figure 11 : Accuracy of different models generated by adding and eliminating variables

Figure 12 : Accuracy indicators for the different models: LDA, CART, RF, RF1 and RF2.

Figure 13 : Map of the RF2 results

Figure 14 : Piper diagram for samples from unknown origin featuring all variables and identified with the RF2 model. 1=Quaternary, 2=Pliocene, 3= Messinian, 4=Tortonian, 5=Triassic

Figure 1

Click here to download high resolution image

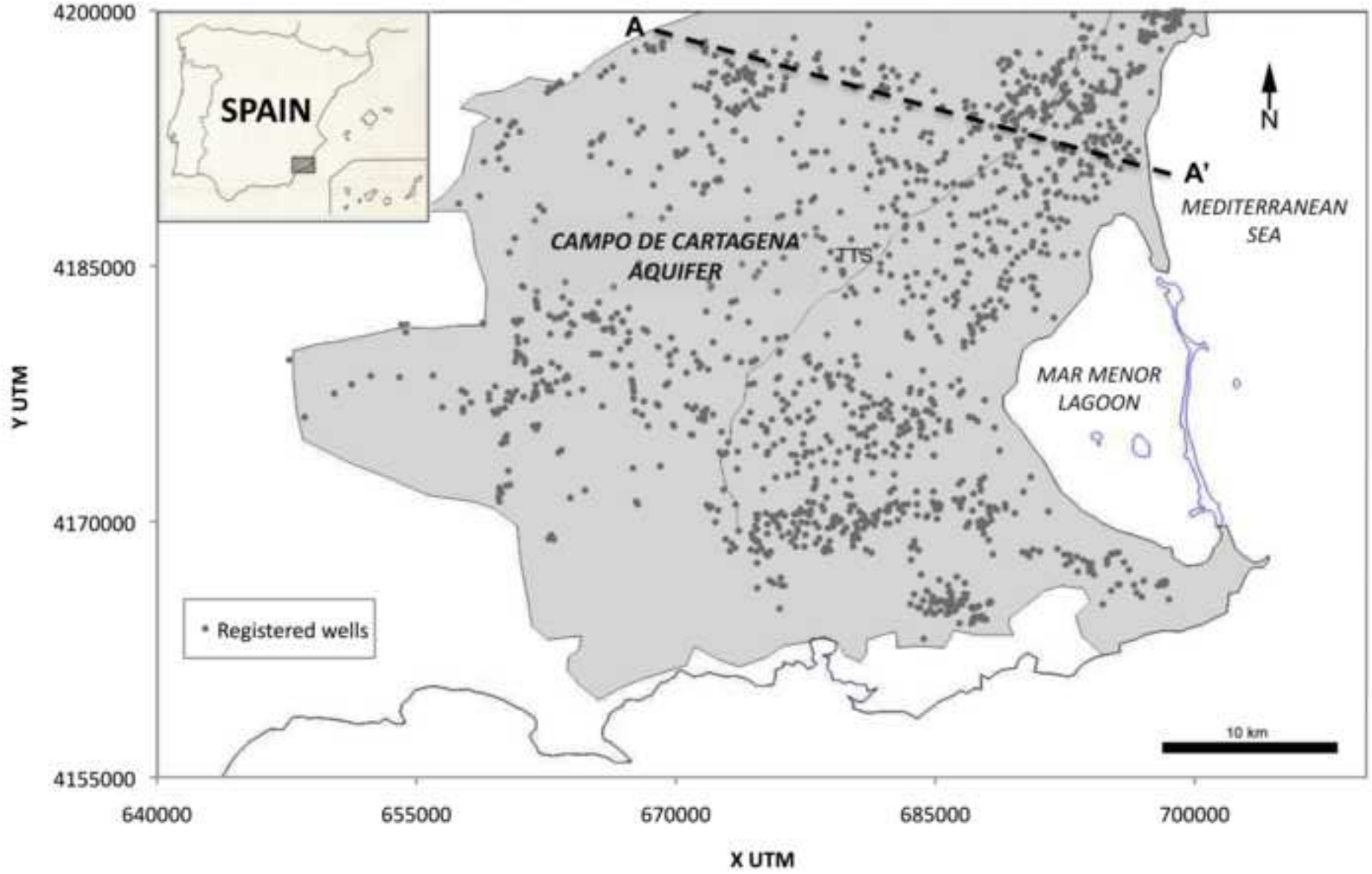

Figure 2

Click here to download high resolution image

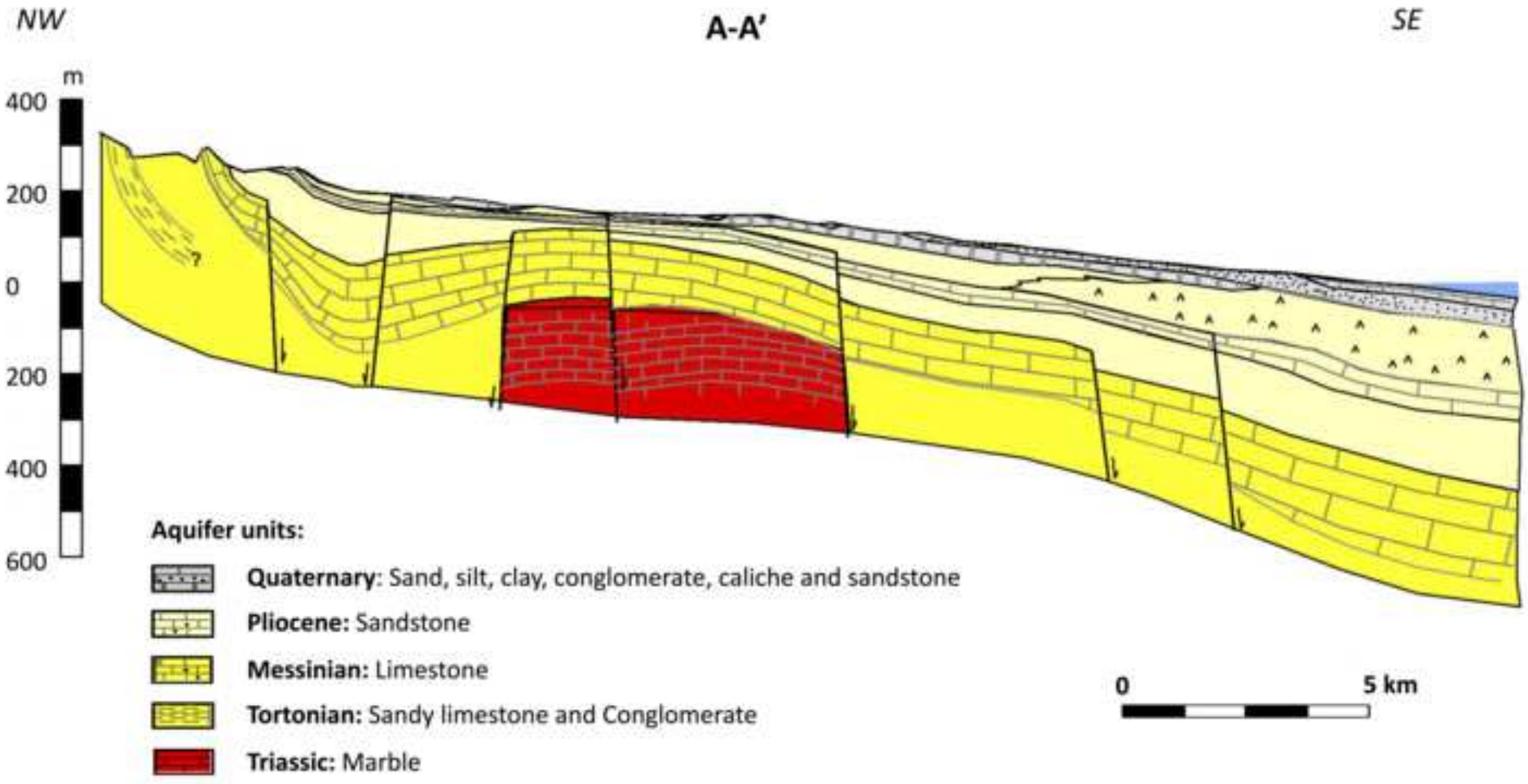

Figure 3

Click here to download high resolution image

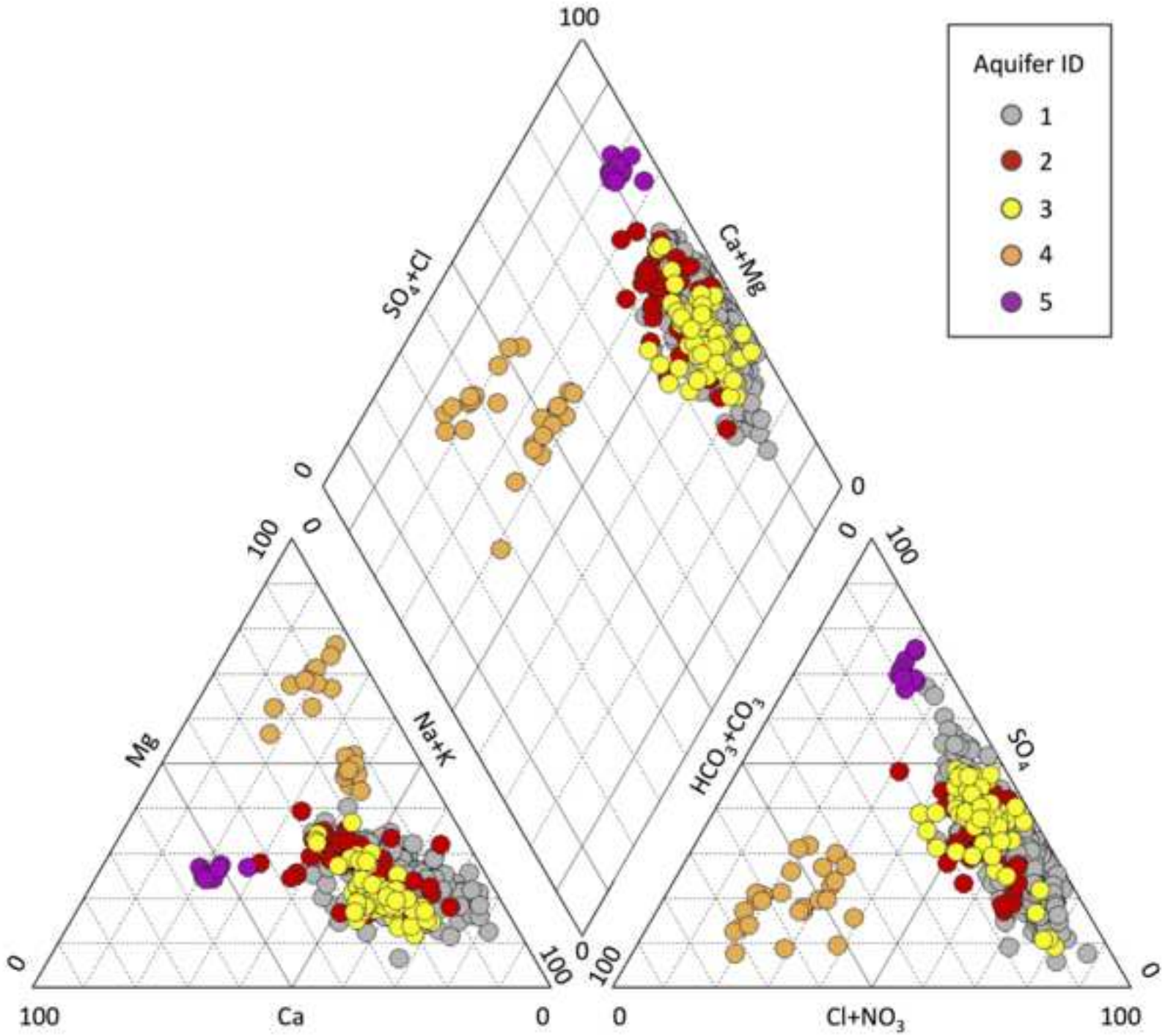

Figure 4

Click here to download high resolution image

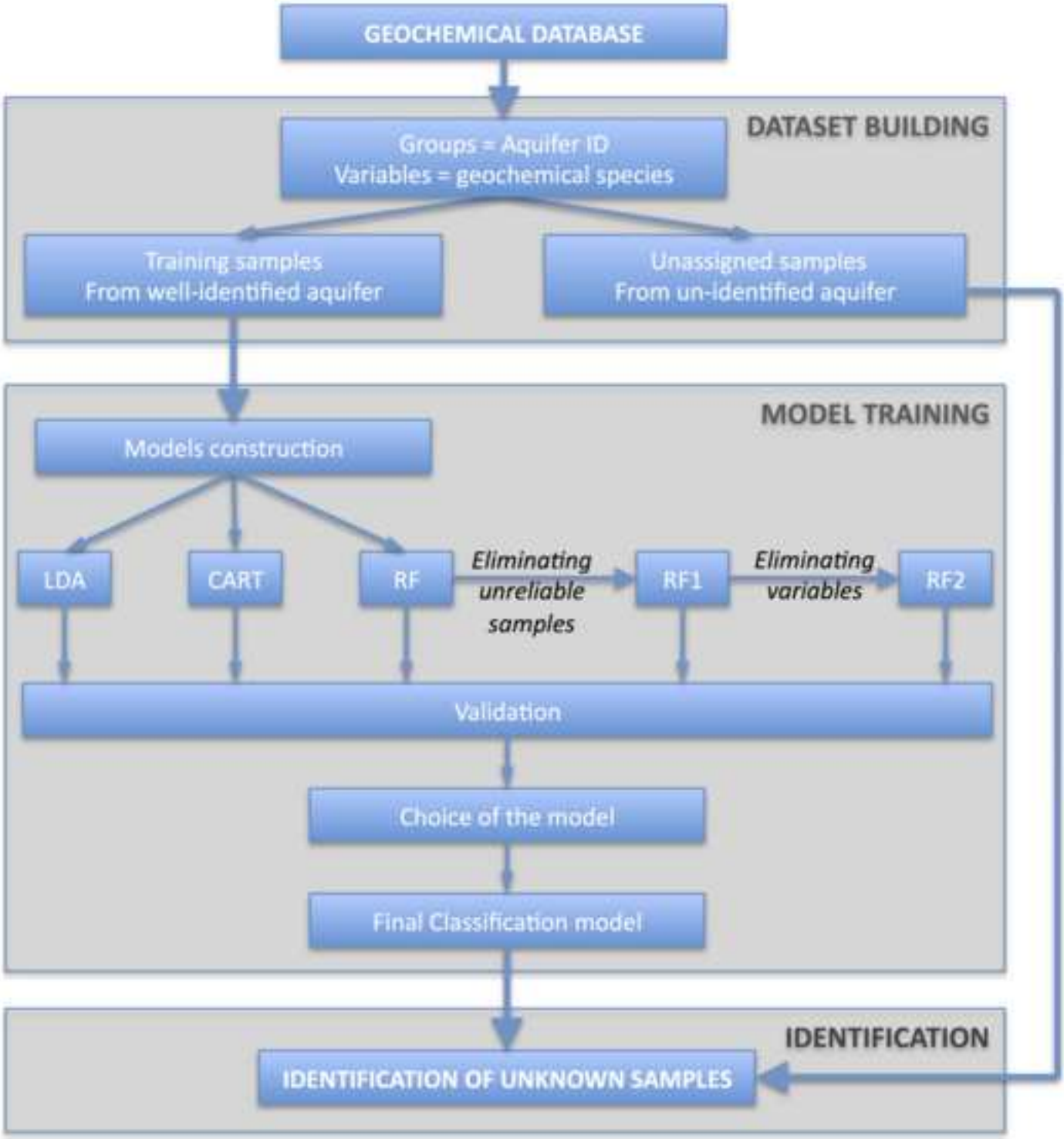

Figure 5

Click here to download high resolution image

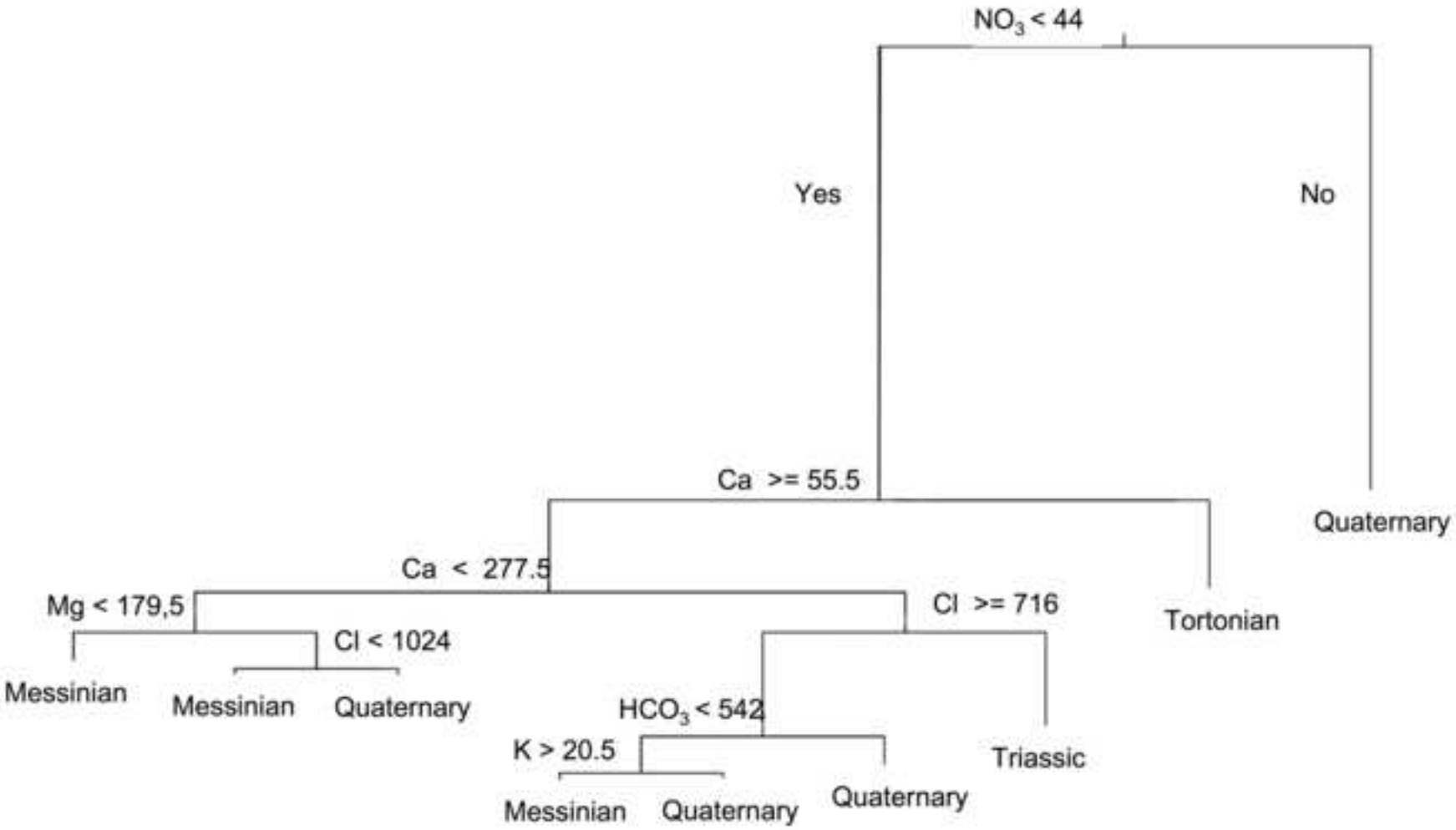

Figure 6

Click here to download high resolution image

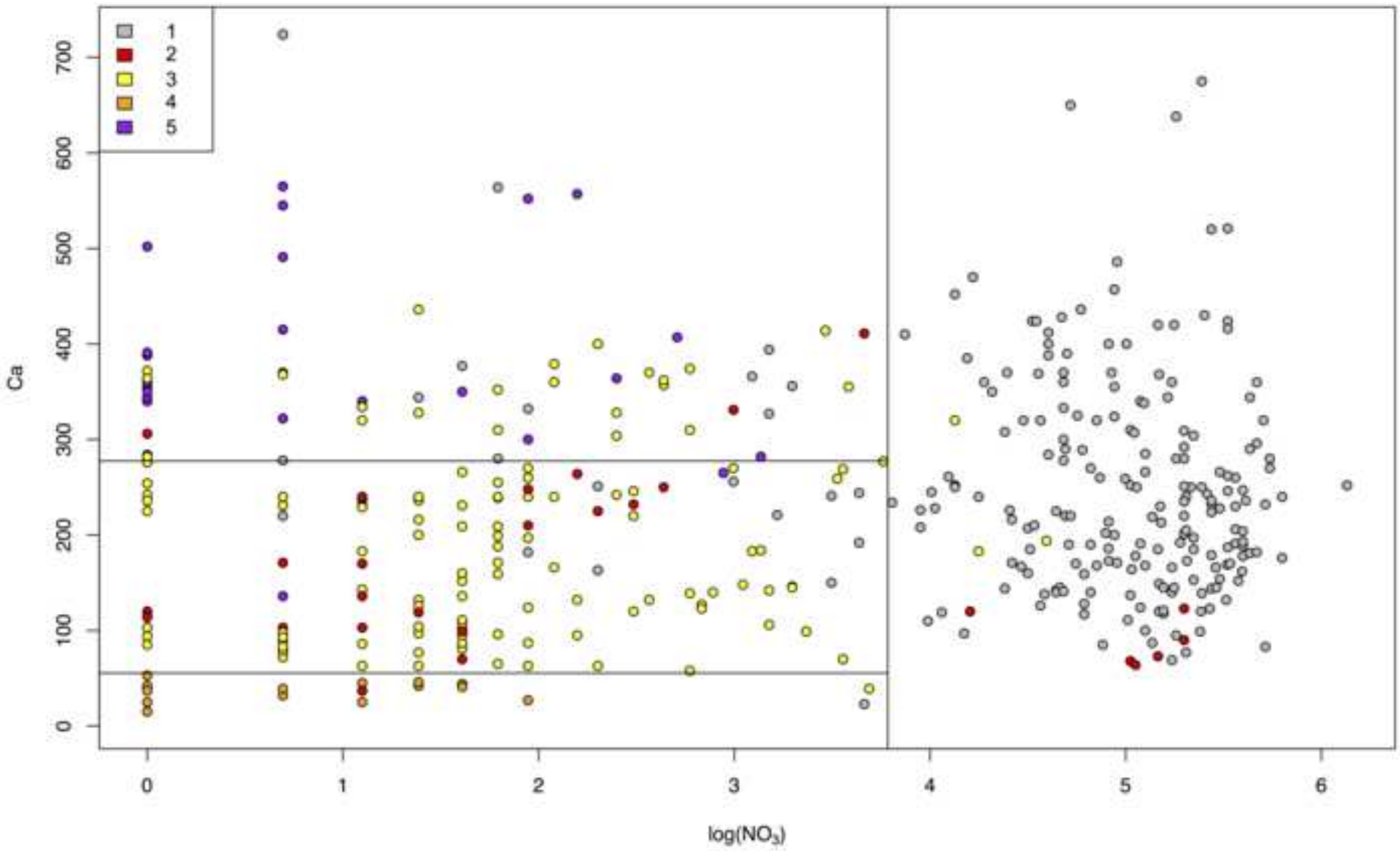

Figure 7

Click here to download high resolution image

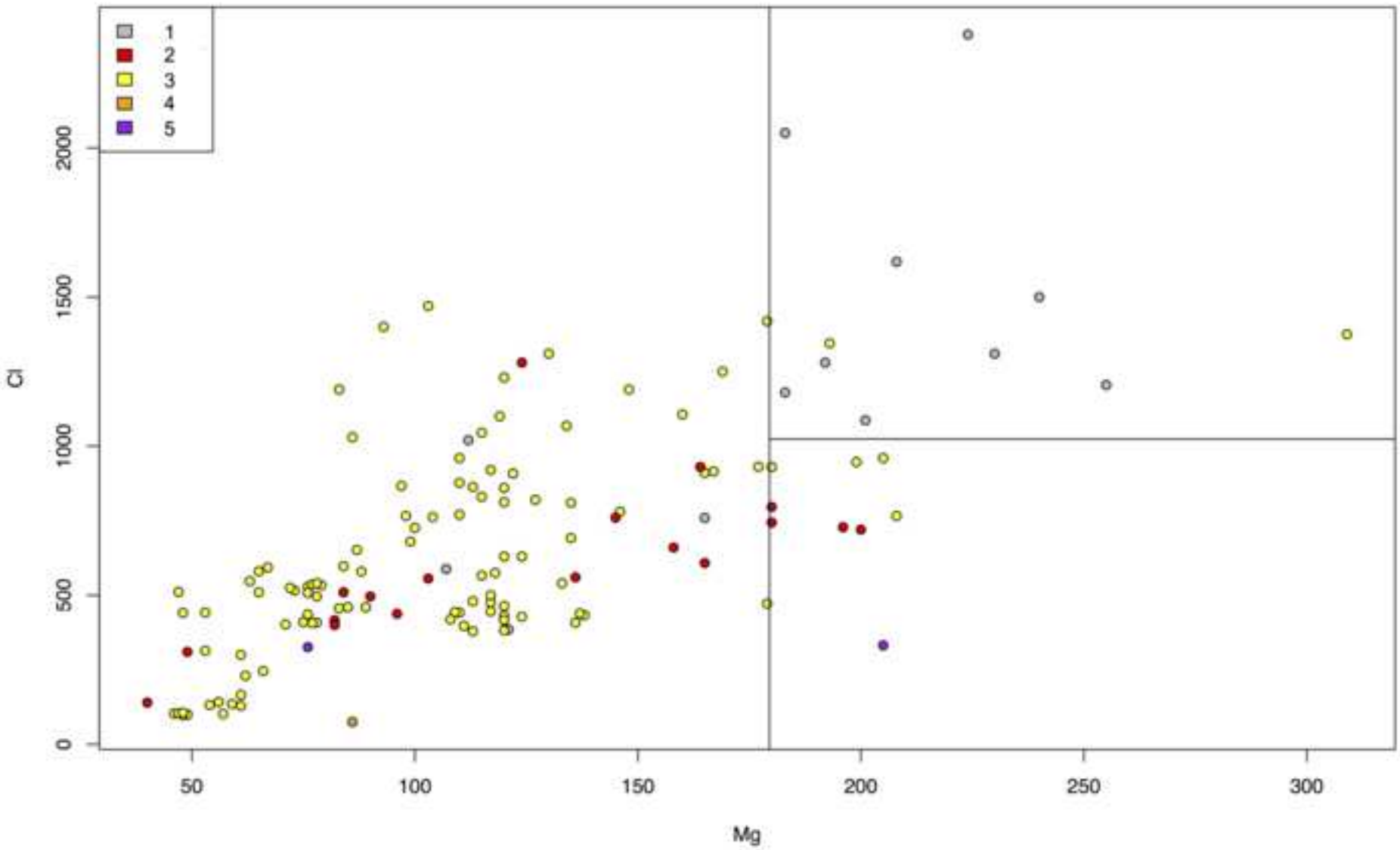

Figure 8

Click here to download high resolution image

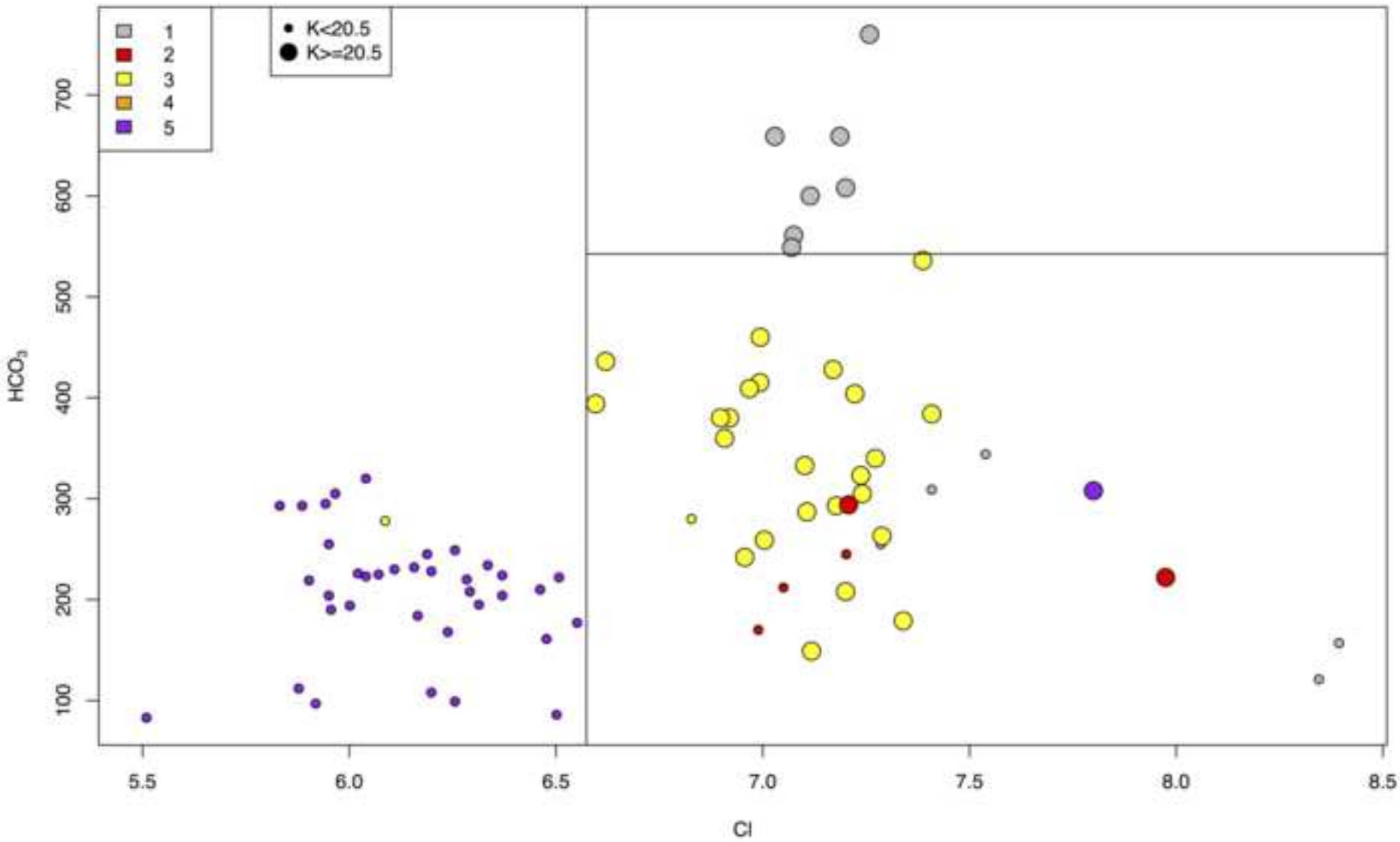

Figure 9

Click here to download high resolution image

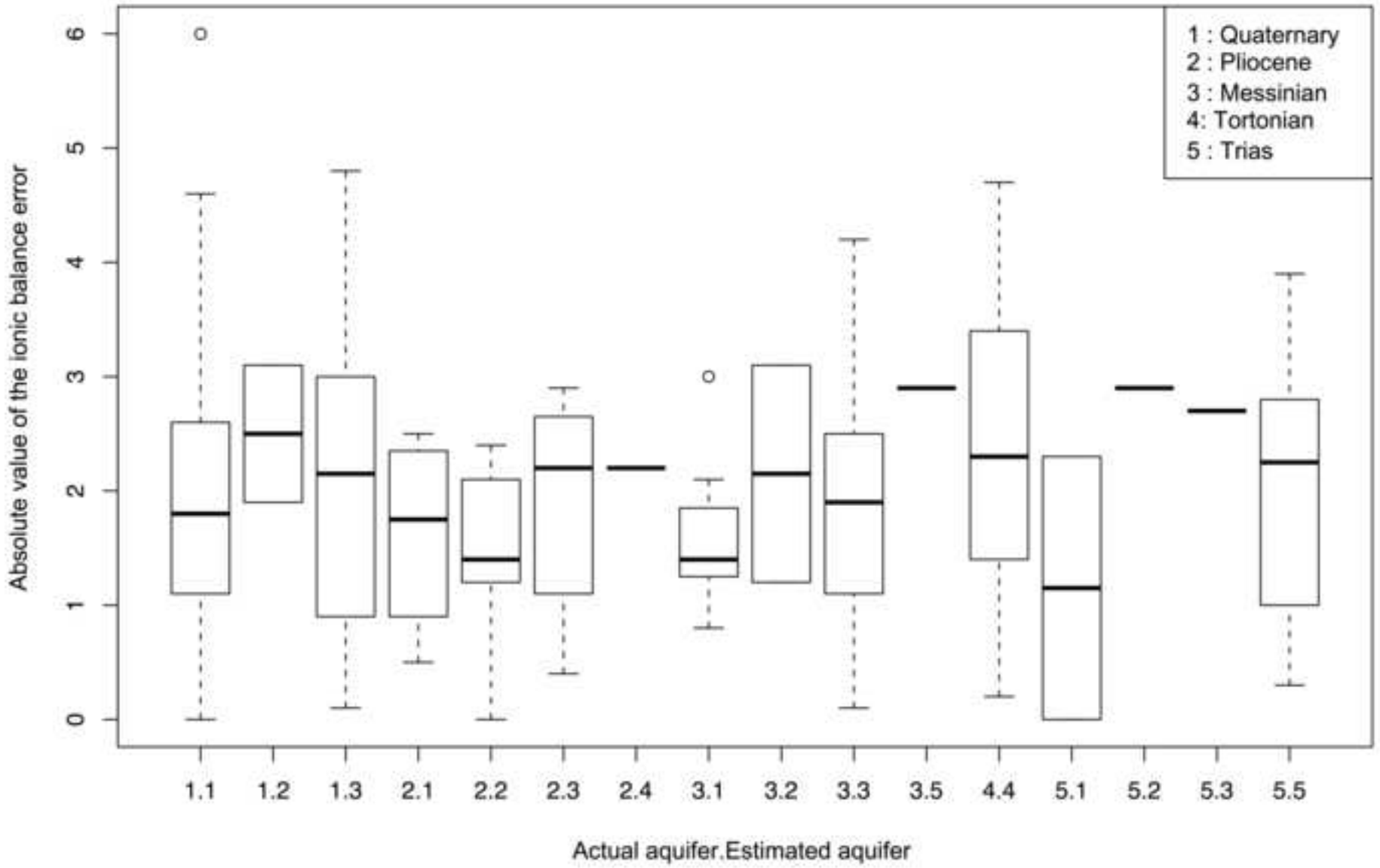

Figure 10

Click here to download high resolution image

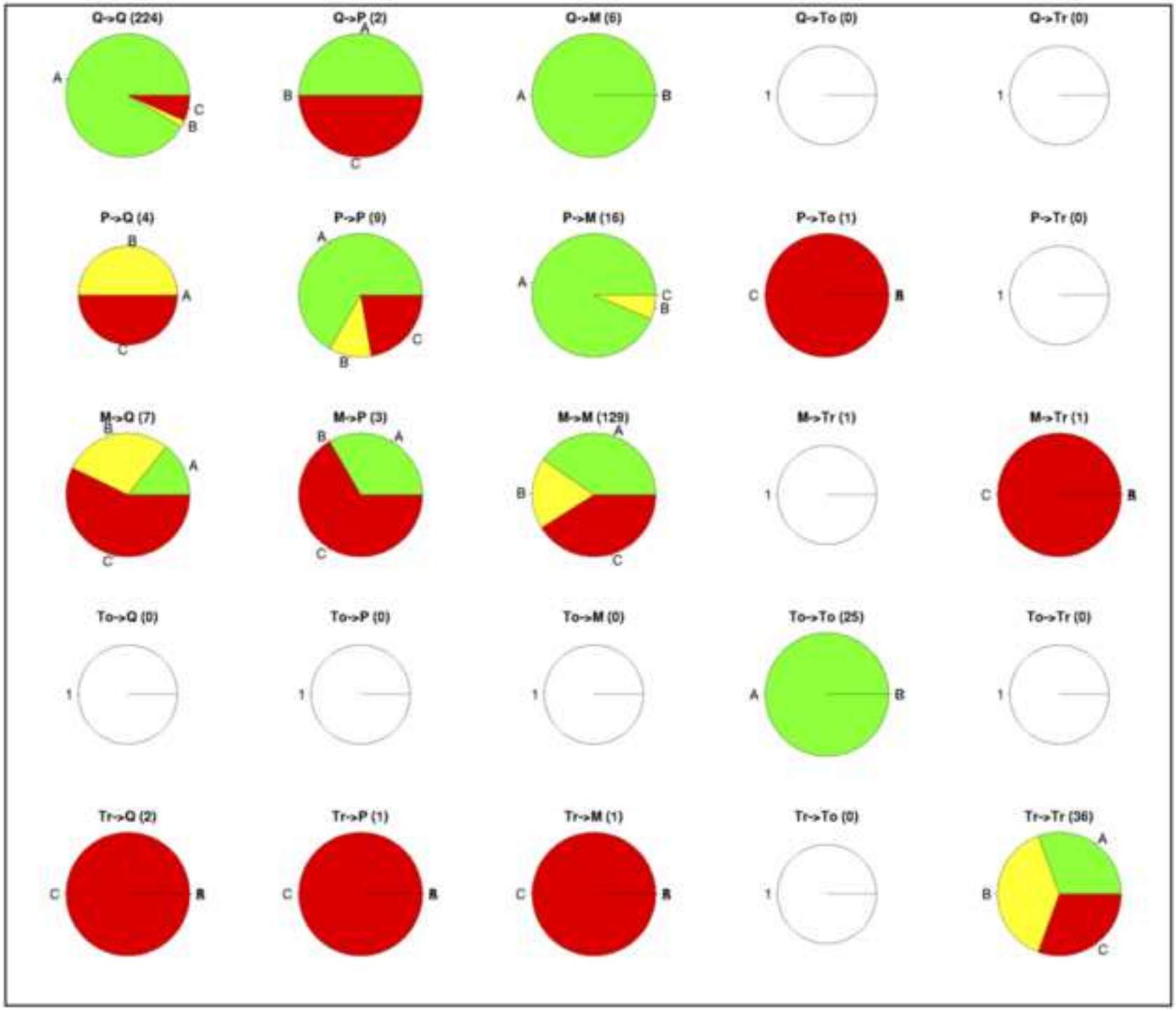

Figure 11

Click here to download high resolution image

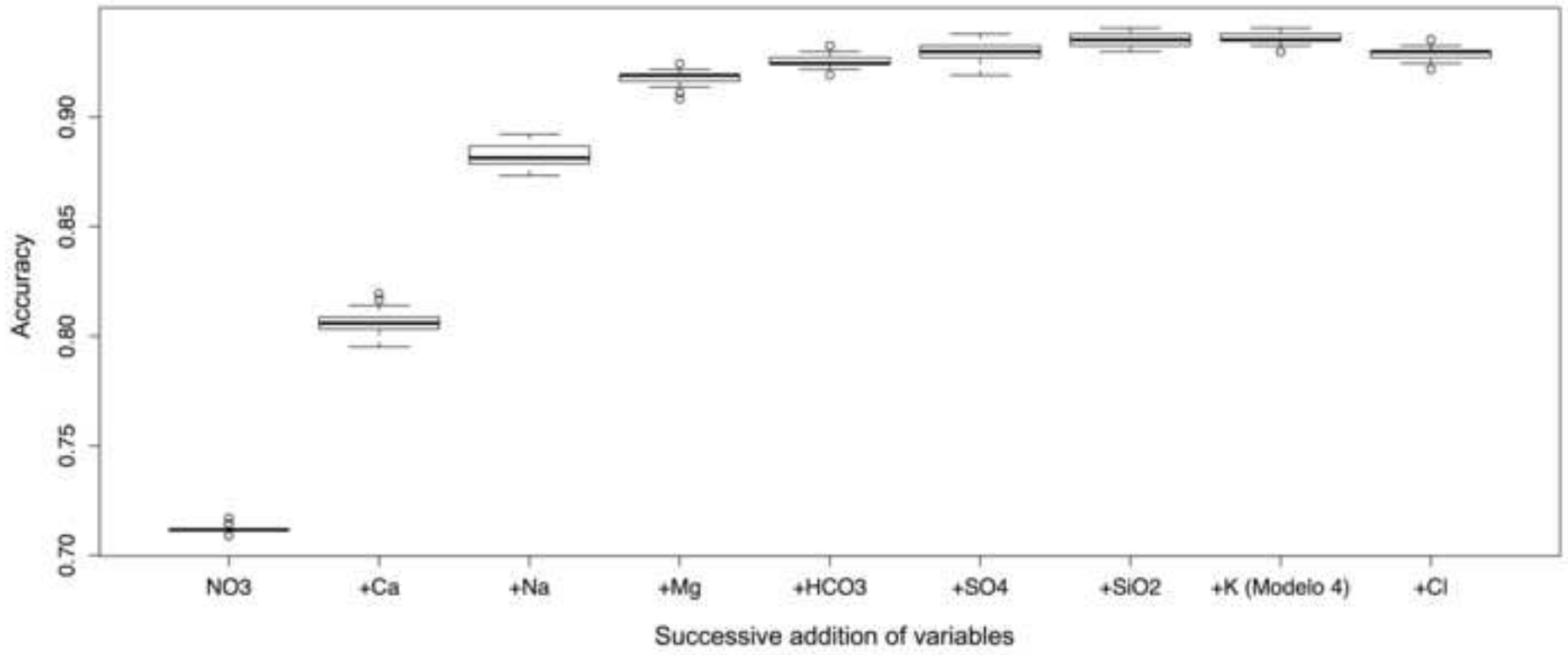

Figure 12

Click here to download high resolution image

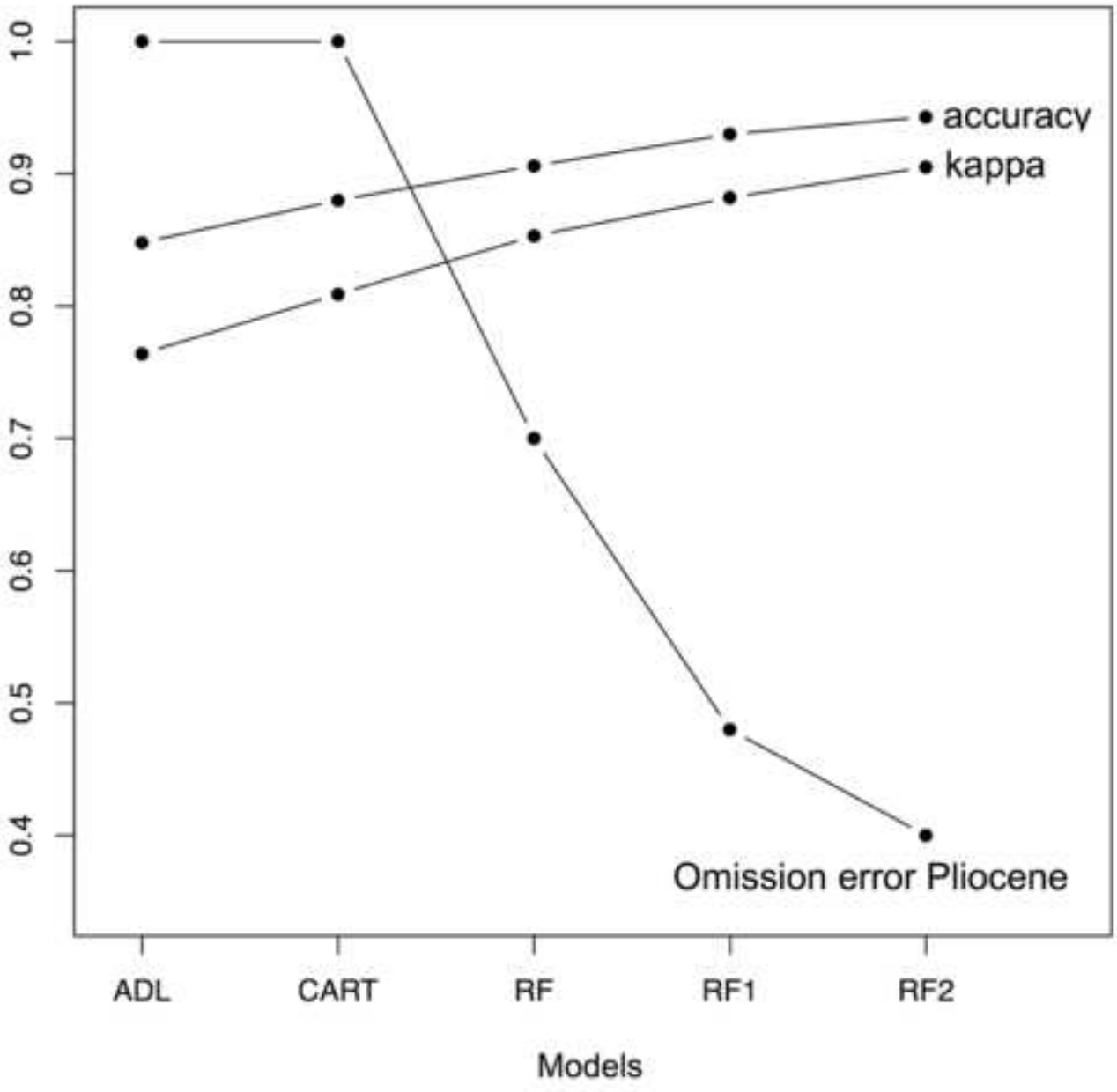

Figure 13

Click here to download high resolution image

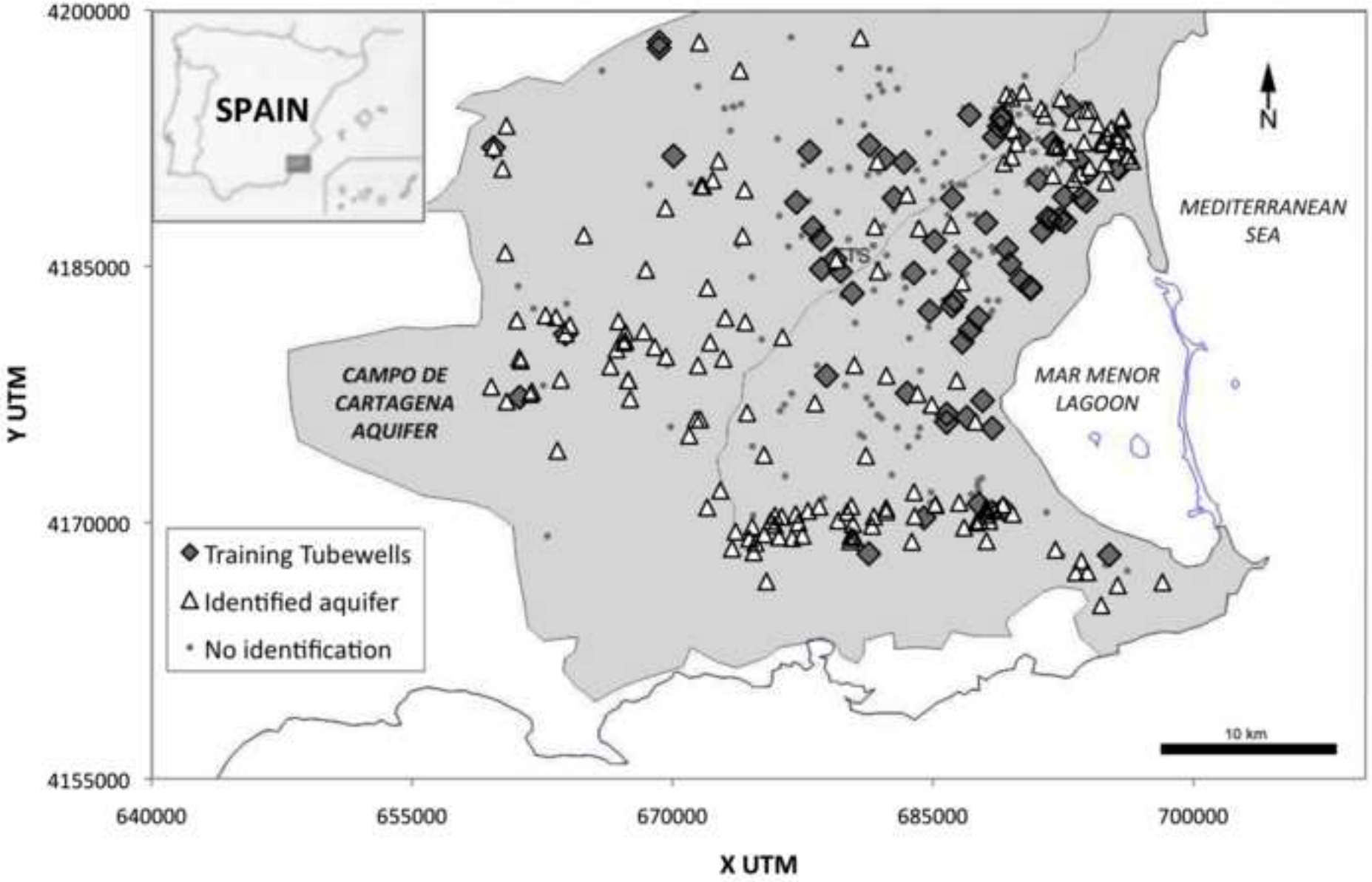

Figure 14

Click here to download high resolution image

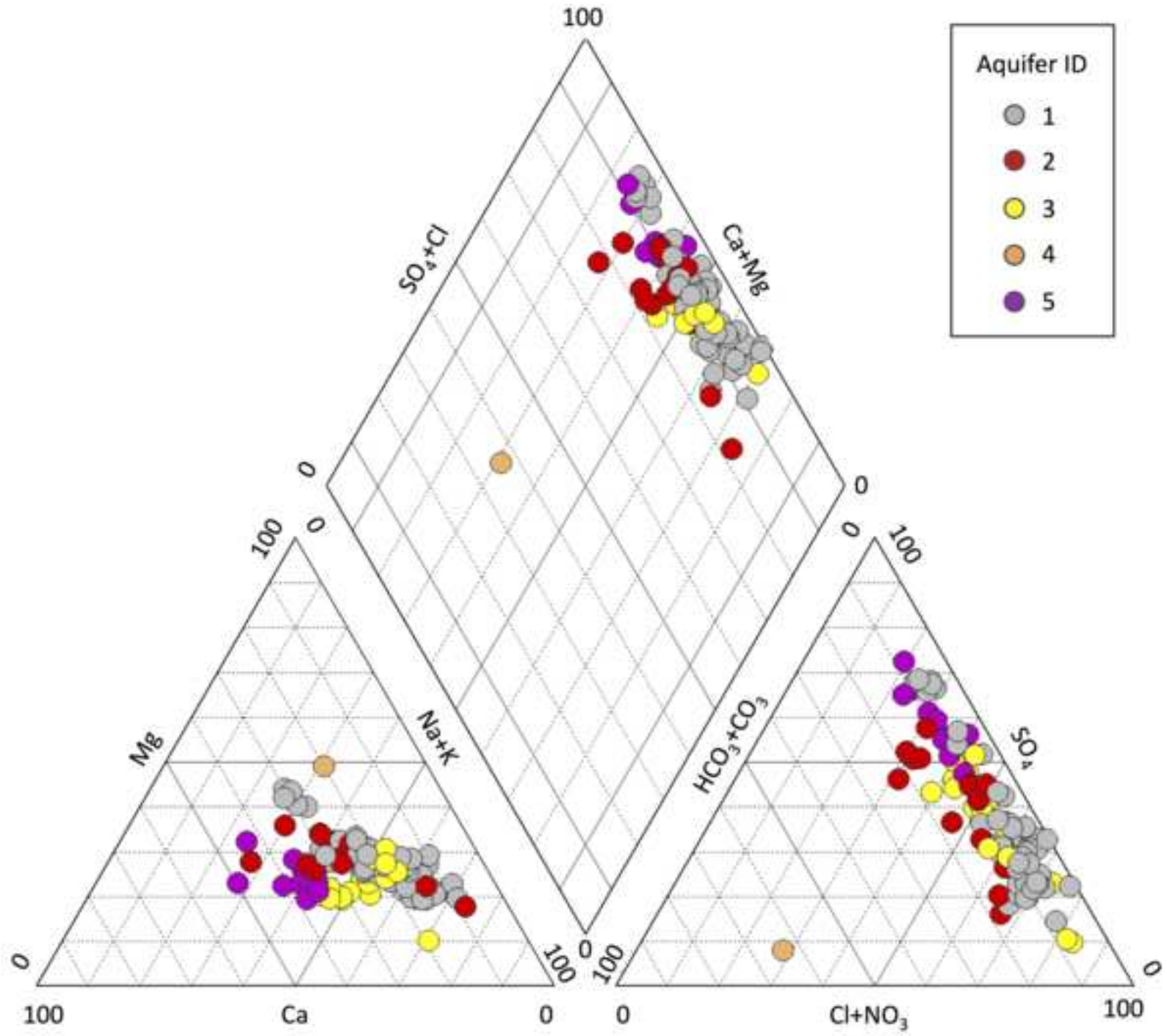

Table 1

Click here to download Table: Table1.doc

<table>
<tr><td>Whole dataset</td><td colspan="2">1592</td></tr>
<tr><td rowspan="2">Samples from one only aquifer</td><td colspan="2">Total 805</td></tr>
<tr><td>Featuring all variables 468</td><td>With missing variables 337</td></tr>
<tr><td rowspan="2">Samples of unknown origin</td><td colspan="2">Total 282</td></tr>
<tr><td>Identified with RF2 107</td><td>Not identified with RF2* 175</td></tr>
<tr><td>Samples identified as mixing</td><td colspan="2">403</td></tr>
</table>

*

Table 2

Click here to download Table: Table2.doc

	j=columns (true class)						**row totals**
i=rows (model class)		**1**	**2**	**3**	…	**J**	***ni*+**
	1	*n*11	*n*12	*n*13	…	*n*1*J*	*n*1+
	2	*n*21	*n*22	*n*23	…	*n*2*J*	*n*2+
	3	*n*31	*n*32	*n*33	…	*n*3*J*	*n*3+
	…	…	…	…	…	…	…
	I	*nI*1	*nI*2	*nI*3	…	*nIJ*	*nI*+
tot. column	***n*+*j***	*n*+1	*n*+2	*n*+3	…	*n*+*J*	*n*

Table 3

Click here to download Table: Table3.doc

	Q	P	M	To	Tr
Quaternary	206	7	5	0	0
Pliocene	0	0	1	0	1
Messinian	26	22	130	1	2
Tortonian	1	1	0	24	0
Trias	0	0	4	0	37
Commission error	5.5	100	28.2	7.7	9.8
Omission error	11.6	100	7.1	4	7.5
κ=0.764					
Overall accuracy=84.8%					

Table 4

Click here to download Table: Table4.doc

	Q	P	M	To	Tr
Quaternary	227	11	11	0	2
Pliocene	0	0	0	0	0
Messinian	5	18	127	2	2
Tortonian	1	1	1	23	0
Trias	0	0	1	0	36
Commission error	2.57	0	10	8	10
Omission error	9.56	100	17.64	11.54	2.7
κ=0.809 Overall accuracy=88%					

Table 5

Click here to download Table: Table5.doc

	Q	P	M	To	Tr
Quaternary	225	4	7	0	2
Pliocene	2	9	16	1	0
Messinian	6	16	129	0	1
Tortonian	0	1	0	25	0
Trias	0	0	1	0	36
Commission error	5.46	40	15.13	3.85	2.7
Omission error	3.43	70	7.86	0	10
κ=0.853 Overall accuracy = 90.6%					

Table 6

Click here to download Table: Table6.doc

ion	NO_3^-	Mg^{2+}	Na^+	Ca^{2+}	Cl^-	SO_4^{2-}	K^+	HCO_3^-	SiO_2
Overall accuracy	0.270	0.107	0.104	0.098	0.085	0.060	0.051	0.035	0.015

Table 7

Click here to download Table: Table7.doc

	Q	P	M	To	Tr
Quaternary	210	3	6	0	0
Pliocene	0	13	2	0	0
Messinian	5	9	72	0	0
Tortonian	0	0	0	25	0
Trias	1	0	0	0	25
Commission error	4.11	13.33	16.28	0	3.85
Omission error	2.78	48	10	0	0
κ =0.882 Overall accuracy = 93%					

Table 8

Click here to download Table: Table8.doc

	Q	P	M	To	Tr
Quaternary	211	3	4	0	0
Pliocene	0	15	2	0	0
Messinian	4	7	74	0	0
Tortonian	0	0	0	25	0
Trias	0	0	0	0	25
Commission error	3.21	11.76	12.94	0	3.85
Omission error	2.31	40	7.5	0	0
κ =0.905					
Overall accuracy = 94.3%					