Random Forest for Environmental Data Mining

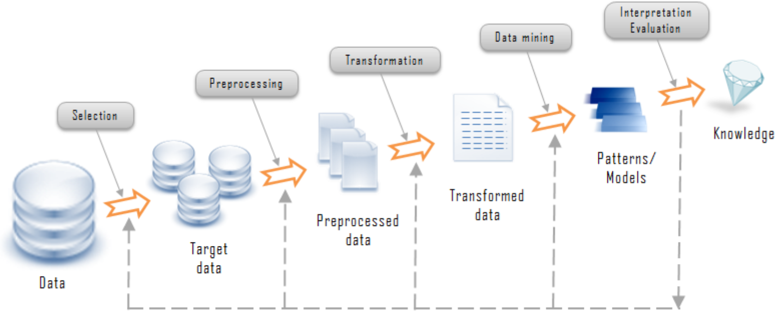
Michael Leuenberger, Carmen D. Vega Orozco
and Mikhaïl Kanevski

CRET - FGSE
University of Lausanne

June 25, 2013

1. Introduction
   - Environmental Data Mining
   - Case study : Forest Fires in a Random Forest
2. Method and Methodology
   - Random Forest
   - Properties
3. Results
   - Prediction
   - Susceptibility map
4. Conclusion

Environmental Data Mining



Data Mining

- Search for relevant patterns for decision making

Motivation

Why Random Forest ?

- Random Forest can deal with high dimensional database

Motivation

Why Random Forest ?

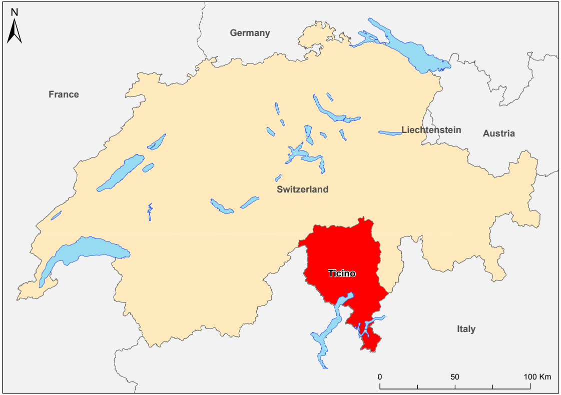
- Random Forest can deal with high dimensional database
- It is a powerful non-linear machine learning algorithm which can deal with the high complexity of phenomena

Motivation

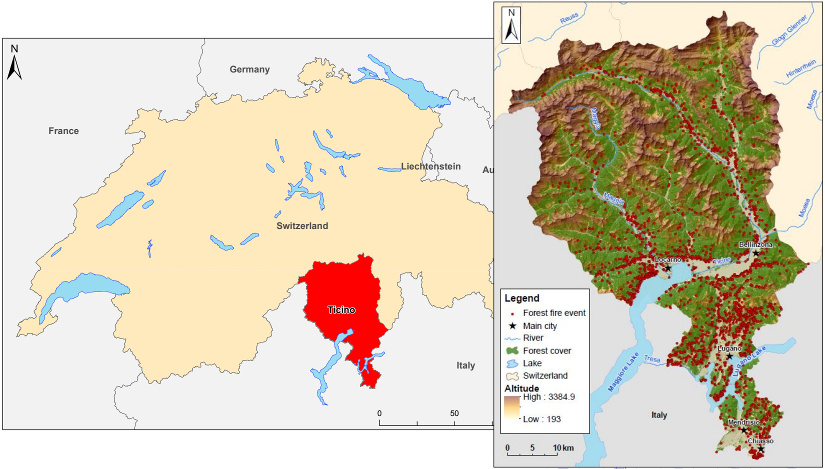
Why Random Forest ?

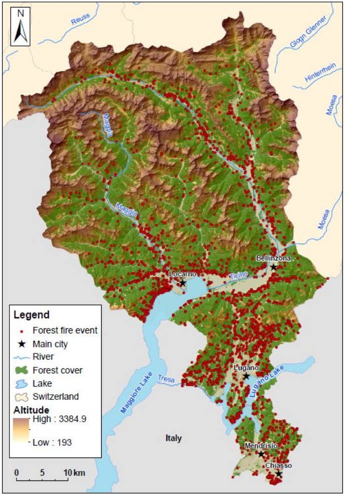
- Random Forest can deal with high dimensional database
- It is a powerful non-linear machine learning algorithm which can deal with the high complexity of phenomena
- It provides directly a measure for both errors and variable importances

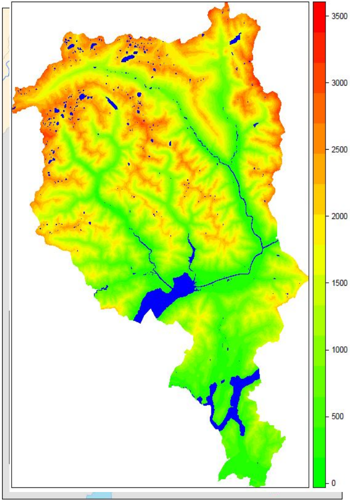
This study is focused on the 2'224 anthropogenic forest fires - ignition points - that have occured from 1969 to 2008. (*Canton Ticino, Swiss Alps, WSL-CH dataset*)



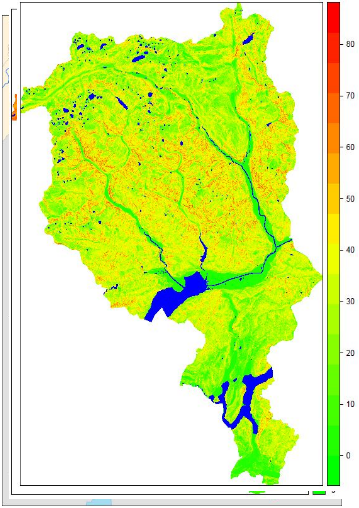
This study is focused on the 2'224 anthropogenic forest fires - ignition points - that have occured from 1969 to 2008. (*Canton Ticino, Swiss Alps, WSL-CH dataset*)



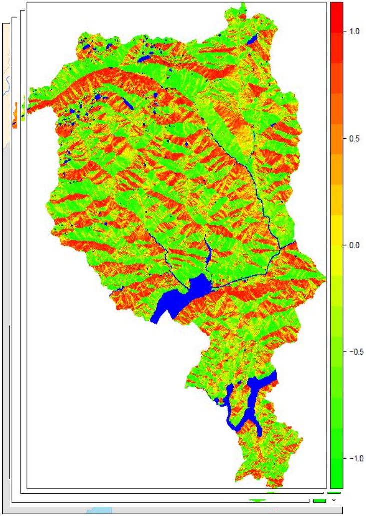

N
Reuss
Glogn Glenner
Hinterrhein
Moesa
Maggia
Maggia
Moesa
Bellinzona
Locarno
Ticino
Lugano
Lugano Lake
Tresa
Maggiore Lake
Italy
Mendrisio
Chiasso
Legend
● Forest fire event
★ Main city
River
Forest cover
Lake
Switzerland
Altitude
High : 3384.9
Low : 193
0 5 10 km



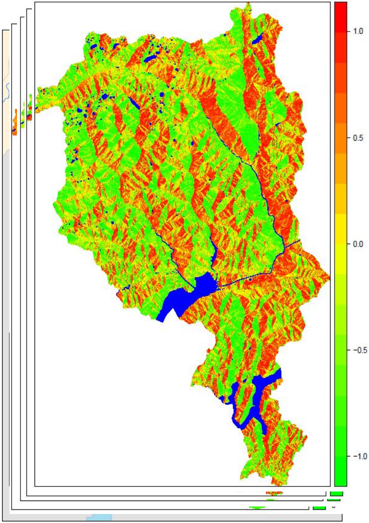
- Altitude



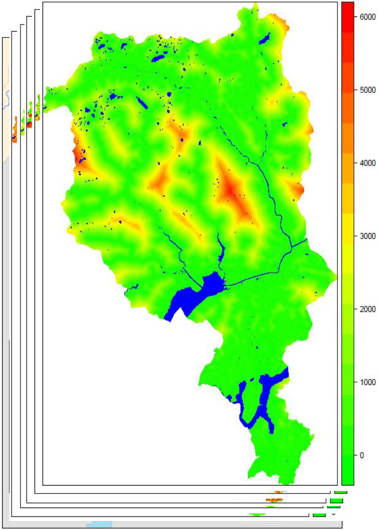
- Altitude
- Slope



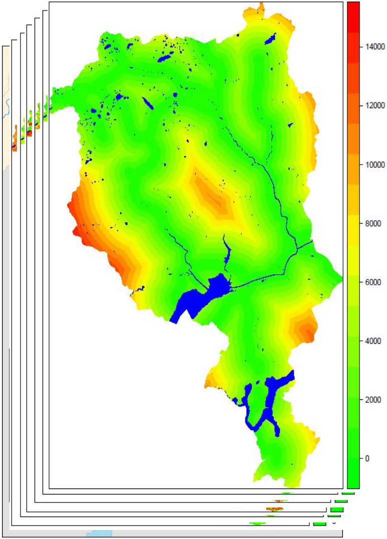
- Altitude
- Slope
- North aspect



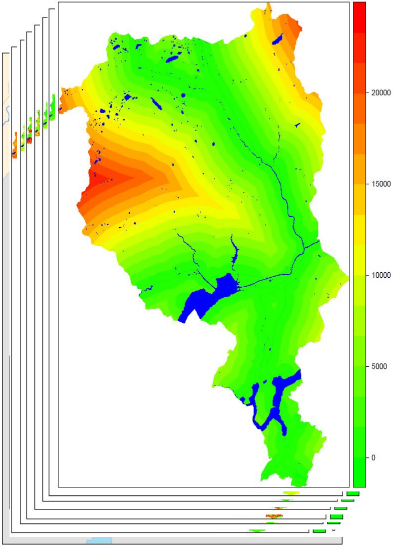
- Altitude
- Slope
- North aspect
- West aspect



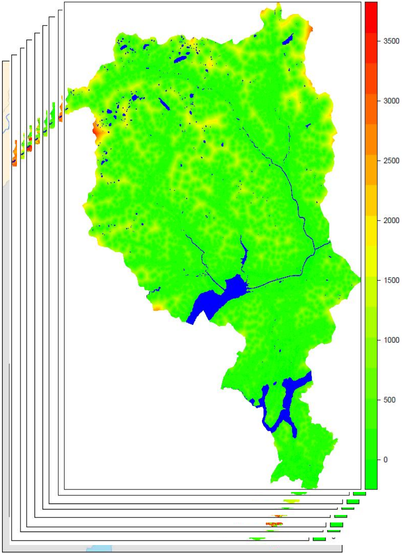
- Altitude
- Slope
- North aspect
- West aspect
- Dist. Streets



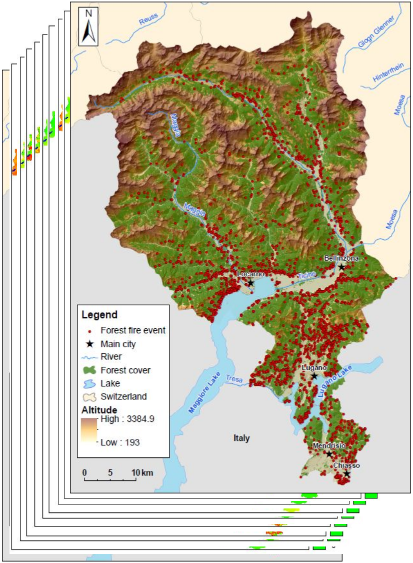
- Altitude
- Slope
- North aspect
- West aspect
- Dist. Streets
- Dist. Hightens



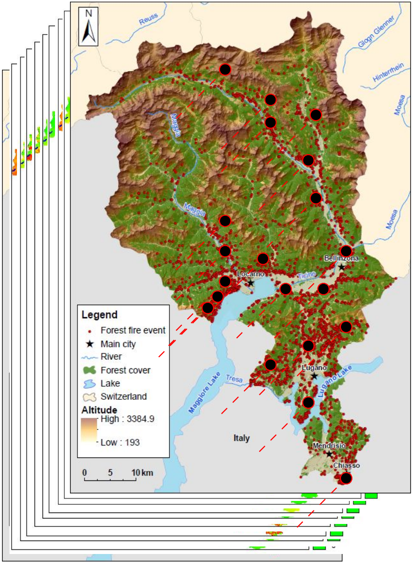
- Altitude
- Slope
- North aspect
- West aspect
- Dist. Streets
- Dist. Hightens
- Dist. Railways

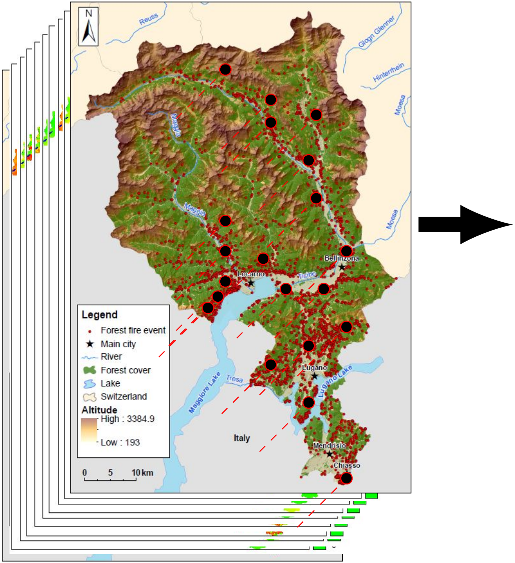


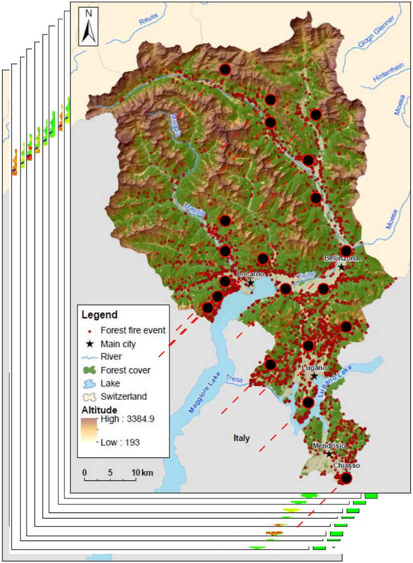
- Altitude
- Slope
- North aspect
- West aspect
- Dist. Streets
- Dist. Hightens
- Dist. Railways
- Dist. Buildings



- Altitude
- Slope
- North aspect
- West aspect
- Dist. Streets
- Dist. Hightens
- Dist. Railways
- Dist. Buildings


N
Reuss
Giogn Glenner
Hinterrhein
Moesa
Maggia
Bellinzona
Locarno
Ticino
Lugano
Lugano Lake
Tresa
Maggiore Lake
Italy
Mendrisio
Chiasso
Legend
Forest fire event
Main city
River
Forest cover
Lake
Switzerland
Altitude
High : 3384.9
Low : 193
0 5 10 km


N
Reuss
Giogn Glenner
Hinterrhein
Moesa
Maggia
Maggia
Moesa
Bellinzona
Locarno
Ticino
Legend
Forest fire event
Main city
River
Forest cover
Lake
Switzerland
Altitude
High : 3384.9
Low : 193
Tresa
Lugano
Lugano Lake
Maggiore Lake
Italy
Mendrisio
Chiasso
0 5 10 km



	V1	V2	V3	...
data1	●	●	●	●
data2	●	●	●	●
data3	●	●	●	●
data4	●	●	●	●
data5	●	●	●	●
data6	●	●	●	●
data7	●	●	●	●
data8	●	●	●	●
...	●	●	●	●

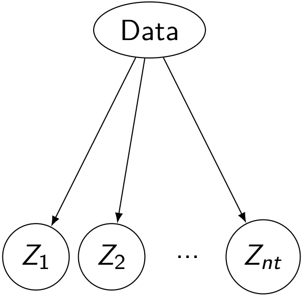
Random Forest

Data

	V1	V2	V3	V4	V5	...
data1	●	●	●	●	●	●
data2	●	●	●	●	●	●
data3	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data6	●	●	●	●	●	●
data7	●	●	●	●	●	●
data8	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data12	●	●	●	●	●	●
data13	●	●	●	●	●	●
data14	●	●	●	●	●	●
data15	●	●	●	●	●	●
...	●	●	●	●	●	●

Random Forest

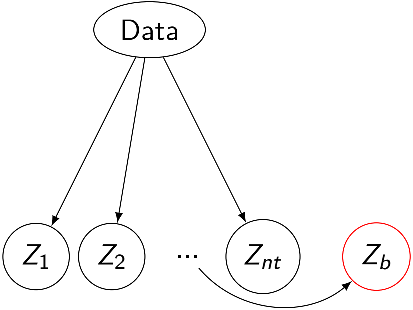
Bootstrapping



	V1	V2	V3	V4	V5	...
data1	●	●	●	●	●	●
data2	●	●	●	●	●	●
data3	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data6	●	●	●	●	●	●
data7	●	●	●	●	●	●
data8	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data12	●	●	●	●	●	●
data13	●	●	●	●	●	●
data14	●	●	●	●	●	●
data15	●	●	●	●	●	●
...	●	●	●	●	●	●

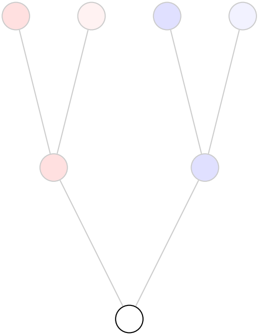
Random Forest

Bootstrapping



	V1	V2	V3	V4	V5	...
data1	●	●	●	●	●	●
data2	●	●	●	●	●	●
data3	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data6	●	●	●	●	●	●
data7	●	●	●	●	●	●
data8	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data12	●	●	●	●	●	●
data13	●	●	●	●	●	●
data14	●	●	●	●	●	●
data15	●	●	●	●	●	●
...	●	●	●	●	●	●

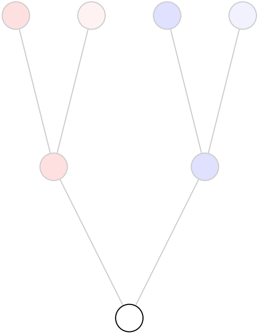
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

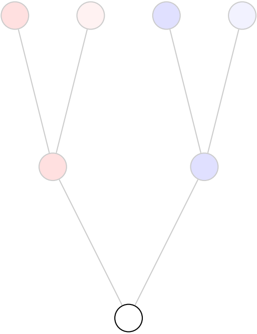
Random Forest



Z_b

	V1	V2	V3	V4	V5	…
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
…	●	●	●	●	●	●

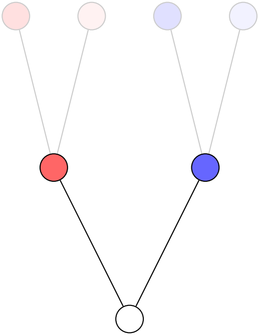
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

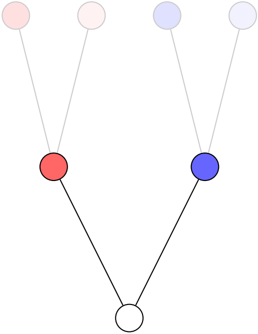
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

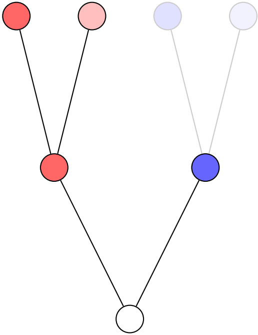
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

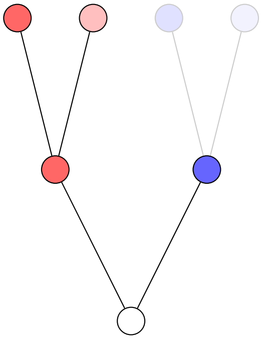
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

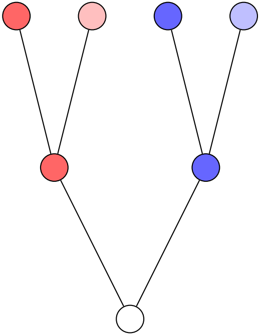
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

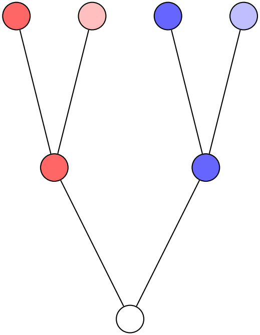
Random Forest



Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

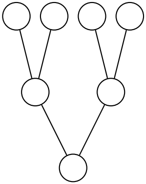
Random Forest



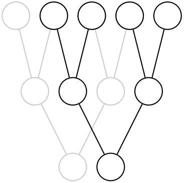
Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
data16	●	●	●	●	●	●
data17	●	●	●	●	●	●
data19	●	●	●	●	●	●
data19	●	●	●	●	●	●
...	●	●	●	●	●	●

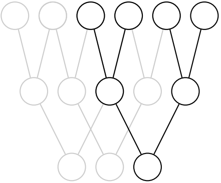
Random Forest



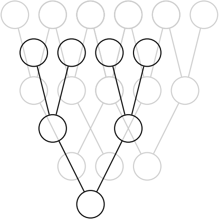
Random Forest



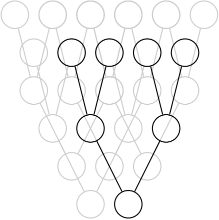
Random Forest



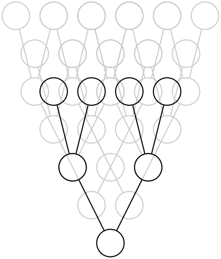
Random Forest



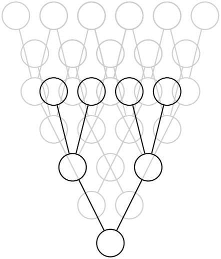
Random Forest



Random Forest

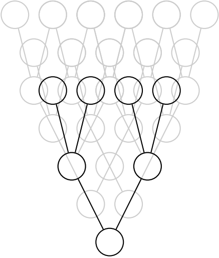


Random Forest



+

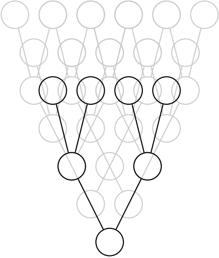
Random Forest



+

Bootstrapping
Variables selection

Random Forest

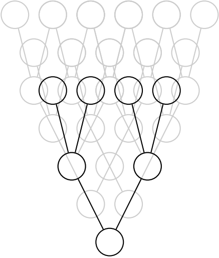


+

=

Bootstrapping
Variables selection

Random Forest



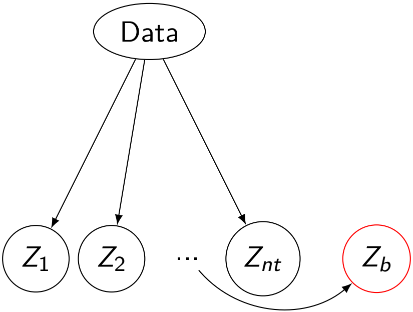
+

Bootstrapping
Variables selection

= **Random Forest**

OOB (Out-Of-Bag)

Bootstrapping



	V1	V2	V3	V4	V5	...
data1	●	●	●	●	●	●
data2	●	●	●	●	●	●
data3	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data6	●	●	●	●	●	●
data7	●	●	●	●	●	●
data8	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data12	●	●	●	●	●	●
data13	●	●	●	●	●	●
data14	●	●	●	●	●	●
data15	●	●	●	●	●	●
...	●	●	●	●	●	●

OOB (Out-Of-Bag)

OOB_b

	V1	V2	V3	V4	V5	...
data1	●	●	●	●	●	●
data3	●	●	●	●	●	●
data6	●	●	●	●	●	●
data8	●	●	●	●	●	●
data12	●	●	●	●	●	●
data13	●	●	●	●	●	●
data15	●	●	●	●	●	●
...	●	●	●	●	●	●

Z_b

	V1	V2	V3	V4	V5	...
data2	●	●	●	●	●	●
data4	●	●	●	●	●	●
data5	●	●	●	●	●	●
data5	●	●	●	●	●	●
data7	●	●	●	●	●	●
data9	●	●	●	●	●	●
data9	●	●	●	●	●	●
data10	●	●	●	●	●	●
data11	●	●	●	●	●	●
data14	●	●	●	●	●	●
data14	●	●	●	●	●	●
...	●	●	●	●	●	●

Measures in Random Forest

OOB error estimation

For each data we compute error over the trees where she appear OOB (E_i). By averaging we obtain an estimation of the global error.

$$OBB_{error} = \frac{1}{N} \sum_{i=1}^{N} E_i$$

Variable Importance Measurement

For each tree we compute the error obtained by the OOB data (E_b). Then we recompute the error after permuting all values in one variable (m) for all OOB data (E_b^m). By averaging over all the trees we obtain :

$$I_m = \frac{1}{B} \sum_{b \in B} E_b^m - E_b$$

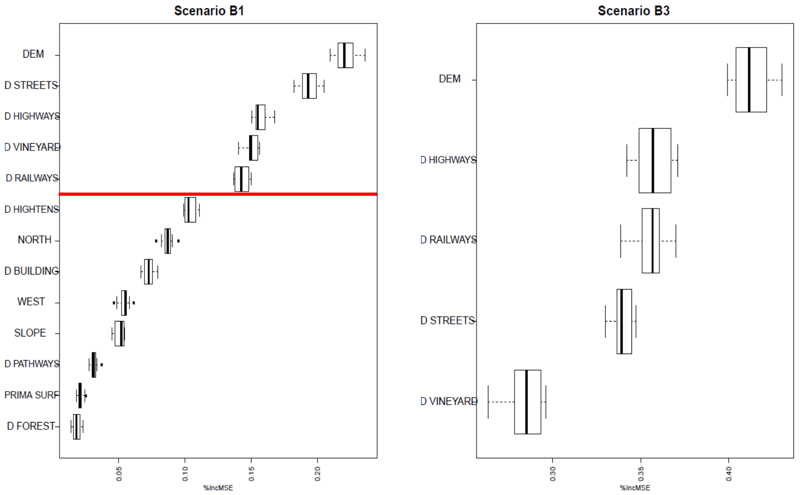
Illustrations

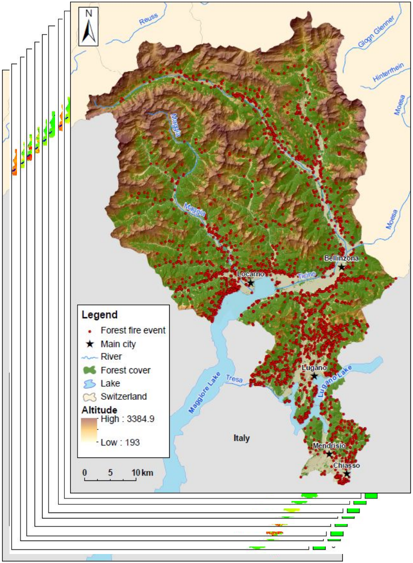
Error assessment

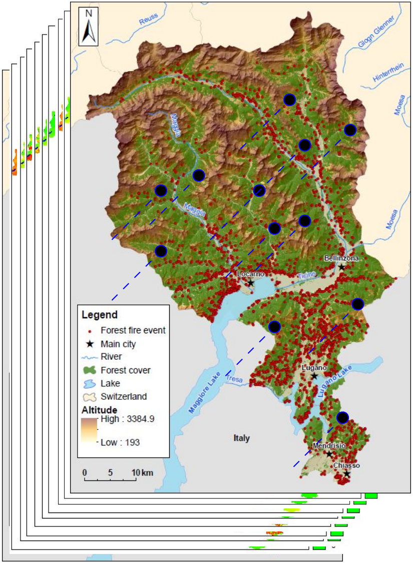
Scn	X, Y (variables)	MSE test set (sd)	MSE oob (sd)
A1	Yes (all variables)	1.309 (±0.0026)	1.299 (±0.0043)
A3	Yes (just 8 best)	1.334 (±0.0041)	1.316 (±0.0050)
B1	No (all variables)	1.316 (±0.0041)	1.333 (±0.0039)
B3	No (just 5 best)	1.363 (±0.0046)	1.381 (±0.0032)

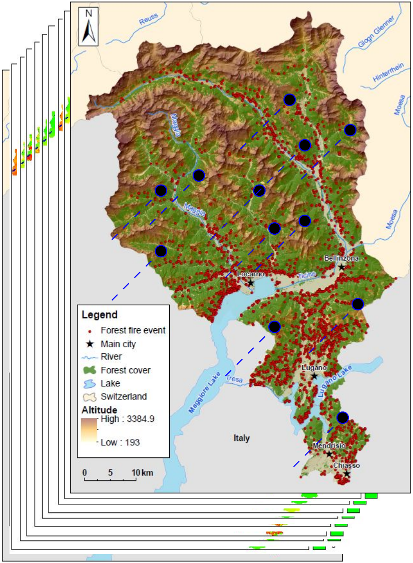
- MSE test $\approx$ MSE oob
- $MSE_{A1}, ..., MSE_{B3} \in [1.309, 1.363]$
- Large decrease in number of variables $\Longrightarrow$ slight increase in the error

Illustrations



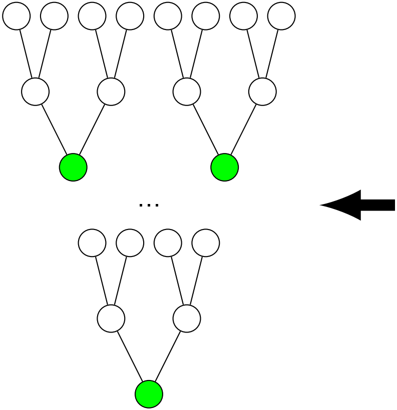

N
Reuss
Giogn Glenner
Hinterrhein
Moesa
Maggia
Ticino
Bellinzona
Locarno
Lugano
Lugano Lake
Tresa
Maggiore Lake
Italy
Mendrisio
Chiasso
Legend
Forest fire event
Main city
River
Forest cover
Lake
Switzerland
Altitude
High : 3384.9
Low : 193
0
5
10 km


N
Reuss
Giogn Glenner
Hinterrhein
Moesa
Maggia
Ticino
Bellinzona
Locarno
Lugano
Lugano Lake
Tresa
Maggiore Lake
Italy
Mendrisio
Chiasso
Legend
Forest fire event
Main city
River
Forest cover
Lake
Switzerland
Altitude
High : 3384.9
Low : 193
0
5
10 km



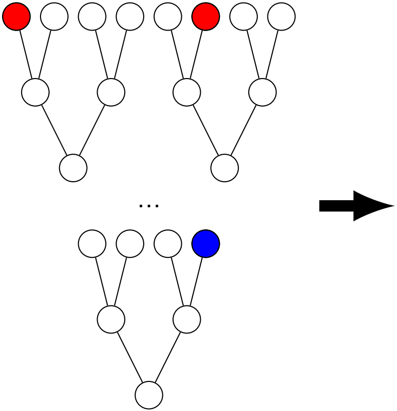
	V1	V2	V3	...
Newdata1	●	●	●	●
Newdata2	●	●	●	●
Newdata3	●	●	●	●
Newdata4	●	●	●	●
Newdata5	●	●	●	●
Newdata6	●	●	●	●
Newdata7	●	●	●	●
Newdata8	●	●	●	●
...	●	●	●	●

Prediction



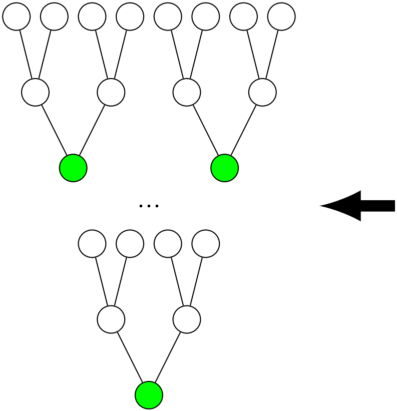
	V1	V2	V3	V4	V5	...
Newdata1	●	●	●	●	●	●
Newdata2	●	●	●	●	●	●
Newdata3	●	●	●	●	●	●
Newdata4	●	●	●	●	●	●
Newdata5	●	●	●	●	●	●
Newdata6	●	●	●	●	●	●
Newdata7	●	●	●	●	●	●
...	●	●	●	●	●	●

Prediction



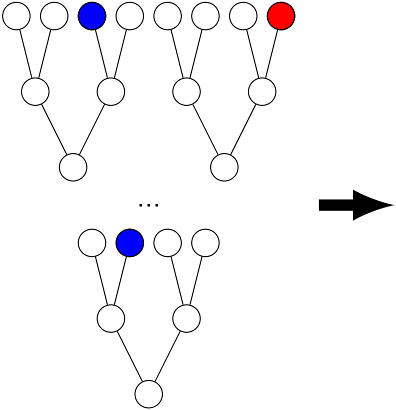
	V1	V2	V3	V4	V5	...
Newdata1	●	●	●	●	●	●
Newdata2	●	●	●	●	●	●
Newdata3	●	●	●	●	●	●
Newdata4	●	●	●	●	●	●
Newdata5	●	●	●	●	●	●
Newdata6	●	●	●	●	●	●
Newdata7	●	●	●	●	●	●
...	●	●	●	●	●	●

Prediction



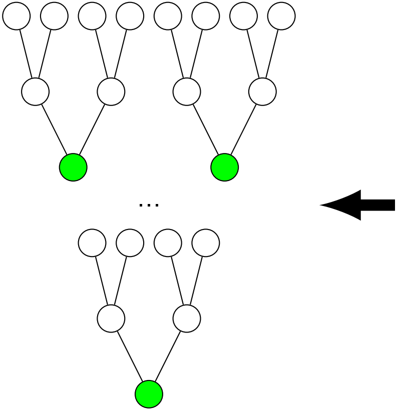
	V1	V2	V3	V4	V5	...
Newdata1	●	●	●	●	●	●
Newdata2	●	●	●	●	●	●
Newdata3	●	●	●	●	●	●
Newdata4	●	●	●	●	●	●
Newdata5	●	●	●	●	●	●
Newdata6	●	●	●	●	●	●
Newdata7	●	●	●	●	●	●
...	●	●	●	●	●	●

Prediction



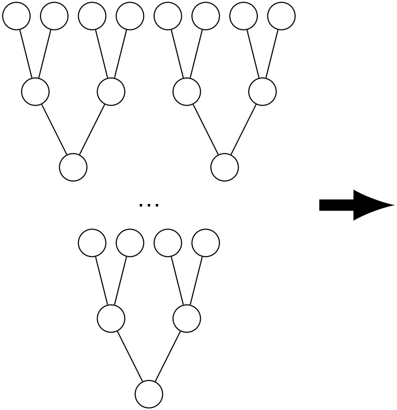
	V1	V2	V3	V4	V5	...
Newdata1	●	●	●	●	●	●
Newdata2	●	●	●	●	●	●
Newdata3	●	●	●	●	●	●
Newdata4	●	●	●	●	●	●
Newdata5	●	●	●	●	●	●
Newdata6	●	●	●	●	●	●
Newdata7	●	●	●	●	●	●
...	●	●	●	●	●	●

Prediction



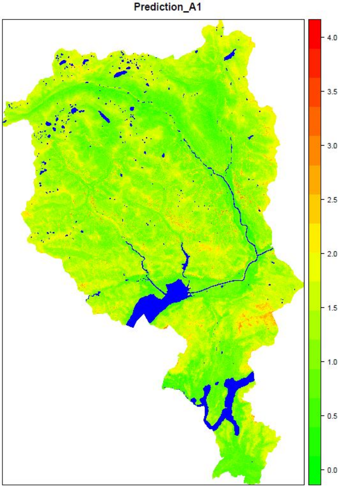
	V1	V2	V3	V4	V5	...
Newdata1	●	●	●	●	●	●
Newdata2	●	●	●	●	●	●
Newdata3	●	●	●	●	●	●
Newdata4	●	●	●	●	●	●
Newdata5	●	●	●	●	●	●
Newdata6	●	●	●	●	●	●
Newdata7	●	●	●	●	●	●
...	●	●	●	●	●	●

Prediction



	V1	V2	V3	V4	V5	...
Newdata1	●	●	●	●	●	●
Newdata2	●	●	●	●	●	●
Newdata3	●	●	●	●	●	●
Newdata4	●	●	●	●	●	●
Newdata5	●	●	●	●	●	●
Newdata6	●	●	●	●	●	●
Newdata7	●	●	●	●	●	●
...	●	●	●	●	●	●

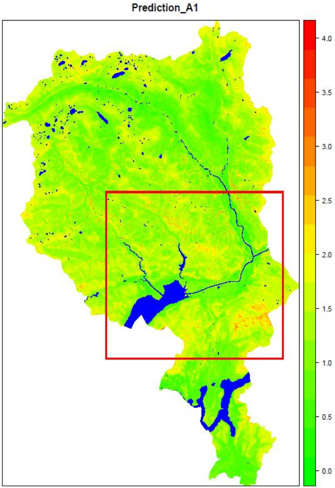
Scenario A1

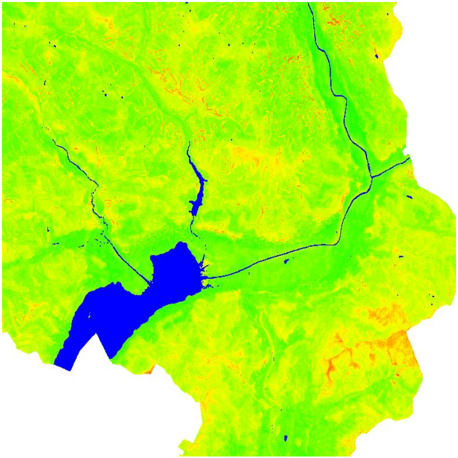


15 Variables :

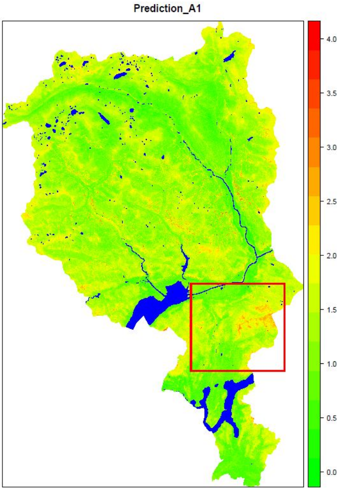
- Altitude (DEM)
- X,Y Coordinates
- Slope
- North, West aspect
- Dist Streets, Pathways
- Dist Railways, Highways
- Dist Hightens, Building
- Dist Forest, Vineyard
- Primary surface

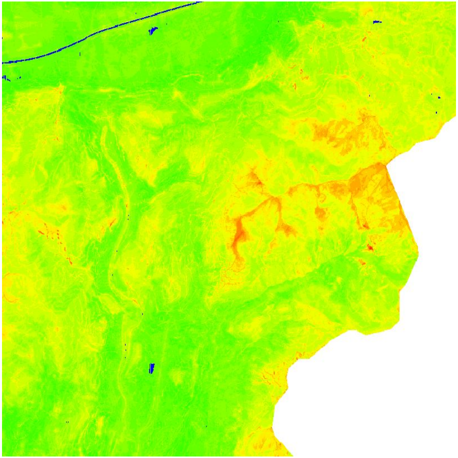
Scenario A1



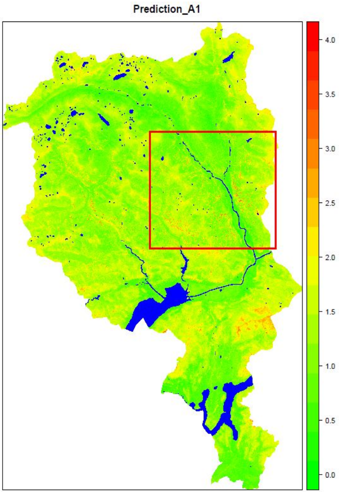


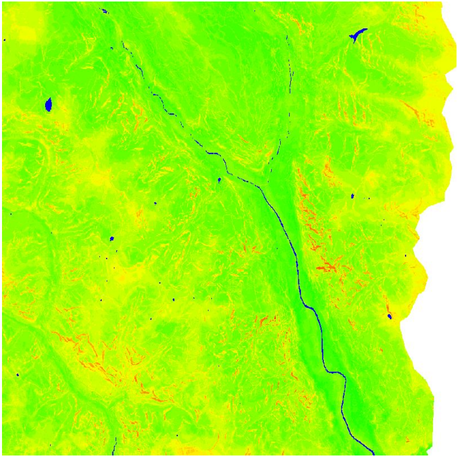
Scenario A1





Scenario A1





Conclusion

Future research

- Further analysis to optimize hyper-parameters

Conclusion

Future research

- Further analysis to optimize hyper-parameters
- Construction of larger input space

Conclusion

Future research

- Further analysis to optimize hyper-parameters
- Construction of larger input space
- Application of other ensemble learning methods and one-class classification methods

Conclusion

Future research

- Further analysis to optimize hyper-parameters
- Construction of larger input space
- Application of other ensemble learning methods and one-class classification methods
- Application to other case studies : Landslide, Permafrost, Avalanche,...

Thank you for your attention !

Acknowledgments

The research was partly supported by the Swiss NSF Project No. 200021-140658, "Analysis and modelling of space-time patterns in complex regions".