Assessing the effect of spatial autocorrelation in predicting groundwater salinity with Machine Learning

Tziachris, P., Arampatzis, G., Aschonitis, V., Sachsamanoglou, E., P. Tziritis, E.

Soil and Water Resources Institute (SWRI), Hellenic Agricultural Organization Demeter (ELGO-DIMITRA)

Machine learning

ML goal:

Predictions to new datasets

(e.g. different time, space etc)

However…

Be careful of the overfitting.

Machine learning

Two cases in this presentation:

1. Whether we can trust models based on training-testing prediction results, in case of strong linear correlations.
2. If spatial autocorrelation interferes with the data, creating biased ML models

Datasets

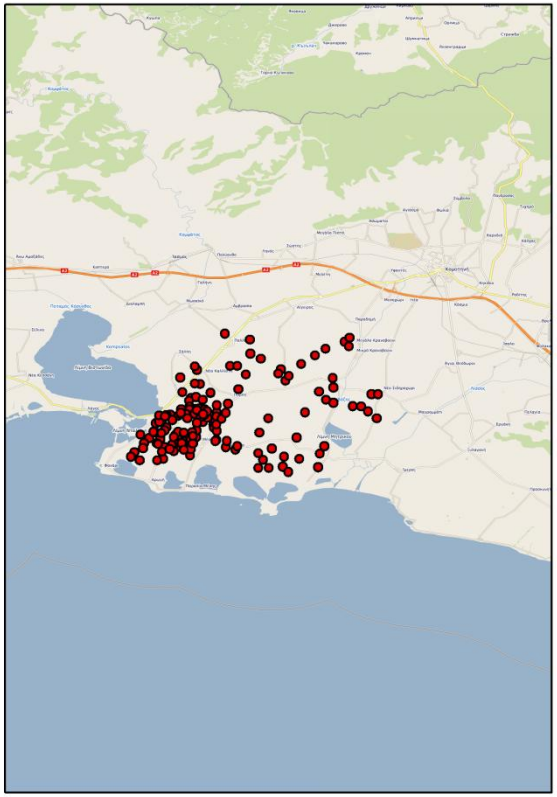
MEDSAL

"Salinization of critical groundwater reserves in coastal Mediterranean areas: Identification, Risk Assessment and Sustainable Management with the use of integrated modelling and smart ICT tools"

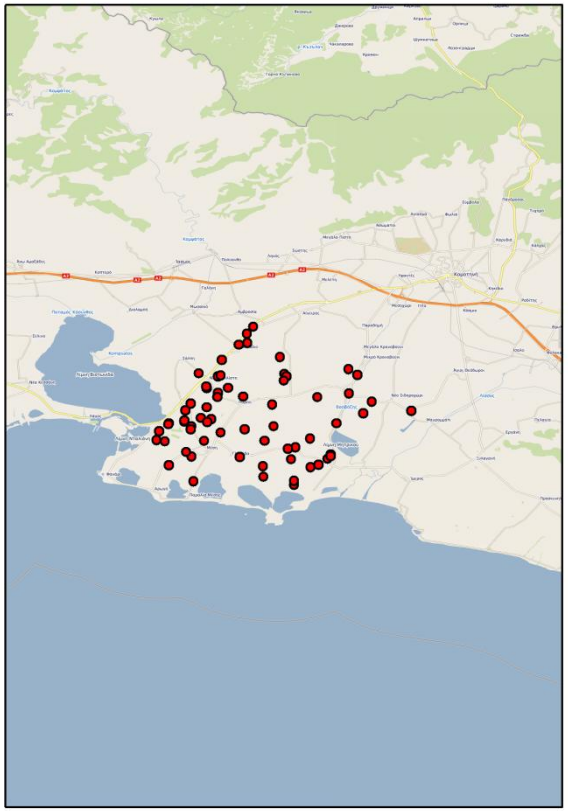
MEDSAL areas

Datasets

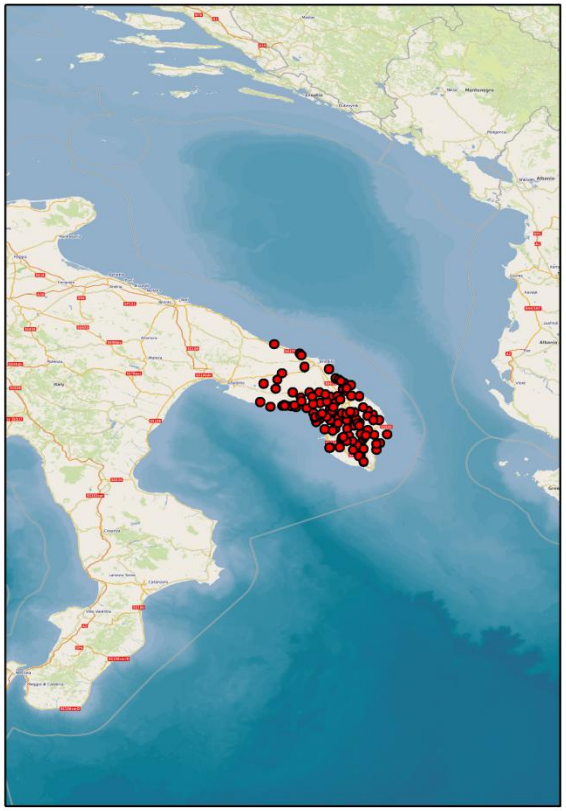
A. Greece, July 2020 (147 points)



B. Greece, June 2019 (65 points)



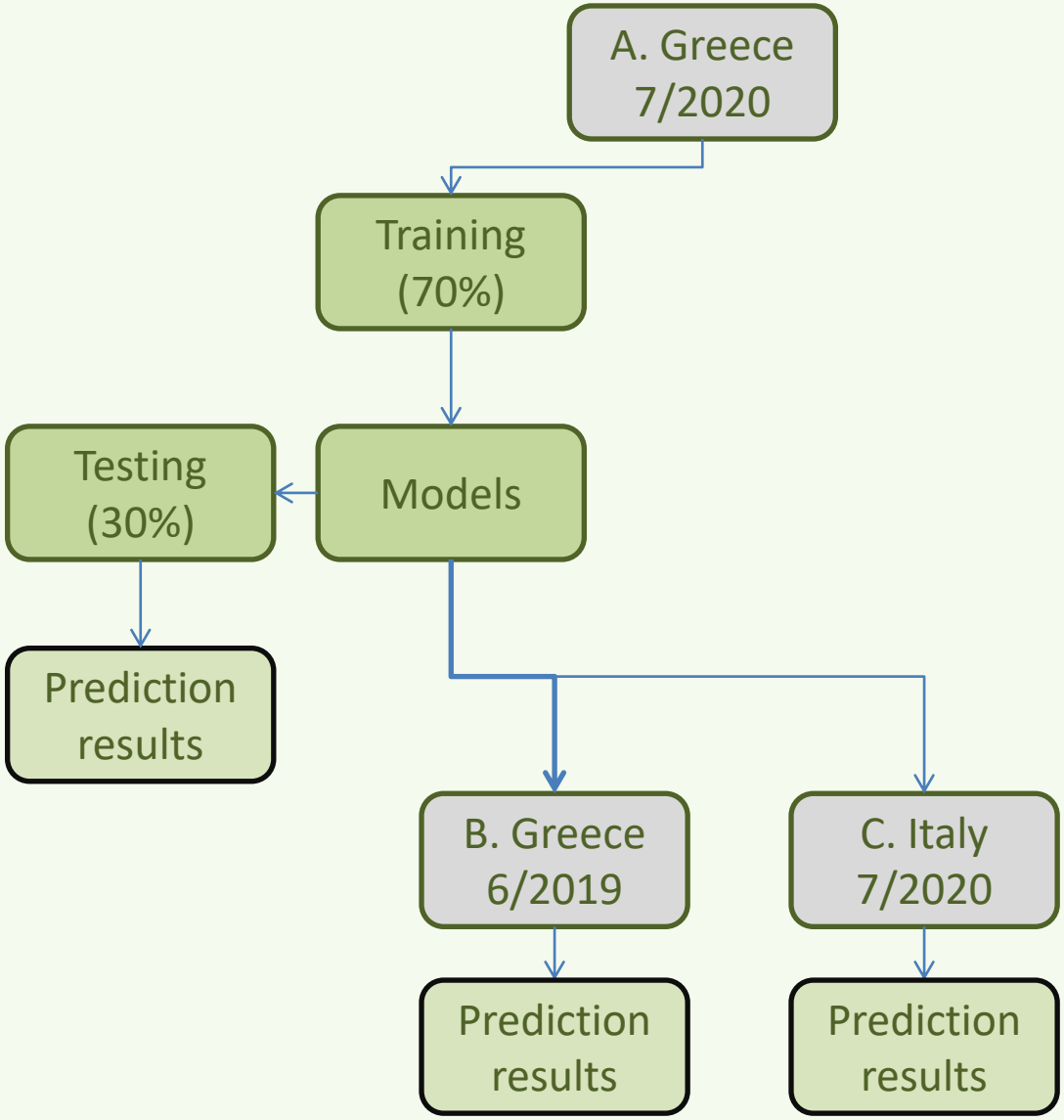
C. Italy, July 2020 (121 points)



Linear regression VS Machine Learning

No spatial parameters

Workflow


A. Greece 7/2020
Training (70%)
Testing (30%)
Models
Prediction results
B. Greece 6/2019
C. Italy 7/2020
Prediction results
Prediction results

Algorithms

- Linear regression (LN)

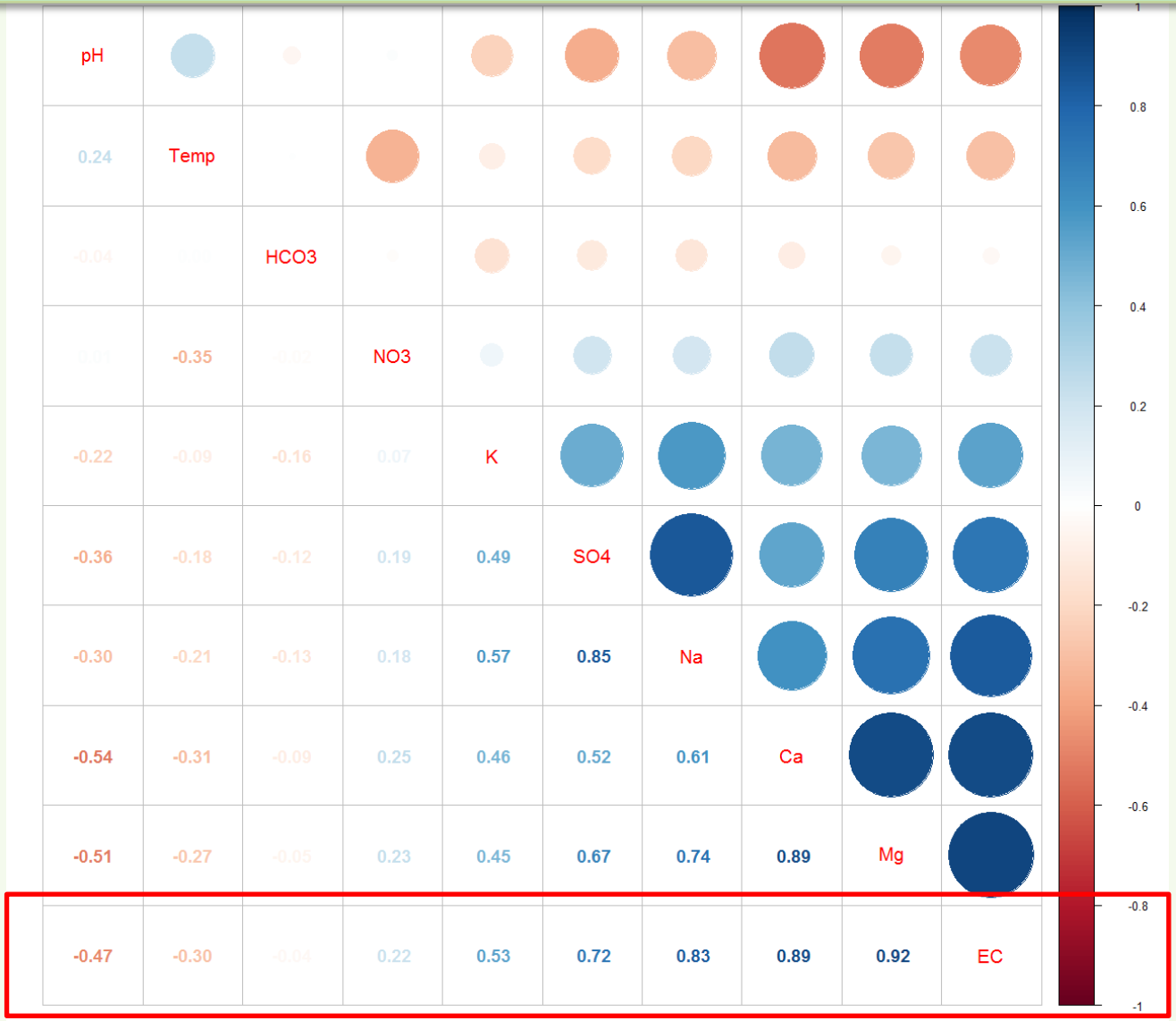
ML – no tuning:

- Random forest (RF)
- Quantile random forest (QRF)
- Extreme gradient boosting (XGB)

Variables

	Covariates	Unit	Method
1	Electric conductivity (EC)	µS/cm	Measured in situ using the YSI ProDSS Multiparameter portable equipment
2	Temperature	C	Measured in situ using the YSI ProDSS Multiparameter portable equipment
3	Bicarbonate **(HCO_3^-)**	mg/L	Measured in the lab using the titration method with neutralization of HCl
4	Nitrate **(NO_3^-)**	mg/L	Measured in the lab using spectrophotometer
5	**pH**	-	Measured in the lab using electrodes
6	Potassium **(K^+)**	mg/L	Measured in the lab using a flame photometer
7	Sulphate **($SO4^{2-}$)**	mg/L	Measured in the lab using spectrophotometer
8	Sodium **(Na^+)**	mg/L	Measured in the lab using a flame photometer
9	Calcium **(Ca^{2+})**	mg/L	Measured in the lab using atomic absorption spectrometer (AAS flame)
10	Magnesium **(Mg^{2+})**	mg/L	Measured in the lab using atomic absorption spectrometer (AAS flame)

Variables


pH
0.24 Temp
-0.04 HCO3
0.01 -0.35 NO3
-0.22 -0.09 -0.16 0.07 K
-0.36 -0.18 -0.12 0.19 0.49 SO4
-0.30 -0.21 -0.13 0.18 0.57 0.85 Na
-0.54 -0.31 -0.09 0.25 0.46 0.52 0.61 Ca
-0.51 -0.27 -0.05 0.23 0.45 0.67 0.74 0.89 Mg
-0.47 -0.30 -0.04 0.22 0.53 0.72 0.83 0.89 0.92 EC
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Results

RMSE	A	B	C
LN	457.90	1143.41	4162.12
QRF	375.09	1250.17	9318.26
RF	389.70	1192.26	9932.52
XGB	685.13	1315.49	9028.03

R^2	A	B	C
LN	0.991	0.912	0.971
QRF	0.994	0.897	0.914
RF	0.994	0.901	0.862
XGB	0.979	0.882	0.894

MAE	A	B	C
LN	345.32	622.19	1601.22
QRF	261.86	723.38	3392.98
RF	288.26	710.47	3626.20
XGB	495.01	805.16	3312.17

A. Greece, July 2020
B. Greece, June 2019
C. Italy, July 2020

Results

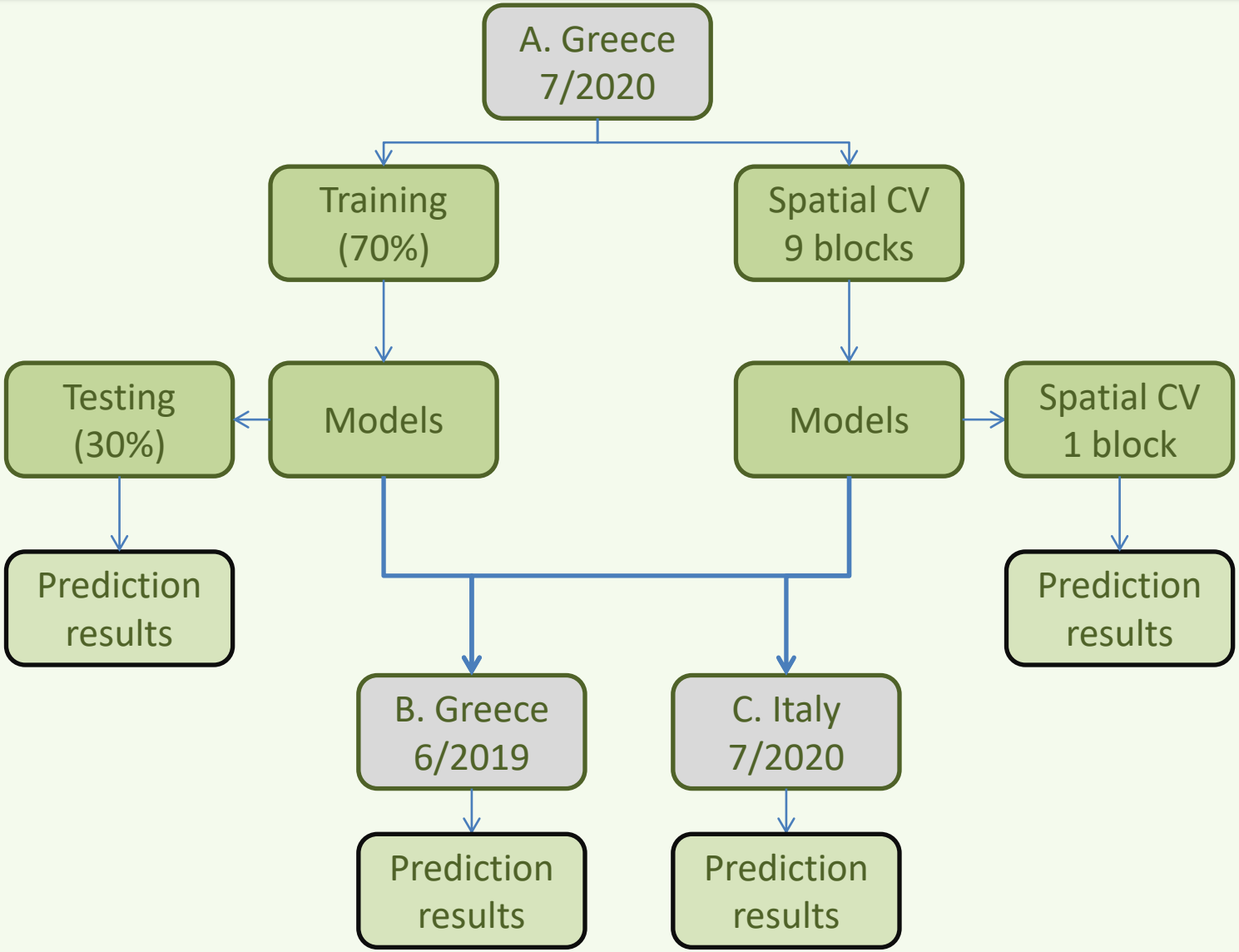
- In case of strong correlation between variables, **Linear Regression** is the best choice even if it has worst results in the training-testing dataset (overfitting).
- You get:
  - Better prediction results with different datasets
  - Interpretability

Spatial autocorrelation and ML

Machine learning spatial autocorrelation

- Robust and efficient ML models should be trained with data that **are independent and identically distributed** (i.i.d.) or else
  - Overfitting,
  - Bias assessment of the model's capability to generalize the learned relationship to new data
- However, **spatial data** are special kind of data that this assumption does not always hold due to their **spatial autocorrelation**

Workflow


A. Greece 7/2020
Training (70%)
Spatial CV 9 blocks
Testing (30%)
Models
Models
Spatial CV 1 block
Prediction results
Prediction results
B. Greece 6/2019
C. Italy 7/2020
Prediction results
Prediction results

Algorithms

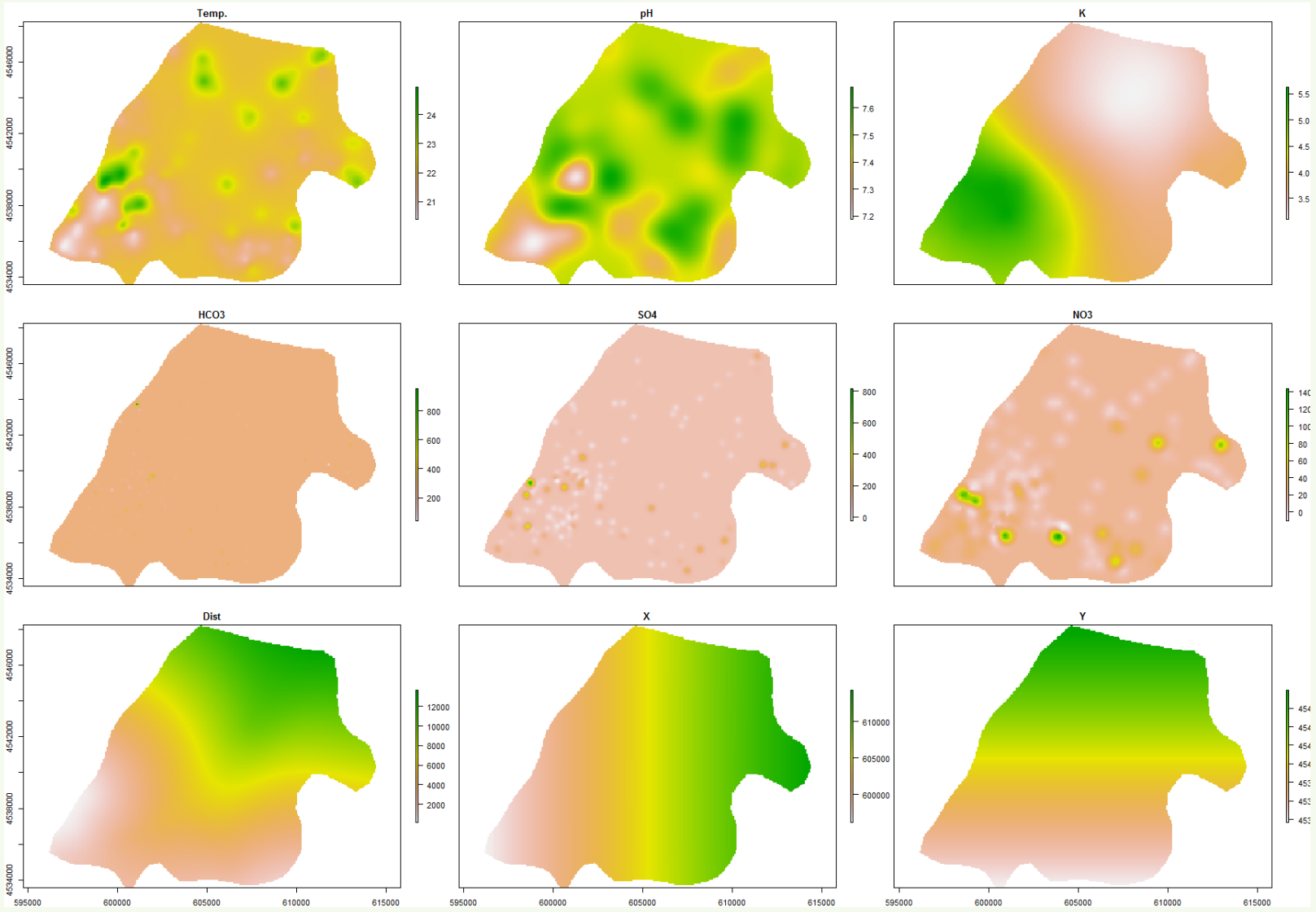
ML – no tuning:

- Random forest (RF)
- Quantile random forest (QRF)
- Extreme gradient boosting (XGB)

Variables

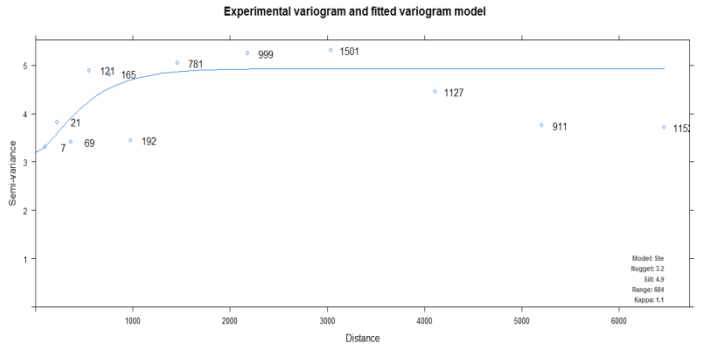
	Covariates	Unit	Method
1	Electric conductivity (EC)	µS/cm	Measured in situ using the YSI ProDSS Multiparameter portable equipment
2	Temperature	C	Measured in situ using the YSI ProDSS Multiparameter portable equipment
3	Bicarbonate **(HCO_3^-)**	mg/L	Measured in the lab using the titration method with neutralization of HCl
4	Nitrate **(NO_3^-)**	mg/L	Measured in the lab using spectrophotometer
5	**pH**	-	Measured in the lab using electrodes
6	Potassium **(K^+)**	mg/L	Measured in the lab using a flame photometer
7	Sulphate **($SO4^{2-}$)**	mg/L	Measured in the lab using spectrophotometer
8	Distance from sea (dist)	meters	Estimated from data
9	x	meters	Coordinates estimated from data
10	y	meters	Coordinates estimated from data

Variables - interpolation

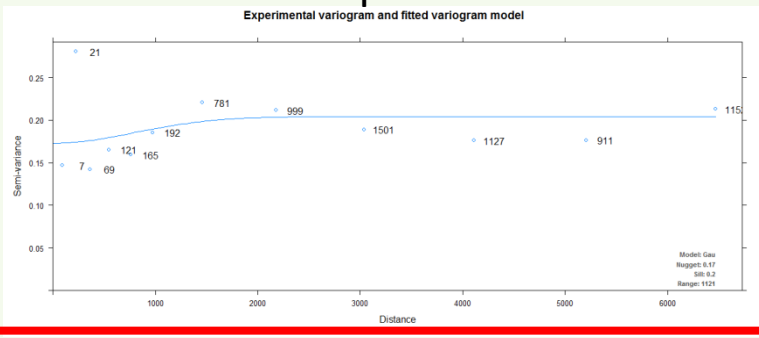


Variables – semivariograms

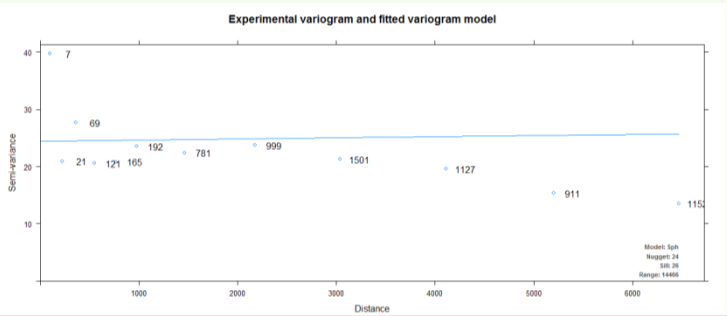
Temperature



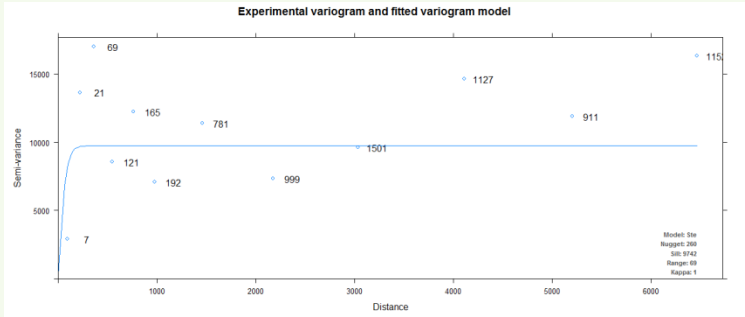
pH



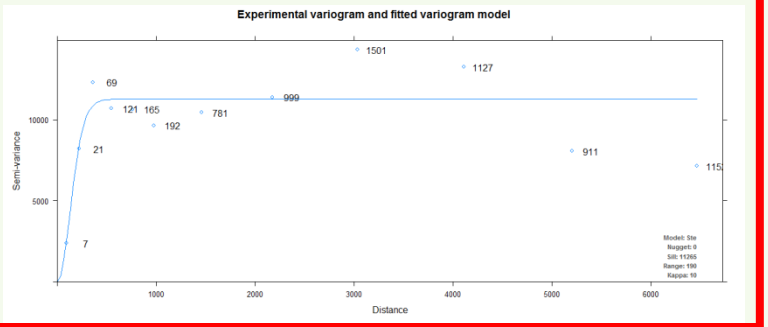
K



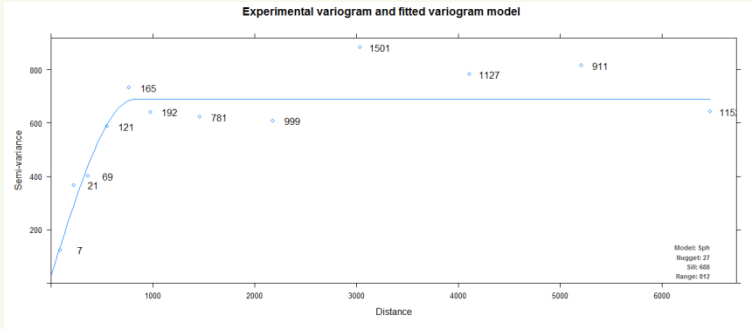
HCO3



SO4

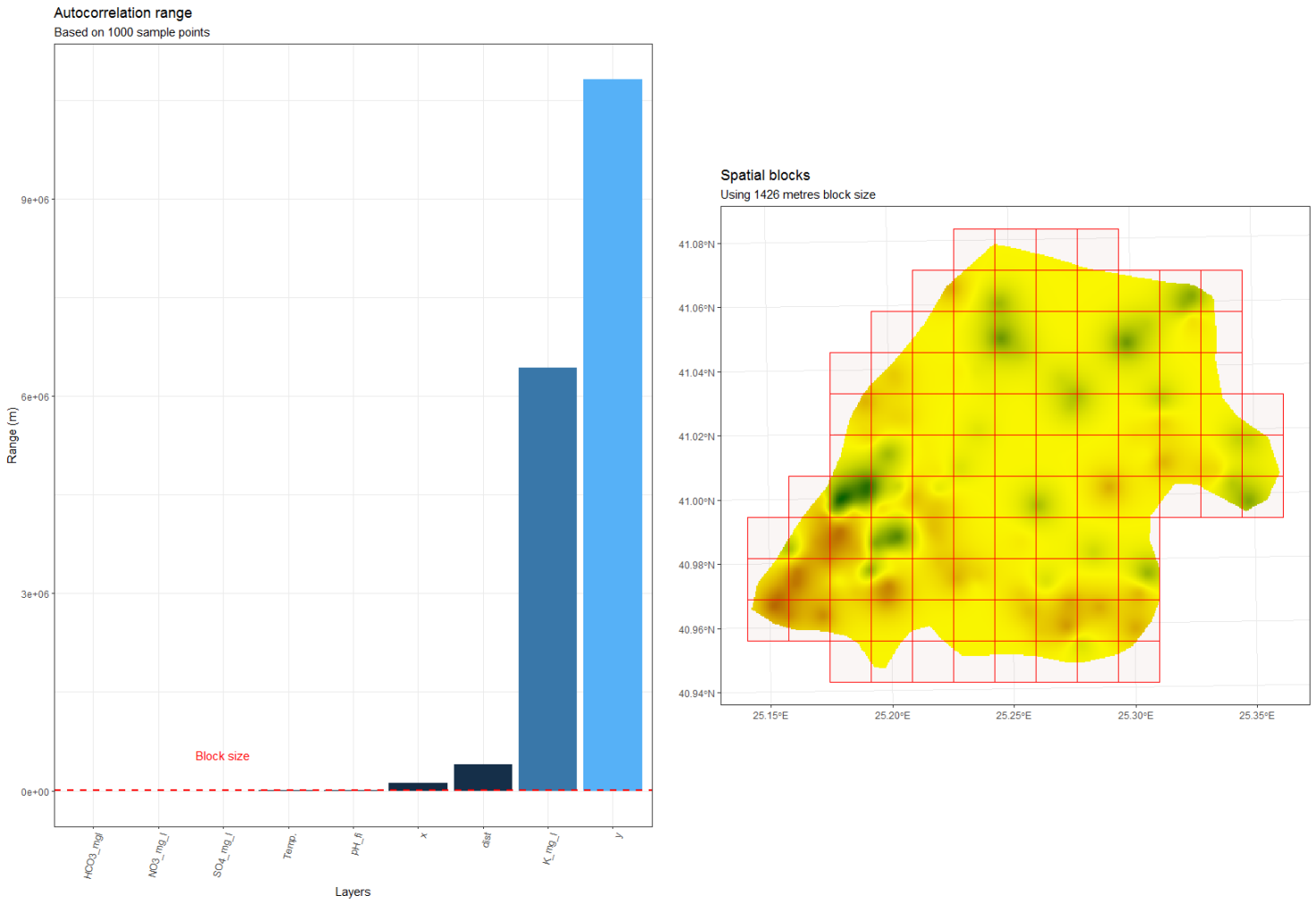


NO3

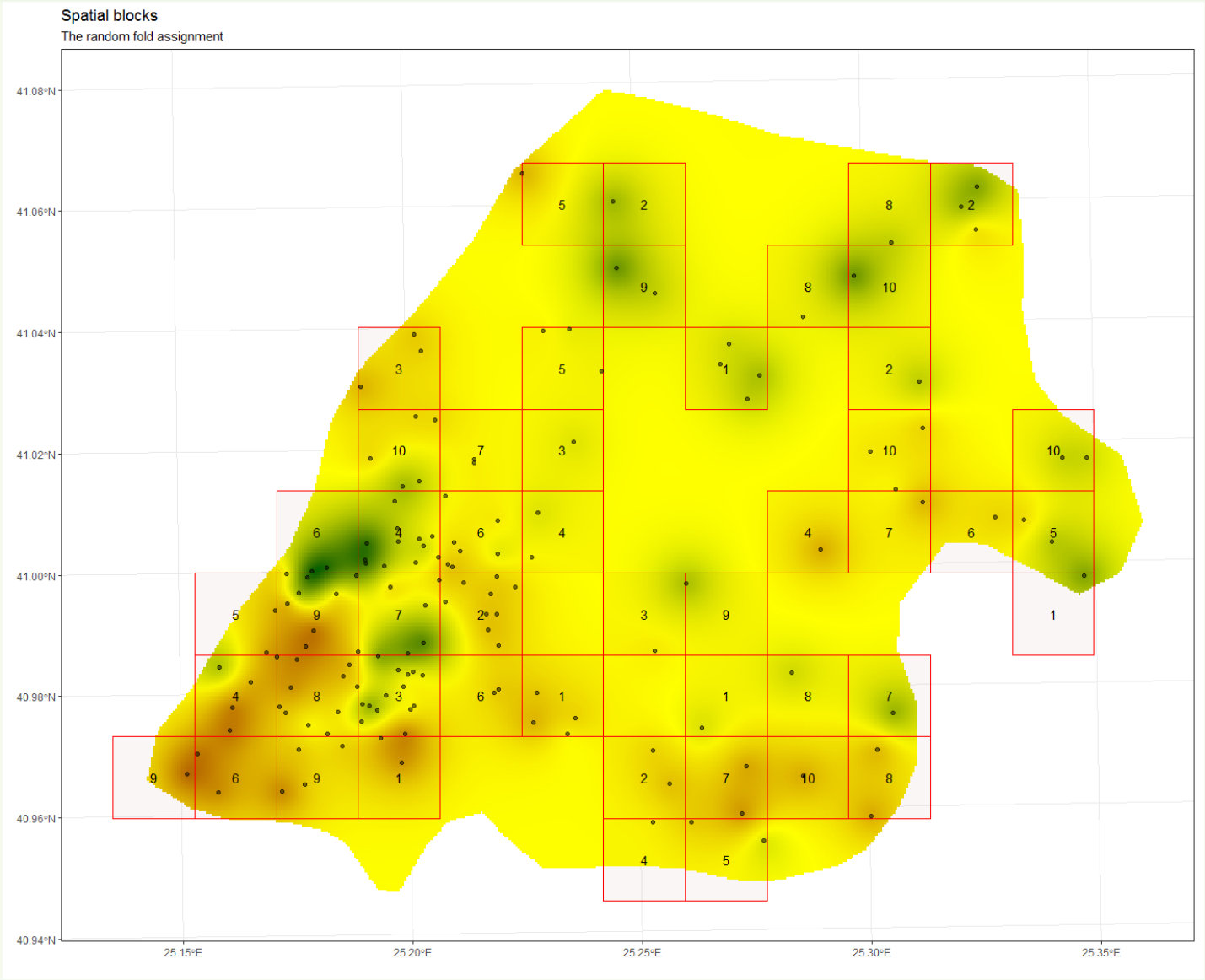


Variables - autocorrelation range

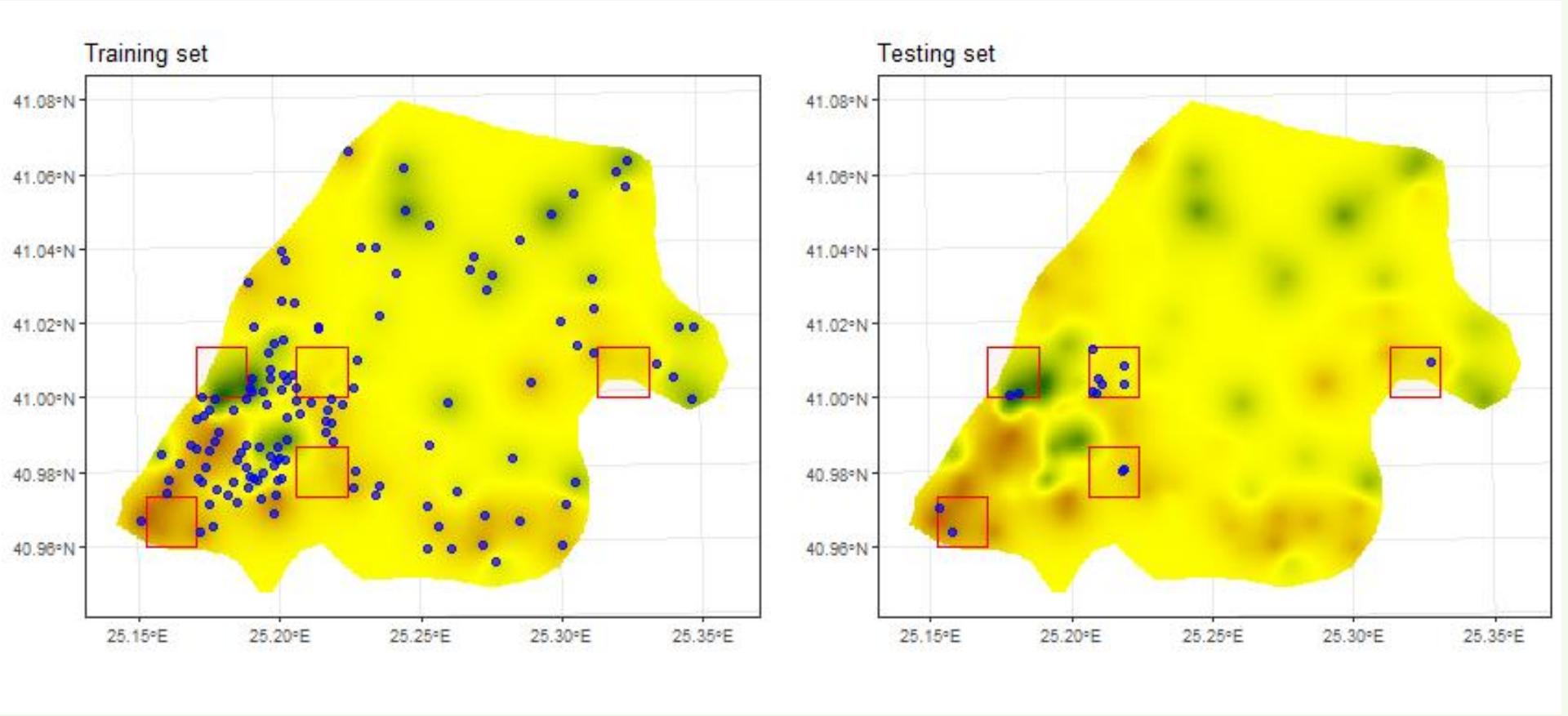
Block CV



Variables - spatial blocks



Variables – sample spatial fold



Results

Random split

RMSE	A	B	C
QRF	1993.95	2236.05	11251.38
RF	2131.56	1996.11	11651.82
XGB	2412.84	2316.77	11081.31

R^2	A	B	C
QRF	0.789	0.696	0.562
RF	0.762	0.721	0.503
XGB	0.665	0.677	0.540

MAE	A	B	C
QRF	1300.20	1395.60	6228.82
RF	1527.33	1414.00	7453.50
XGB	1546.68	1543.84	7231.58

Spatial split

RMSE	A	B	C
QRF	2655.63	2157.02	11611.89
RF	2699.40	1864.35	11992.96
XGB	2575.22	2263.28	11935.79

R^2	A	B	C
QRF	0.662	0.720	0.544
RF	0.624	0.764	0.475
XGB	0.646	0.720	0.603

MAE	A	B	C
QRF	1572.03	1323.18	5887.62
RF	1841.38	1299.85	6849.17
XGB	1593.91	1412.30	6781.84

A. Greece, July 2020
B. Greece, June 2019
C. Italy, July 2020

Results

- **Random split**: better results in the training-testing dataset (same dataset). In new datasets?
- **Spatial split:** could produce better results in new datasets in case of spatial autocorrelation (In our case we have mixed results)

Future work

- Assess more cases (datasets from other time/places, stronger spatial autocorrelation)
- Spatial tuning (hyperparameters optimization)
- Compare different spatial split methods

Be careful with ML overfitting !

Thank you