

CF program

User manual

(for working with RandomForest projects)

Changes:

Date (program version)	Chapter	Description
03.02.09 (1.27)		First release version.
27.02.09 (1.28)		Predicted values for oob-set compounds can now be viewed on the "Forest statistics" tab.
	1	Working with case set files was improved.
		Specified model can be deleted from the model list (menu FOREST / DELETE FOREST)
	6	New chapter was inserted. "Options" menu with various settings was added to the program.
	4	Loading of multiple models to the same forest list are allowed now.
18.03.09 (1.29)	1	Y randomization procedure was implemented.
11.06.09 (2.00)		Possibility of analysis of multi-target models was added. Each Y (property) can have its own weight at model construction process. Menu "Statistics" has been removed. Menu "Rebuild forest" has been disabled. Visualization of model statistics and details has been changed and can be displayed for each Y (property) separately. Data-files can now contain missing values marked as NAN.
25.09.09 (2.03)		RF algorithm speed was significantly boosted Some interface elements were optimized for working with numerous data
05.11.09 (2.04)		Two domain applicability measures were implemented: 1) based on variable importance values (in descriptor space considering their relative importance) 2) based on each tree prediction (in space of models)
21.11.09 (2.05)		Multi-threads calculation was implemented, which can speed up very intensive calculation steps
10.01.10 (2.06)		Improve statistics calculation. Found memory leaks were eliminated Structure of the manual was considerably revised, new chapters were added and obsolete ones were deleted.

Content

1.	Creation of the first RandomForest project.	4
1.1.	Load data file.....	4
1.2.	Build RF model	7
1.2.1.	Variables tab	7
1.2.2.	Cases tab	11
1.2.3.	Forest tab	12
1.2.4.	Possible warning messages.....	14
2.	View model results	16
2.1.	General statistics.....	16
2.2.	View single trees composing RF model	17
2.3.	Detailed statistics and results	17
3.	Model (forest) routines.....	20
3.1.	Variable importance calculation	20
3.2.	Domain of applicability calculation.....	21
4.	“Preset mode” of model construction.....	23
5.	Model (files) routines.....	24
5.1.	Saving model.....	24
5.2.	Opening model(s).....	24
6.	Prediction of compounds properties which are in an external data-file.	25
7.	General information.	26
8.	Afterword.....	27



Important remarks are marked in such style.



Advices are marked in such style.

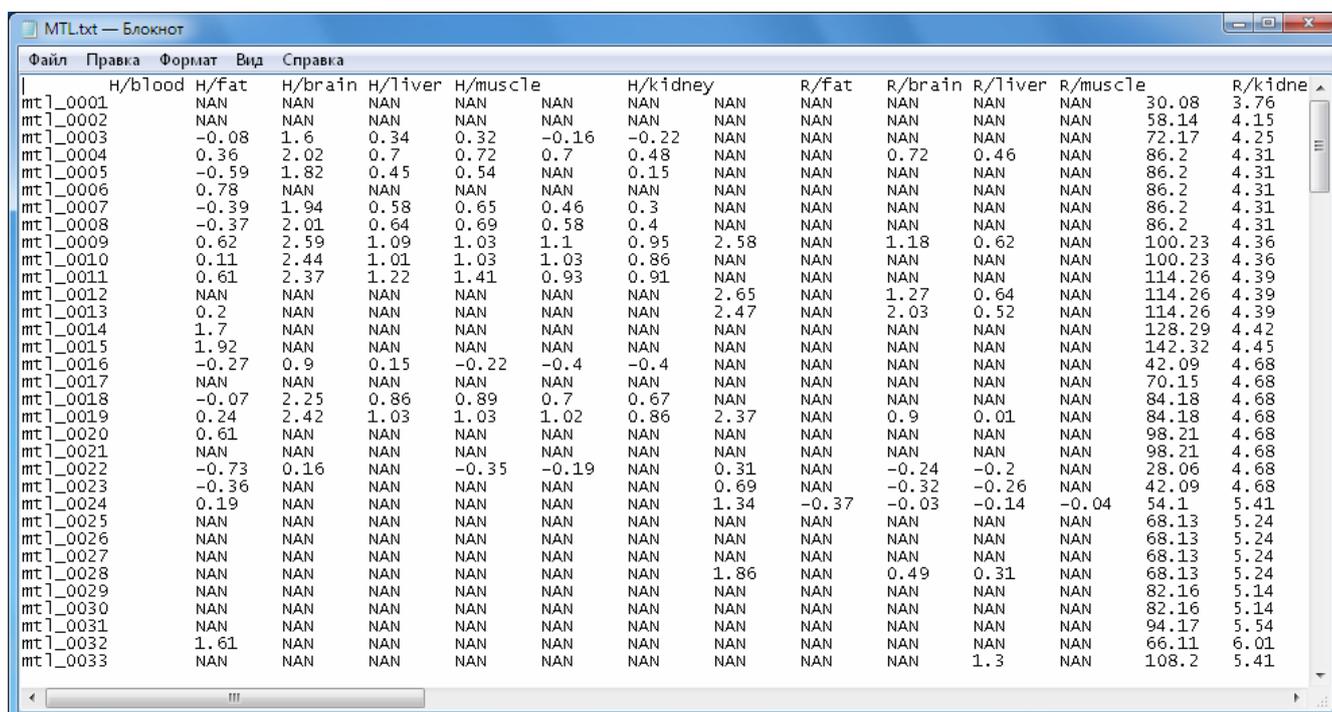
1. Creation of the first RandomForest project.

1.1. Load data file

To create RandomForest project choose menu FILE / NEW PROJECT / NEW RANDOM FOREST PROJECT

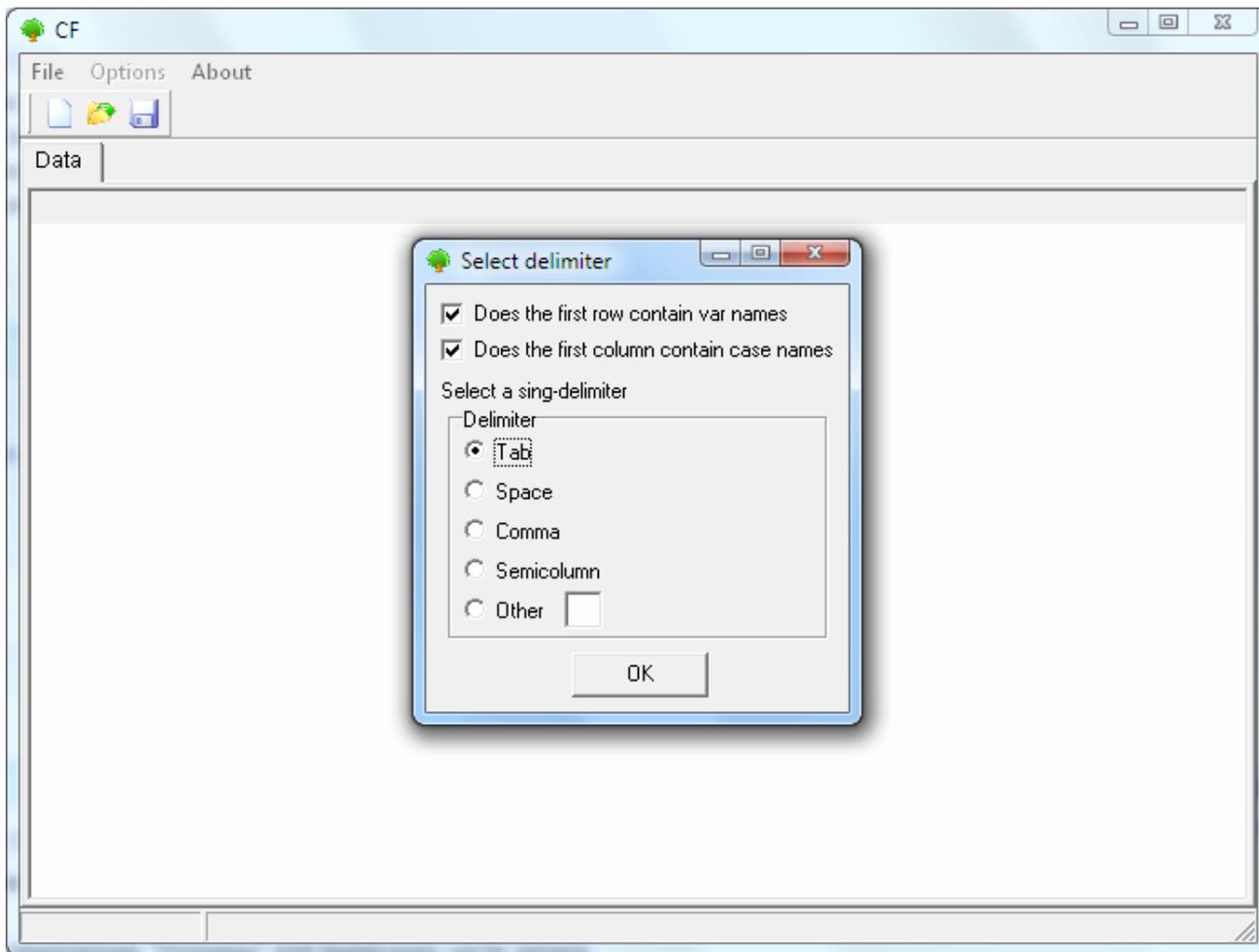
Select a file with source data in the dialog:

- rfd-file, this is own file format of CF program,
- dat-file, this is file format of MDA1 program from HiT-QSAR Software package,
- txt-file, plain text format, descriptors are in columns, cases (compounds) are in rows (see example below). First row and column contain descriptors names and molecules names correspondingly. If some values are missing then they should be represented as NAN textual value or leave empty. Such missed descriptor values automatically replaced with special NAN value. Descriptor values should be numerical only (restriction of the current version) else an error message will be displayed and file will not be opened. (Program does not check all possible errors in txt-file, so be careful and be sure that there are no errors in your data file).



	H/blood	H/fat	H/brain	H/liver	H/muscle	H/kidney	R/fat	R/brain	R/liver	R/muscle	R/kidney		
mt1_0001	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	30.08	3.76	
mt1_0002	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	58.14	4.15	
mt1_0003	-0.08	1.6	0.34	0.32	-0.16	-0.22	NAN	NAN	NAN	NAN	72.17	4.25	
mt1_0004	0.36	2.02	0.7	0.72	0.7	0.48	NAN	NAN	0.72	0.46	NAN	86.2	4.31
mt1_0005	-0.59	1.82	0.45	0.54	NAN	0.15	NAN	NAN	NAN	NAN	NAN	86.2	4.31
mt1_0006	0.78	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	86.2	4.31
mt1_0007	-0.39	1.94	0.58	0.65	0.46	0.3	NAN	NAN	NAN	NAN	NAN	86.2	4.31
mt1_0008	-0.37	2.01	0.64	0.69	0.58	0.4	NAN	NAN	NAN	NAN	NAN	86.2	4.31
mt1_0009	0.62	2.59	1.09	1.03	1.1	0.95	2.58	NAN	1.18	0.62	NAN	100.23	4.36
mt1_0010	0.11	2.44	1.01	1.03	1.03	0.86	NAN	NAN	NAN	NAN	NAN	100.23	4.36
mt1_0011	0.61	2.37	1.22	1.41	0.93	0.91	NAN	NAN	NAN	NAN	NAN	114.26	4.39
mt1_0012	NAN	NAN	NAN	NAN	NAN	NAN	2.65	NAN	1.27	0.64	NAN	114.26	4.39
mt1_0013	0.2	NAN	NAN	NAN	NAN	NAN	2.47	NAN	2.03	0.52	NAN	114.26	4.39
mt1_0014	1.7	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	128.29	4.42
mt1_0015	1.92	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	142.32	4.45
mt1_0016	-0.27	0.9	0.15	-0.22	-0.4	-0.4	NAN	NAN	NAN	NAN	NAN	42.09	4.68
mt1_0017	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	70.15	4.68
mt1_0018	-0.07	2.25	0.86	0.89	0.7	0.67	NAN	NAN	NAN	NAN	NAN	84.18	4.68
mt1_0019	0.24	2.42	1.03	1.03	1.02	0.86	2.37	NAN	0.9	0.01	NAN	84.18	4.68
mt1_0020	0.61	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	98.21	4.68
mt1_0021	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	98.21	4.68
mt1_0022	-0.73	0.16	NAN	-0.35	-0.19	NAN	0.31	NAN	-0.24	-0.2	NAN	28.06	4.68
mt1_0023	-0.36	NAN	NAN	NAN	NAN	NAN	0.69	NAN	-0.32	-0.26	NAN	42.09	4.68
mt1_0024	0.19	NAN	NAN	NAN	NAN	NAN	1.34	-0.37	-0.03	-0.14	-0.04	54.1	5.41
mt1_0025	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	68.13	5.24
mt1_0026	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	68.13	5.24
mt1_0027	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	68.13	5.24
mt1_0028	NAN	NAN	NAN	NAN	NAN	NAN	1.86	NAN	0.49	0.31	NAN	68.13	5.24
mt1_0029	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	82.16	5.14
mt1_0030	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	82.16	5.14
mt1_0031	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	94.17	5.54
mt1_0032	1.61	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	66.11	6.01
mt1_0033	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	1.3	NAN	108.2	5.41

If txt-file has been chosen to create new project following dialog window would be displayed. One should select appropriate settings to load txt-file. If variables (descriptors) names are absent in the first line of the file (uncheck corresponding box) program will give names automatically (Var1, Var2 etc). Analogous procedure will be executed if case names are absent.



After successful loading of source data it will be displayed on “Data” tab.

! There is no possibility to edit data.

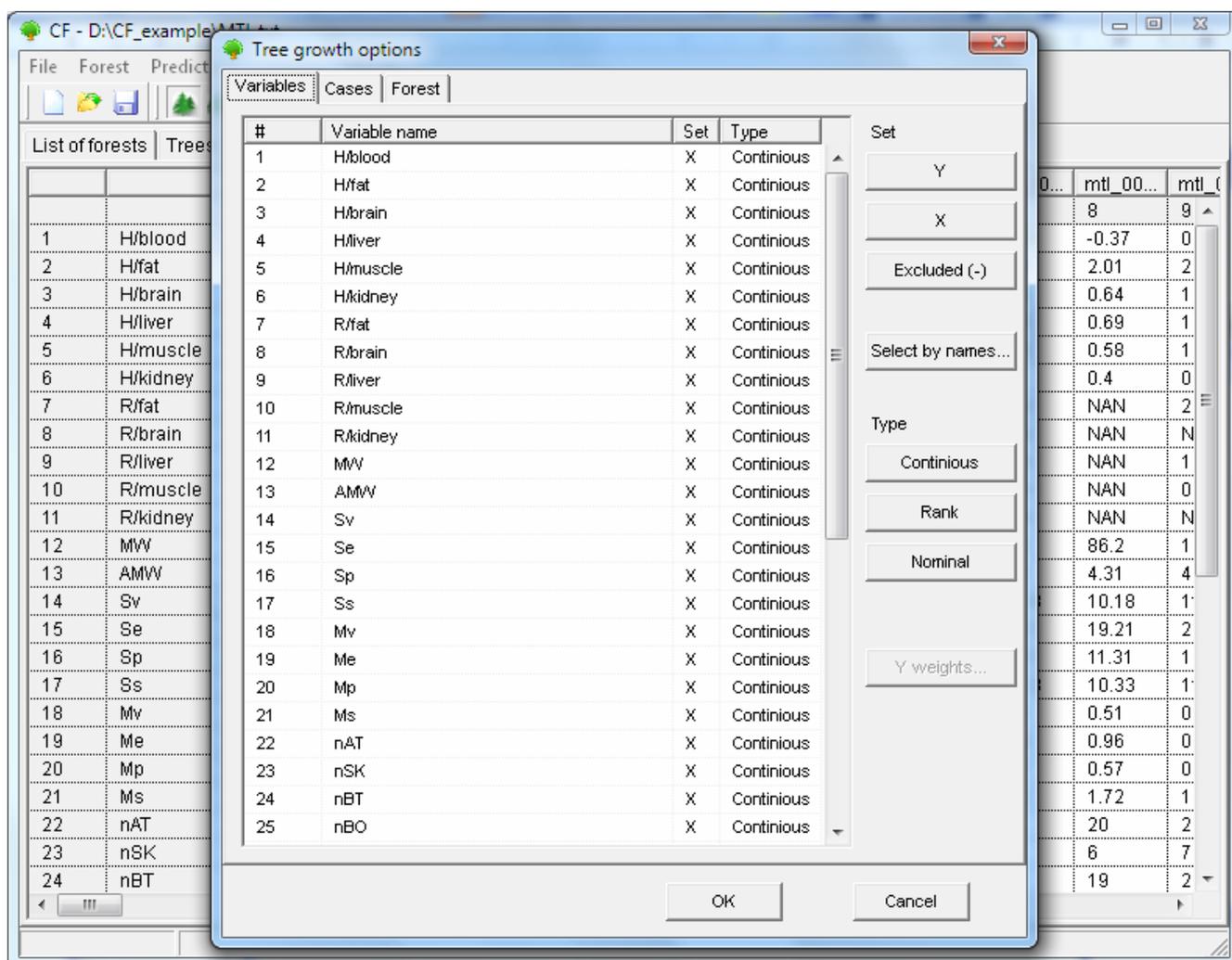
The screenshot shows a software window with a menu bar (File, Forest, Prediction, Options, About) and a toolbar. The 'Data' tab is active, displaying a table with 24 rows and 11 columns. The columns are labeled 'mtl_00...' and the rows contain numerical values and categorical labels. The table data is as follows:

		mtl_00...								
		1	2	3	4	5	6	7	8	9
1	H/blood	NAN	NAN	-0.08	0.36	-0.59	0.78	-0.39	-0.37	0
2	H/fat	NAN	NAN	1.6	2.02	1.82	NAN	1.94	2.01	2
3	H/brain	NAN	NAN	0.34	0.7	0.45	NAN	0.58	0.64	1
4	H/liver	NAN	NAN	0.32	0.72	0.54	NAN	0.65	0.69	1
5	H/muscle	NAN	NAN	-0.16	0.7	NAN	NAN	0.46	0.58	1
6	H/kidney	NAN	NAN	-0.22	0.48	0.15	NAN	0.3	0.4	0
7	R/fat	NAN	2							
8	R/brain	NAN	N							
9	R/liver	NAN	NAN	NAN	0.72	NAN	NAN	NAN	NAN	1
10	R/muscle	NAN	NAN	NAN	0.46	NAN	NAN	NAN	NAN	0
11	R/kidney	NAN	N							
12	MW	30.08	58.14	72.17	86.2	86.2	86.2	86.2	86.2	1
13	AMW	3.76	4.15	4.25	4.31	4.31	4.31	4.31	4.31	4
14	Sv	3.79	6.99	8.59	10.18	10.18	10.18	10.18	10.18	1
15	Se	7.66	13.44	16.33	19.21	19.21	19.21	19.21	19.21	2
16	Sp	4.27	7.79	9.55	11.31	11.31	11.31	11.31	11.31	1
17	Ss	4	7	8.5	10	10.75	10.67	10.33	10.33	1
18	Mv	0.47	0.5	0.51	0.51	0.51	0.51	0.51	0.51	0
19	Me	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0
20	Mp	0.53	0.56	0.56	0.57	0.57	0.57	0.57	0.57	0
21	Ms	2	1.75	1.7	1.67	1.79	1.78	1.72	1.72	1
22	nAT	8	14	17	20	20	20	20	20	2
23	nSK	2	4	5	6	6	6	6	6	7
24	nBT	7	13	16	19	19	19	19	19	2

1.2. Build RF model

1.2.1. Variables tab

To grow forest (build model) choose menu FOREST / GROW FOREST.

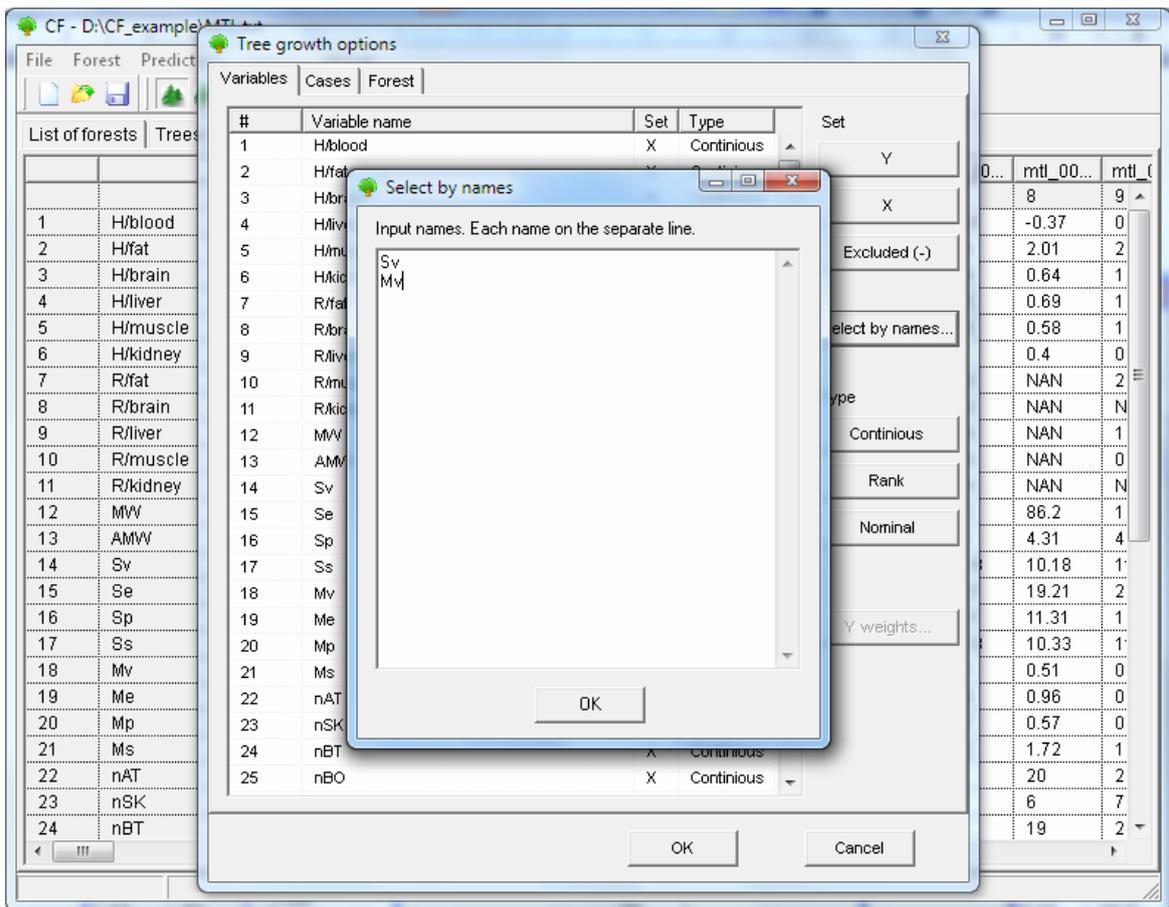


The following window will appear. Select variables which will be used for model construction on “Variables” tab. Variables can be dependent (Y, several Y’s are allowed), independent (X) and excluded (which will not take part in model construction). Also variables type should be chosen. Y variable can possess all three types (but each Y should have identical variable type), X variables can be continuous type only (restriction of the current program version). To do these operation simply select variable(s) in the list and click on the appropriate button.

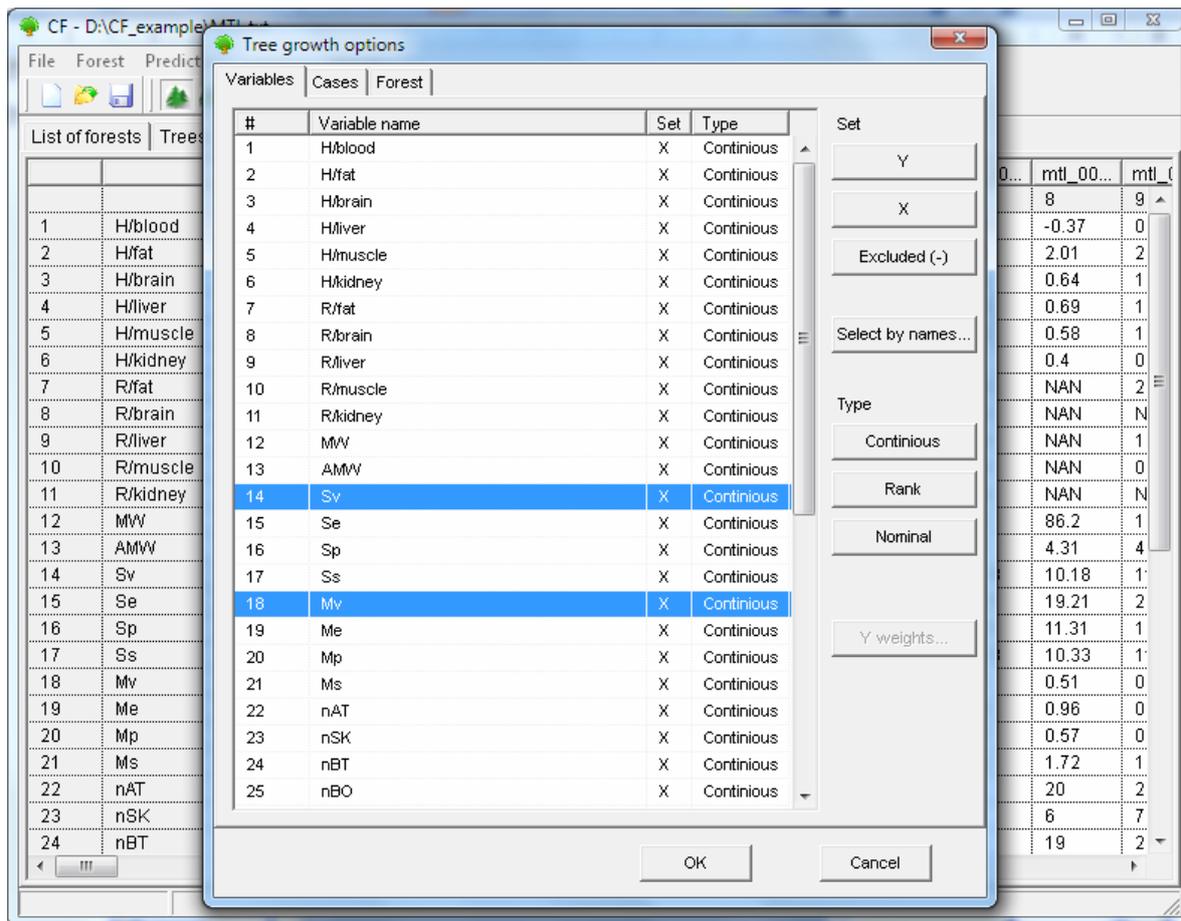


Buttons Y, X and Excluded have keyboard shortcuts - y, x and space correspondingly.

One can select variables in the list by its names. Click on the “Select by names” button and input variables names (one variable name per line).



After button OK clicked specified variables would be selected (and you can set all of them as excluded for example).



If you choose several Y's then "Y weights..." button will be enabled and weights for each Y (property) will be able to be assigned. All positive numbers are allowed.

The screenshot displays a software interface with a 'Tree growth options' dialog box. The dialog has three tabs: 'Variables', 'Cases', and 'Forest'. The 'Variables' tab is active, showing a list of variables with their corresponding 'Set' and 'Type'.

#	Variable name	Set	Type
1	H/blood	Y	Continuous
2	H/fat	Y	Continuous
3	H/brain	Y	Continuous
4	H/liver	Y	Continuous
5	H/muscle	Y	Continuous
6	H/kidney	Y	Continuous
7	R/fat	Y	Continuous
8	R/brain	Y	Continuous
9	R/liver	Y	Continuous
10	R/muscle	Y	Continuous
11	R/kidney	Y	Continuous
12	MW	Y	Continuous
13	AMW	Y	Continuous
14	Sv	Y	Continuous
15	Se	Y	Continuous
16	Sp	Y	Continuous
17	Ss	Y	Continuous
18	Mv	Y	Continuous
19	Me	Y	Continuous
20	Mp	X	Continuous
21	Ms	X	Continuous
22	nAT	X	Continuous
23	nSK	X	Continuous
24	nBT	X	Continuous
25	nBO	X	Continuous

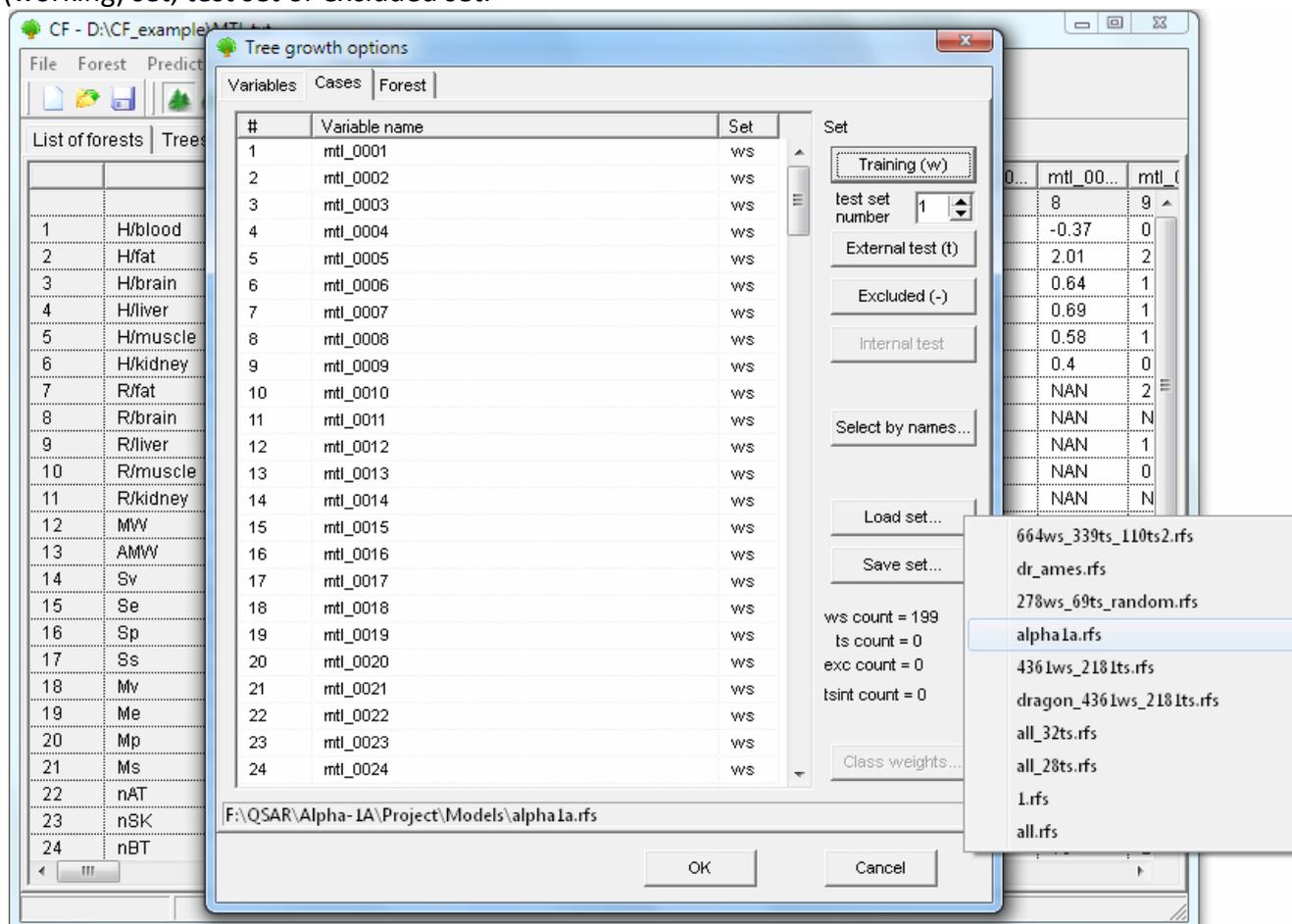
The 'Properties weights' sub-dialog box is open, showing a table with 'Property' and 'Weight' columns.

Property	Weight
H/blood	1
H/fat	1
H/brain	1
H/liver	1
H/muscle	1
H/kidney	1
R/fat	1
R/brain	1

The background shows a 'List of forests' table with columns for forest ID and variable names, and a data table with columns 'mtl_00...' and 'mtl_...'.

1.2.2. Cases tab

Select appropriate set of each case (compound) on the “Cases” tab. Possible values are training (working) set, test set or excluded set.



The program allows to define up to 10 separate test sets. To set a case to the wanted test set (second for example) one should specify corresponding number in “test set number” field (in our case it is 2) and then select the case and click “External test” button.



Buttons Training, External test и Excluded have keyboard shortcuts – w, t and space correspondingly.

It is possible to load and save case sets. Case sets saves simultaneously in two formats:

- rfs, internal format of CF program (it supports multiple test sets);
- wsf, format of MDA1 program from HiT-QSAR Software package for backward compatibility purpose (it supports only one test set, all test sets (if more than one) are saved as one entire test set).

Program keeps 10 latest loaded and saved set-files. To view list of them click by right mouse button on “Load set...” button. Latest used files will be on the top of the list.

Full path to selected set file in popup menu are displayed in the status bar just under the list of cases. If opened set file was not find in its location the respective message would be appeared in the status bar.

Statistics of compound numbers in each set are displayed below:

- ws – number of compounds in the training set;
- ts – number of compounds in all test sets;
- exc – number of compounds in the excluded set.

If one Y variable selected and it has ranked or nominal type (“Variables” tab) then button “Class weights...” will be enabled. Click it and following window will appear where one can define weights of each compound class. Case weights can be integers only. This window is analogous to previously described “Y weights...” dialog from “Variable” tab.

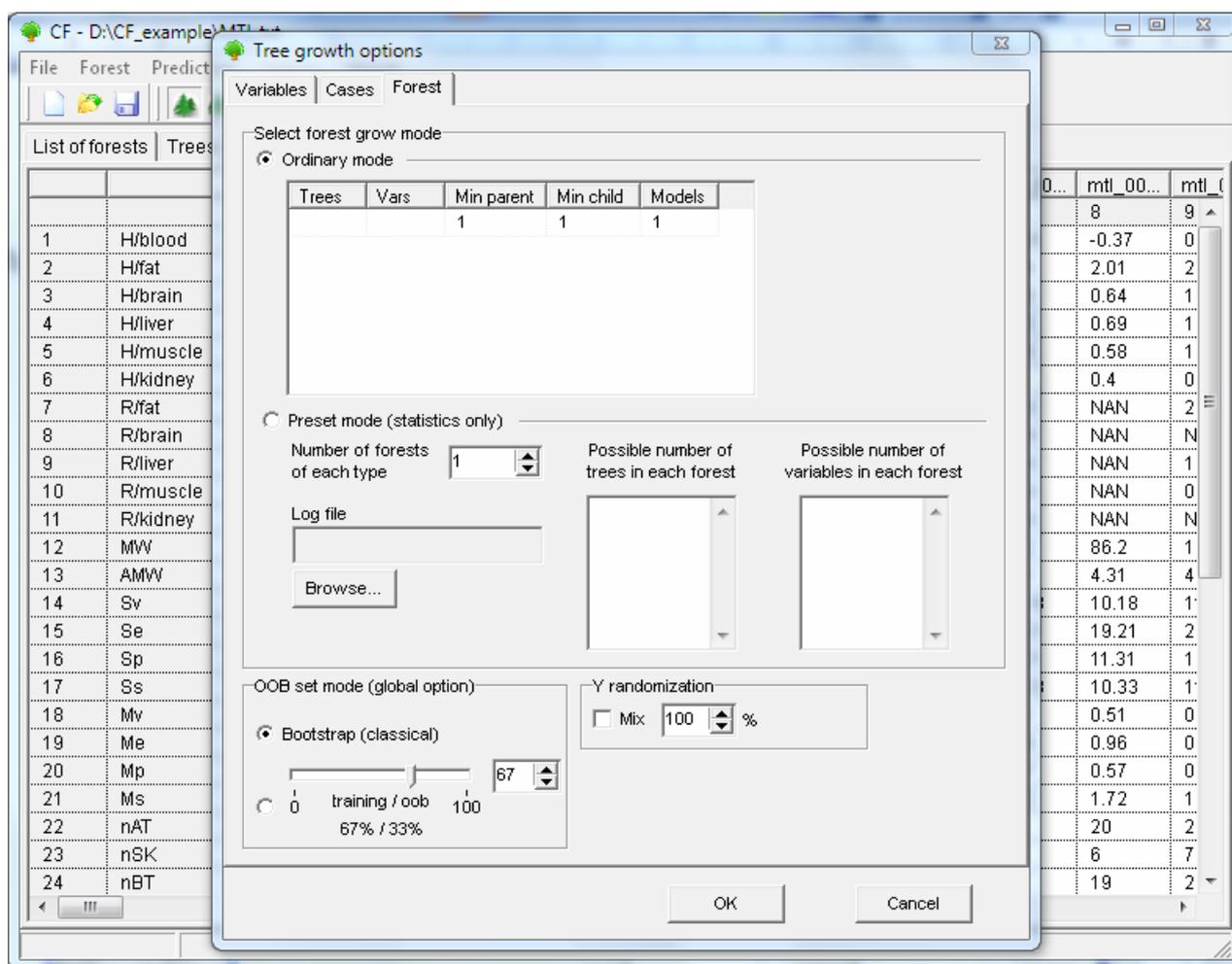


It is recommended to leave all values equal to 1 because testing of this option is in progress now.

Function of “Select by names” button is absolutely analogous to the same button on “Variables” tab.

1.2.3. Forest tab

Model building settings are defined on “Forest” tab.



Here “Ordinary mode” » is described only. “Preset mode” will be described below in a separate chapter.

It should be input in the table:

- **Trees** – the number of trees in the Random Forest model;
- **Vars** – the number of variables (descriptors) which will be used for splitting in each node of trees. If one input this value which will be greater than available descriptors number this value will be reduced automatically at the calculation step.
- **Min parent** and **Min child** – it is a minimum number of cases (compounds) in the parent or child nodes. It can not be greater than 1/3 from the number of training set compounds. Otherwise warning message will appear and this model will not be constructed.



In the original algorithm there are no such restriction parameters. All trees are growing for their maximum size. So we recommend to use 1 as a value of “Min child” and “Min parent” fields for classification tasks. For regression task to greater numbers can be assigned for these values to increase calculation speed (for example Min parent = 5), usually it has no influence on model quality

- **Models** – it is the number of models which will be constructed according to specified settings.

When all fields in one row are filled with non-zero values another row is appeared. This new row one can fill with new settings. Thus a queue (package) of tasks is formed. Press Ctrl+Del to delete selected row in the table.



In the case of very big datasets (thousands of cases and variables) models construction consumes considerable memory size. So be careful when you choose forest growth settings. And be sure that you have enough memory to complete all your needed operations.

A method of training set formation of each tree is specified in the **OOB set mode options** dialog:

- **Bootstrap** – it as a classical mode of formation of training and out-of-bag sets for each tree construction (with replacement).
- **Custom** – user can specify parts of cases of training and out-of-bag sets (without replacement).

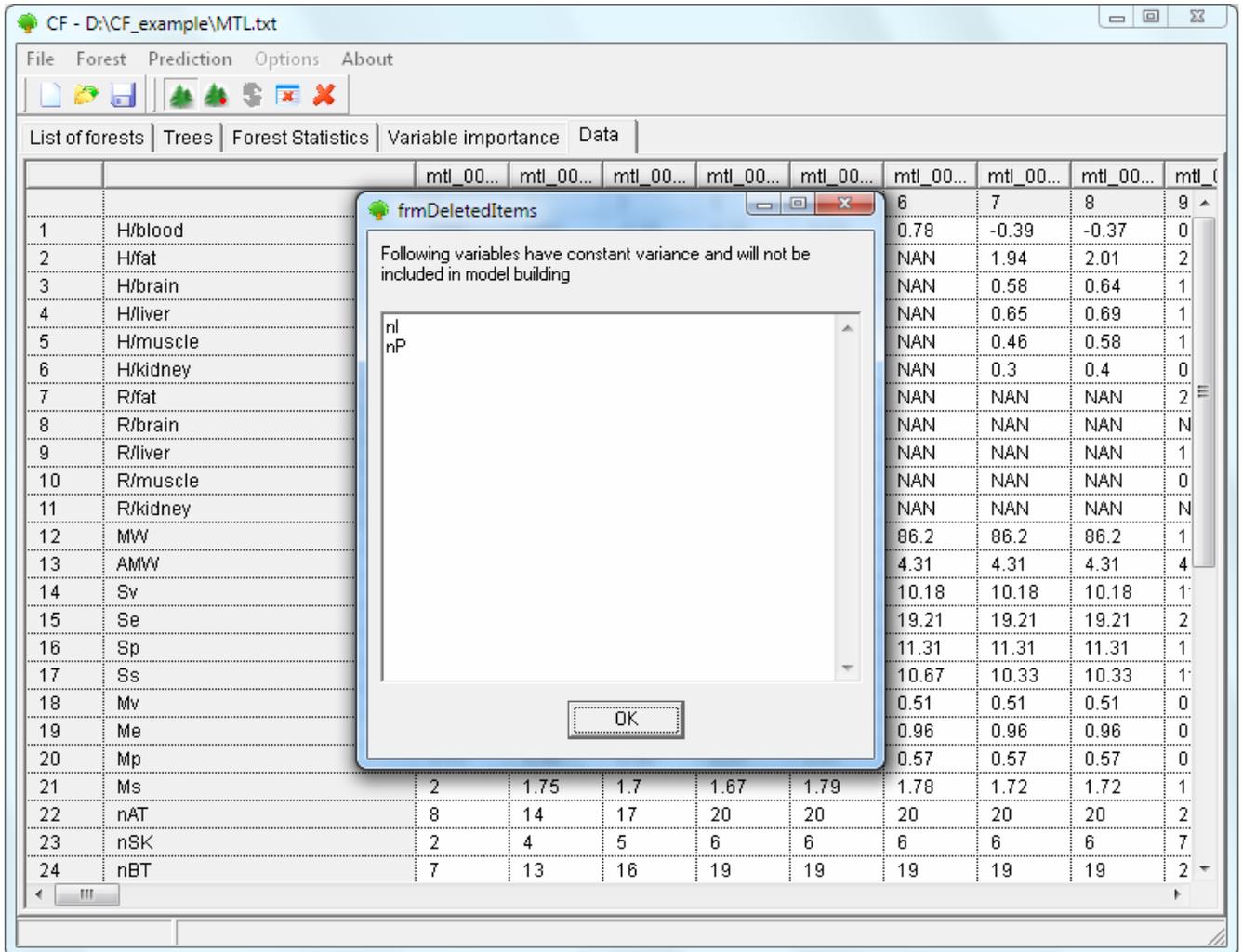


Experience is shown that models which constructed in the second (custom) mode have not appreciable changes in their quality. In addition there is only little difference in model construction time. So we recommend to choose the first (classical) mode (bootstrap).

Each model can be constructed with randomized Y values (**Y randomization**). To define part of Y values which will be shuffled at model building one should check “**Mix**” field and choose corresponding value from the range 0-100. If 100% value was chosen it would be Y scrambling procedure. This procedure is used to prove that obtained model isn’t random.

1.2.4. Possible warning messages

After OK button is pressed, if there are descriptors with constant and/or missing values among X's then a list with those descriptors names will be appeared in separate windows. All these descriptors will be removed from the model construction process.



The screenshot shows a software window titled "CF - D:\CF_example\MTL.txt". The window has a menu bar with "File", "Forest", "Prediction", "Options", and "About". Below the menu bar is a toolbar with icons for file operations and data management. The main area is divided into tabs: "List of forests", "Trees", "Forest Statistics", "Variable importance", and "Data". The "Data" tab is active, displaying a table with columns for various variables and rows for different cases. A dialog box titled "frmDeletedItems" is overlaid on the table, displaying a list of cases that have missing values for all selected properties and will be excluded from the training set. The dialog box has an "OK" button.

		mtl_00...								
1	H/blood							7	8	9
2	H/fat							-0.39	-0.37	0
3	H/brain							1.94	2.01	2
4	H/liver							0.58	0.64	1
5	H/muscle							0.65	0.69	1
6	H/kidney							0.46	0.58	1
7	R/fat							0.3	0.4	0
8	R/brain							NAN	NAN	2
9	R/liver							NAN	NAN	N
10	R/muscle							NAN	NAN	1
11	R/kidney							NAN	NAN	0
12	MW							NAN	NAN	N
13	AMW							86.2	86.2	1
14	Sv							4.31	4.31	4
15	Se							10.18	10.18	1
16	Sp							19.21	19.21	2
17	Ss							11.31	11.31	1
18	Mv							10.33	10.33	1
19	Me							0.51	0.51	0
20	Mp							0.96	0.96	0
21	Ms							0.57	0.57	0
22	nAT	2	1.75	1.7	1.67	1.79	1.78	1.72	1.72	1
23	nSK	8	14	17	20	20	20	20	20	2
24	nBT	2	4	5	6	6	6	6	6	7
		7	13	16	19	19	19	19	19	2

Deleted items list:

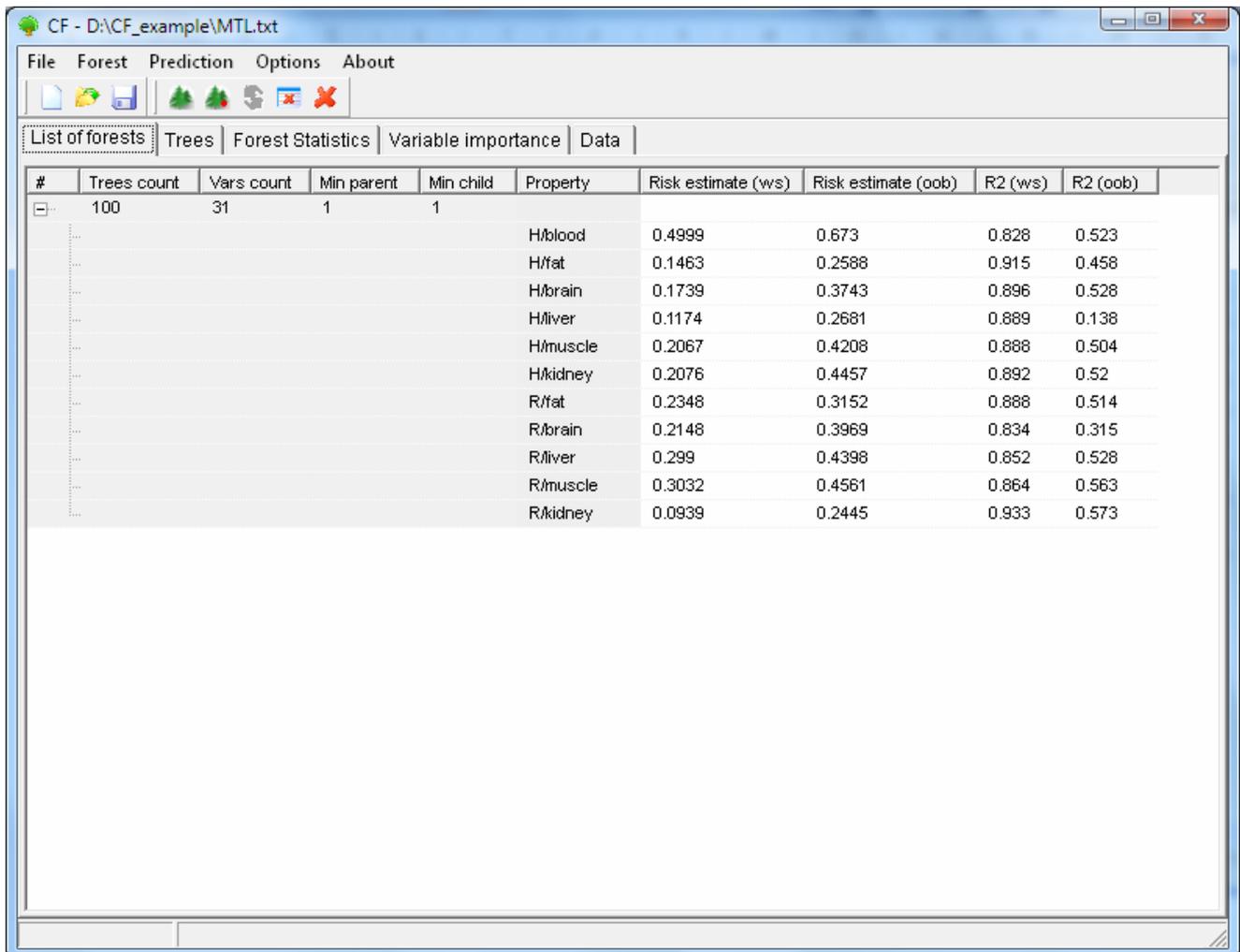
- mtl_0001
- mtl_0002
- mtl_0017
- mtl_0021
- mtl_0025
- mtl_0026
- mtl_0027
- mtl_0029
- mtl_0030
- mtl_0031
- mtl_0039
- mtl_0040
- mtl_0091
- mtl_0094
- mtl_0097
- mtl_0098
- mtl_0100
- mtl_0103

Progress of model construction is displayed in the bottom of main window. After that statistics of obtained model is calculated for each case set.

2. View model results

2.1. General statistics

General obtained results can be looked on Forest list tab. Statistics for each property are displayed.



#	Trees count	Vars count	Min parent	Min child	Property	Risk estimate (ws)	Risk estimate (oob)	R2 (ws)	R2 (oob)
100	31	1	1						
					H/blood	0.4999	0.673	0.828	0.523
					H/fat	0.1463	0.2588	0.915	0.458
					H/brain	0.1739	0.3743	0.896	0.528
					H/liver	0.1174	0.2681	0.889	0.138
					H/muscle	0.2067	0.4208	0.888	0.504
					H/kidney	0.2076	0.4457	0.892	0.52
					R/fat	0.2348	0.3152	0.888	0.514
					R/brain	0.2148	0.3969	0.834	0.315
					R/liver	0.299	0.4398	0.852	0.528
					R/muscle	0.3032	0.4561	0.864	0.563
					R/kidney	0.0939	0.2445	0.933	0.573

All data from this table can be copied by right mouse button click. A case set is shown into the brackets after the value name in the column caption (**ws**–training set, **oob**–out-of-bag set, **ts1**–first test set, **ts2**–second test set and so on).

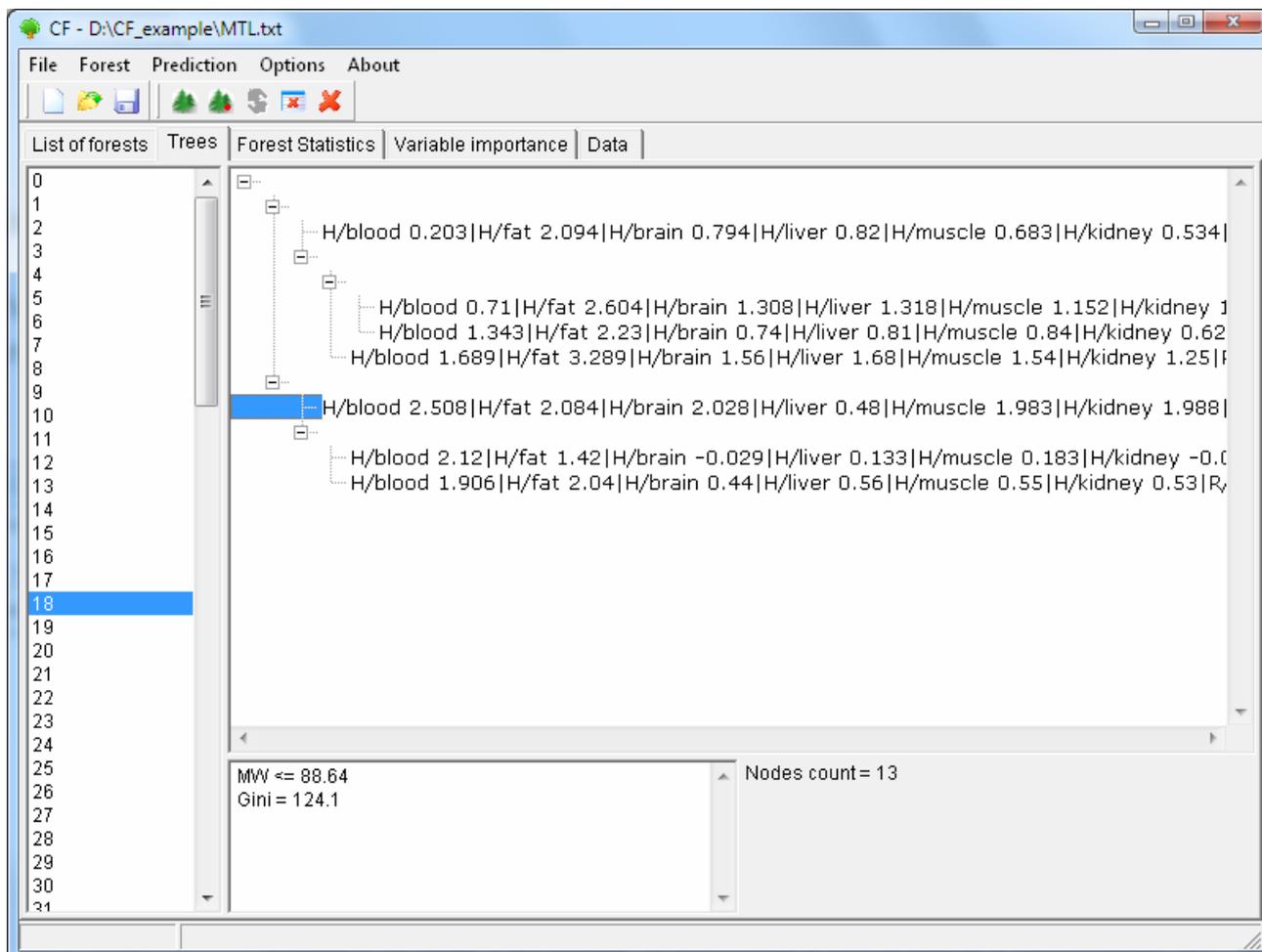
Risk estimate value is a **misclassification error** for classification models and **mean square error** for regression ones. Values of **coefficients of determination** (R²) are calculated only for regression models. R² for out-of-bag (OOB) and test sets are calculated by the formula $1 - \text{PRESS} / \text{SS}$.



New obtaining models are added to the end of the models list until the list will not be cleared. To clear the models list choose menu FOREST / CLEAR FOREST LIST. To delete selected model from the list choose menu FOREST / DELETE FOREST

2.2. View single trees composing RF model

To do that select model in the list by left-click and switch to Trees tab. Each tree in the list can be selected and viewed. Due to of a little importance of such information only general information is displayed.



2.3. Detailed statistics and results

To do that make double click on the model in the list or select model in the list by left-click and switch to Forest Statistics tab.

The following information are displayed:

- 1) compound name
- 2) set to which compound belongs
- 3) observed values of investigated properties
- 4) predicted values of investigated properties
 - for **regression models** it is a mean of all single tree predictions;
 - for **classification models** it is a class having majority of votes (one tree—one vote).
- 5) is compound inside (sing "+") or outside (sing "-") of domain of applicability (several domain of applicability measures were implemented and will discussed separately)

Additional **regression model** specific information:

- 1) standard deviation (StdDev) – it is calculated from set of predicted values by each tree

Additional **classification model** specific information:

- 1) number of each class predictions (in separate columns)
- 2) misclassification matrix (on the bottom of the window)

#	Compound name	Set	Observed (H/blo...	Predicted (H/blo...	Pred. StdDev (H/blood)	Observed (H/f...	Predicted (H/f...	Pred. StdDev (H/f...
	mtl_0011	oob	0.61	0.875	0.58	2.37	2.386	0.505
	mtl_0101	oob	-1.099	0.42	0.267	-	1.753	0.475
	mtl_0102	oob	-0.249	0.284	0.24	-	1.71	0.417
	mtl_0104	oob	0.87	0.652	0.663	-	2.164	0.622
	mtl_0106	oob	-0.359	0.432	0.488	-	1.913	0.478
	mtl_0107	oob	-0.589	0.564	0.586	-	1.976	0.647
	mtl_0108	oob	-1.519	0.544	0.45	-	1.994	0.503
	mtl_0012	oob	-	0.898	0.559	-	2.455	0.473
	mtl_0117	oob	-	0.329	0.339	-	1.434	0.559
	mtl_0118	oob	-	0.231	0.258	-	1.471	0.575
	mtl_0119	oob	-	0.678	0.42	-	1.941	0.578
	mtl_0120	oob	-0.079	0.421	0.343	-	1.834	0.484

There is a possibility to filter results by property and/or set. Selecting certain property from the list allows to see detailed model property corresponding specified property (see figure below).

#	Compound name	Set	Observed (H/blo...	Predicted (H/blo...	Pred. StdDev (H/blo...
mtl_0011		oob	0.61	0.875	0.58
mtl_0101		oob	-1.099	0.42	0.267
mtl_0102		oob	-0.249	0.284	0.24
mtl_0104		oob	0.87	0.652	0.663
mtl_0106		oob	-0.359	0.432	0.488
mtl_0107		oob	-0.589	0.564	0.586
mtl_0108		oob	-1.519	0.544	0.45
mtl_0012		oob	-	0.898	0.559
mtl_0117		oob	-	0.329	0.339
mtl_0118		oob	-	0.231	0.258
mtl_0119		oob	-	0.678	0.42
mtl_0120		oob	-0.079	0.421	0.343
mtl_0013		oob	0.2	0.833	0.566

	WS	OOB
R2	0.838	0.757
R2 test	0.645	0.534
MSE	0.501	0.657
RMSE	0.708	0.811

For regression models following measures are calculated:

- 1) R^2 – determination coefficient (reliable for training set only)
- 2) R^2_{test} – coefficient is calculated as $1 - PRESS/SS$ (reliable for OOB, test and external sets)
- 3) MSE – mean standard error
- 4) RMSE –root mean square error

For classification models following measures are calculated:

- 1) Misclassification error – ratio of number of erroneous predictions to the whole number of predictions

When domain of applicability was calculated corresponding values based on set of compounds inside of domain of applicability are displayed.

3. Model (forest) routines.

Unlimited number of trees can be added to the selected forest. To make this choose menu FOREST / ADD TREES TO FOREST and specify the desired number of trees.

3.1. Variable importance calculation

To calculate variable importances choose menu FOREST / CALC VAR IMPORTANCE.

#	Trees count	Vars count	Min parent	Min child	Property	Risk estimate (ws)	Risk estimate (oob)	R2 (ws)	R2 (oob)
100	31	1	1						
					H/blood	0.5011	0.657	0.838	0.534
					H/fat	0.1244	0.2528	0.921	0.471
					H/brain	0.1841	0.4506	0.901	0.431
					H/liver	0.1038	0.2544	0.921	0.181
					H/muscle	0.218	0.4944	0.894	0.417
								0.898	0.411
								0.895	0.543
								0.845	0.293
								0.855	0.519
								0.873	0.557
								0.928	0.444

User has to define calculation type of variable importances (selection of both simultaneously are allowed).

Sum coefficients for each descriptor – it is a very fast and very rough estimate (temporarily disabled).



We do not recommend to choose this mode due to very low adequacy of obtaining results. Due to this option is disabled now.

Permutation mode – it is a more time-consuming process (especially for very large sets of compounds). But obtaining results are highly adequate. This calculation based on estimation of influence of randomization of each descriptor values on out-of-bag prediction ability of the forest. The greater statistic values for out-of-bag set decrease the greater importance of the descriptor. Due to

randomness of permutation process it is more reliable to make several iterative calculation and average of obtained result.



Numbers of iterations is a fully arbitrary parameter. However we can give an advice – the more compounds in the training set the less number of iterations is needed. For huge data sets (about 1000 compounds and more) one iteration can be enough.

To view results of calculation switch to Variable importance tab.

Variables importance for each property is calculated separately.

3.2. Domain of applicability calculation

To calculate domain of applicability measures choose FOREST / CALC DOMAIN APPLICABILITY

The screenshot shows a software window titled "CF - D:\CF_example\MTL.txt" with a menu bar (File, Forest, Prediction, Options, About) and a toolbar. Below the toolbar are tabs: "List of forests", "Trees", "Forest Statistics", "Variable importance", and "Data". The "Variable importance" tab is active, displaying a table with the following data:

#	Trees count	Vars count	Min parent	Min child	Property	Risk estimate (ws)	Risk estimate (oob)	R2 (ws)	R2 (oob)
100	31	1	1		H/blood	0.5011	0.657	0.838	0.534
					H/fat	0.1244	0.2528	0.921	0.471
					H/brain	0.1841	0.4506	0.901	0.431
					H/liver	0.1038	0.2544	0.921	0.181
								0.894	0.417
								0.898	0.411
								0.895	0.543
								0.845	0.293
								0.855	0.519
								0.873	0.557
								0.928	0.444

Overlaid on the table is a dialog box titled "Choose DA type calculation". It contains the text "Choose domain applicability calculation type" and three radio button options: "based on trees predictions" (which is selected), "based on variable importance", and "based on proximities". An "OK" button is located at the bottom of the dialog.

In the opened dialog you can select desired domain applicability measure.

Measure **based on trees prediction** calculated by creation minimum-cost-tree. Distance s between pairs of training set compounds in models space are considered. That is each model has T number of predictions made by each tree in the model (T - total number of trees). Each prediction is considered as a separate dimension. Thus Euclidean distance can be calculated.

Measure **based on variable importance** is calculated by creation minimum-cost-tree. Euclidean distances between pairs of training set compounds in descriptors space are calculated, but additionally variables importance are considered. So the more important variable is the lesser variability of descriptor value is allowed. This procedure is more time-consuming than previous one.

Measure **based on proximities** is under testing and disabled now.



In all calculations of domain applicability only training set compounds having observed values are considered.

To change domain applicability ranges one should change the number in the field “DA in sigma units” (Forest statistics tab), which represents the coefficient k in the following equation (this coefficient can be a real non-negative number).

$$DA\ limit = mean\ distance\ value + k \times standard\ deviation\ distance\ value$$

After “Recalc” button clicked DA limit will be recalculated and all corresponding statistics too.

CF - D:\CF_example\saves\1.rf

File Forest Prediction About

List of forests | Trees | Forest Statistics | Variable importance | Data

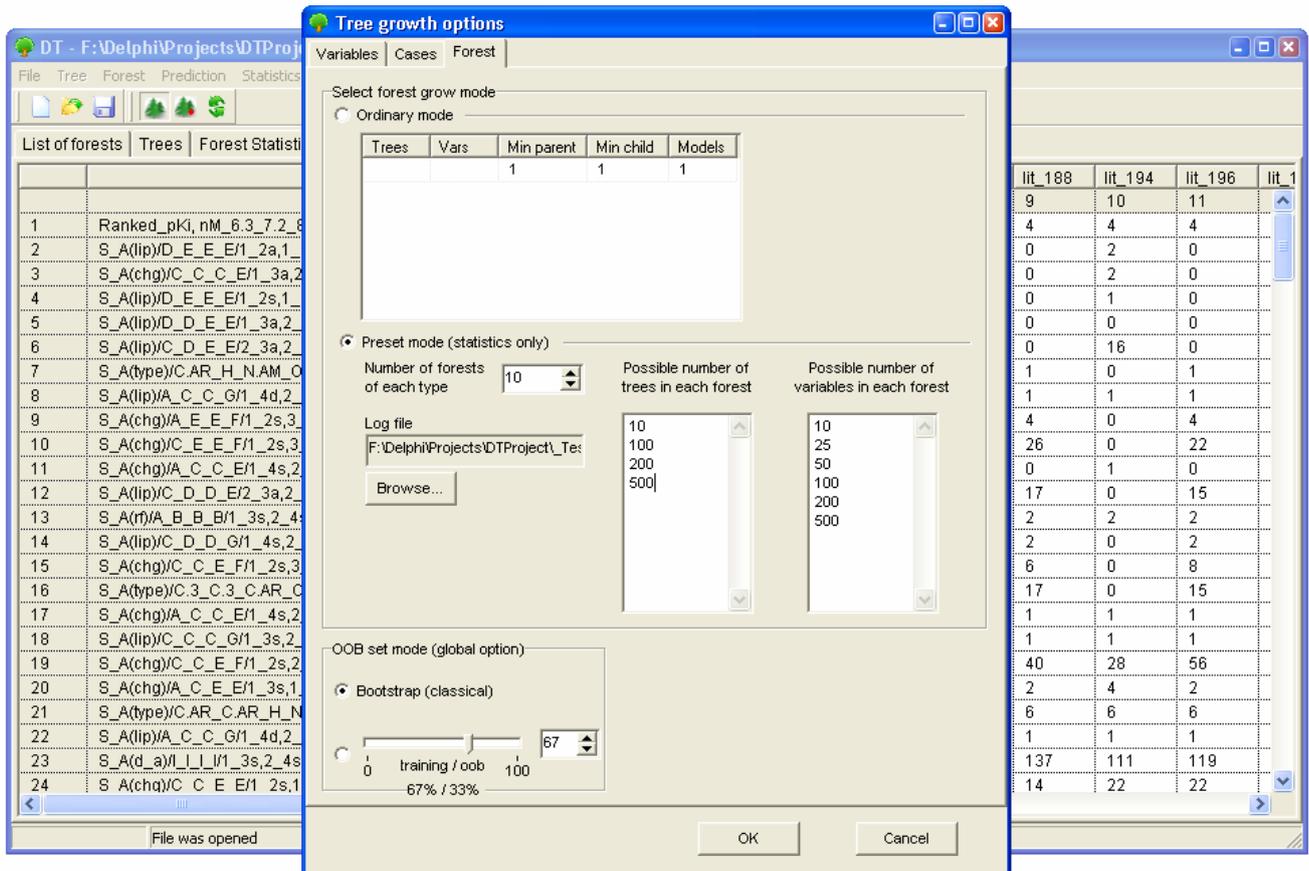
Property: log(IGC50-1) Set: ts2 DA in sigma units: 3 Recalc

#	Compound name	Set	Observed (log(IGC50-...))	Predicted (log(IGC50-...))	DA ...	Pred. StdDev (log(IGC50-...))
tp1086_ethyl_p...	ts2	1.67	0.724	+	0.468	
tp1087_a-chloro...	ts2	1.73	0.467	+	0.54	
tp1088_4-nitrop...	ts2	1.81	1.045	+	0.709	
tp1089_phenylp...	ts2	2.02	0.319	+	0.413	
tp1090_4-chloro...	ts2	2.11	1.499	-	1.159	
tp1091_1-bromo...	ts2	2.31	0.96	+	0.392	
tp1092_1_4-dic...	ts2	2.55	1.344	+	0.811	
tp1093_Benzyl...	ts2	2.74	1.45	-	1.301	
tp0984_3-hydro...	ts2	-1.019	-0.757	+	0.433	
tp0985_4-hydro...	ts2	-0.969	-0.356	+	0.59	
tp0986_4-amino...	ts2	-0.909	-0.322	+	0.637	
tp0987_Benzam...	ts2	-0.909	-0.257	+	0.387	
tp0988_Resorci...	ts2	-0.869	-0.392	+	0.796	
tp0989_4-aceta...	ts2	-0.819	0.051	+	0.779	

	WS	OOB	TS	TS2
R2	0.989	0.9	0.912	0.81
R2test	0.974	0.807	0.828	0.729
MSE	0.0293	0.215	0.192	0.293
RMSE	0.171	0.464	0.438	0.541
R2 (DA)	0.989	0.9	0.923	0.805
R2test (DA)	0.974	0.807	0.848	0.723
MSE (DA)	0.0293	0.215	0.159	0.282
RMSE (DA)	0.171	0.464	0.399	0.531
DA Coverage	1	1	0.962	0.973

DA calculation complete

4. “Preset mode” of model construction.



This option is needed to collect statistics of huge number of models on the base of predefined settings (possibility of saving of individual models is absent in this mode). This procedure is useful to investigate forest behavior in a wide range of setup variables (number of trees and number of descriptors).

There should be defined:

- number of models of each type;
- possible number of trees and descriptors for splitting (one value per line);
- log-file name, where all results are saved.



In this mode “Min parent” and “Min child” parameters equal 1 and cannot be changed.



Data is saved in the log-file as soon as it is produced. So there is no risk to lost data.

5. Model (files) routines.

5.1. Saving model

One can save model in a file by choosing FILE / SAVE PROJECT. Model saves into several separate files:

- 1) .rf file – has a plain text format and contains general information, which can be useful for user
- 2) .t file – has a binary file format and contains all trees composing the model
- 3) .bin - has a binary file format and contains all data concerning the model and all statistics for training, OOB and test sets (information and statistics of external set doesn't save in the file)
- 4) .imp - has a binary file format and contains information concerning variable importances (if they are calculated of course)

All these files are needed for model opening and should be stored in the same directory.

If the source file of the data set is not an rfd-file then at saving one should specify rfd-file name (which will be contain a data set) and then rf-file name (which will be contain model information).



Rfd-file has an associated rfn-file of the same name. Both of them are store source data and needed to successful data loading.



If rfd-file was created once try to use only it to create new projects for the same data set. This can keep free space on HDD. Otherwise each time new rfd-dile will be created.

5.2. Opening model(s)

To open model use standard menu FILE / OPEN PROJECT

To open model it is necessary that data file (rfd-file) is in its initial directory (where it has been saved first time) or in the same directory with rf-file.

- 1) one can freely move models on the computer, if place of corresponding rfd-file will be initial
- 2) one can copy model to USB stick and transfer it to another computer, but it is necessary to copy all model files and associated rfd/rfn-files into the same directory

One could add saved models to the current forest list if they have identical associated data-file (rfd-file).

- 1) if one try to open model file and data-file name will be identical to already opened model then the new model will be added to the list.
- 2) if one model has been already opened than one can select menu FILE / ADD MODELS TO THE CURRENT LIST to proceed. In opened dialog only models having according associated data-file will be displayed. Selection of multiple files is allowed.

6. Prediction of compounds properties which are in an external data-file.

To make prediction of compounds in an external data-file select the desired model in the model list and choose menu PREDICTION / PREDICT DATA FROM FILE.

If the open file has a variable with the same name as a target property then this file will be recognized as an external test set and the corresponding statistics will be calculated.

After prediction process was complete new set named "ext1" will be added to the list of model sets on Forest statistics tab. There one can select this set from the list, or select certain property to look for detailed statistics. As results of external data prediction don't save to model file one can find it useful to copy and paste this information in external editor.

The screenshot shows the 'Forest Statistics' tab in the CF software. The 'Property' dropdown is set to 'H/blood' and the 'Set' dropdown is set to 'all'. The 'DA in sigma units' is set to 3. The 'Recalc' button is visible. The main table displays the following data:

#	Compound name	Set	Observed (H/blo...)	Predicted (H/blo...)	Pred. StdDev (H/blo...)
mtl_0001		ext1	-	0.89	1.11
mtl_0002		ext1	-	1.158	1.206
mtl_0011		ext1	0.61	0.752	0.466
mtl_0101		ext1	-1.099	0.31	0.356
mtl_0102		ext1	-0.249	0.316	0.362
mtl_0103		ext1	-	0.374	0.423
mtl_0104		ext1	0.87	0.568	0.576
mtl_0105		ext1	-	0.364	0.421
mtl_0106		ext1	-0.359	0.467	0.514
mtl_0107		ext1	-0.589	0.467	0.514
mtl_0108		ext1	-1.519	0.469	0.514
mtl_0109		ext1	-	1.395	0.582
mtl_0110		ext1	-	1.385	0.551

Below the table, the 'Set' dropdown is set to 'EXT1'. The summary table shows the following statistics:

	WS	OOB	EXT1
R2	0.84	0.738	0.84
R2 test	0.643	0.506	0.643
MSE	0.503	0.697	0.503
RMSE	0.71	0.835	0.71

7. General information.

To copy data from various lists and tables one can often use right-mouse clicking and choosing appropriate item in popup menu.

Current program version is displayed in window which is called via menu ABOUT.

8. Afterword.

Do not hesitate to contact us if you found mistakes, faults, unusual program behavior or program failure or had any questions or ideas to improve program algorithm or interface! Any advices are welcome and will be taking in consideration at next version development!

ABOUT	25
ADD MODELS TO THE CURRENT LIST	23
ADD TREES TO FOREST	19
CALC DOMAIN APPLICABILITY	20
CALC VAR IMPORTANCE	19
CLEAR FOREST LIST	15
DELETE FOREST	15
GROW FOREST	7

NEW FANDOM FOREST PROJECT	4
NEW PROJECT	4
OPEN PROJECT	23
Ordinary mode	11
PREDICT DATA FROM FILE	24
Preset mode	22
SAVE PROJECT	23