Imputeren en beoordelen meetreeksen RIWA-base

Aart Smits (eauQstat), Eit C.J. van der Meulen (AMO)
Gerrit van de Haar (RIWA), Paul K. Baggelaar (Icastat)

Waar gaat het om?

- RIWA
- Meetnet Rijn en Maas
- RIWA-base
- Jaarrapportages (al circa 45 jaar)
- Kengetallen toestand en trend
- Belemmeringen door **ontbrekende waarden**
- Vraag: **kan imputeren oplossing bieden?**

Bevindingen voorstudies

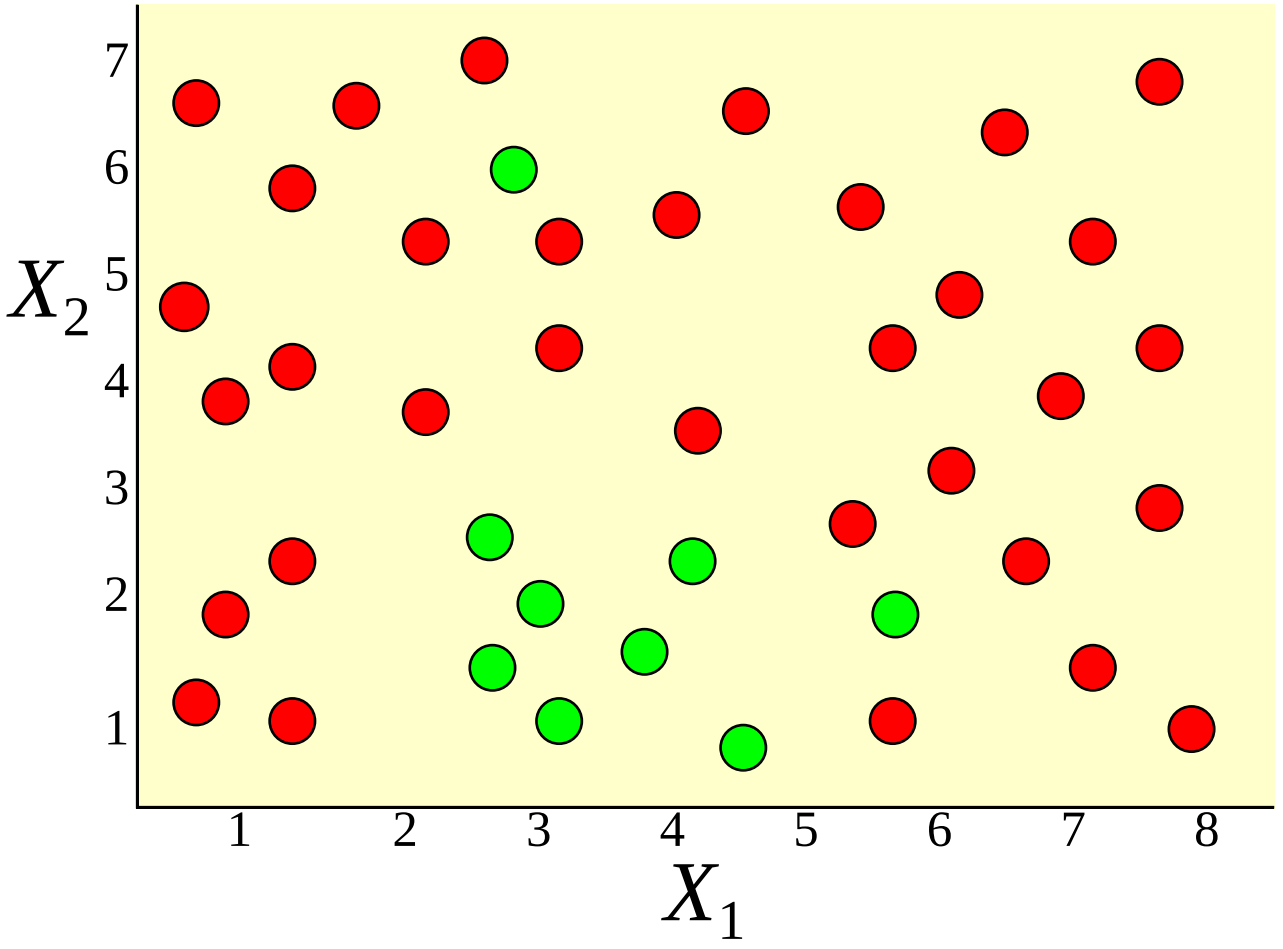
- Gebruik statistische relaties met andere parameters en/of zelfde parameter op andere locaties
- Beoordeel deze met Spearman-rangcorrelatie-coëfficiënt
- Gebruik statistische relaties over korte, recente periode

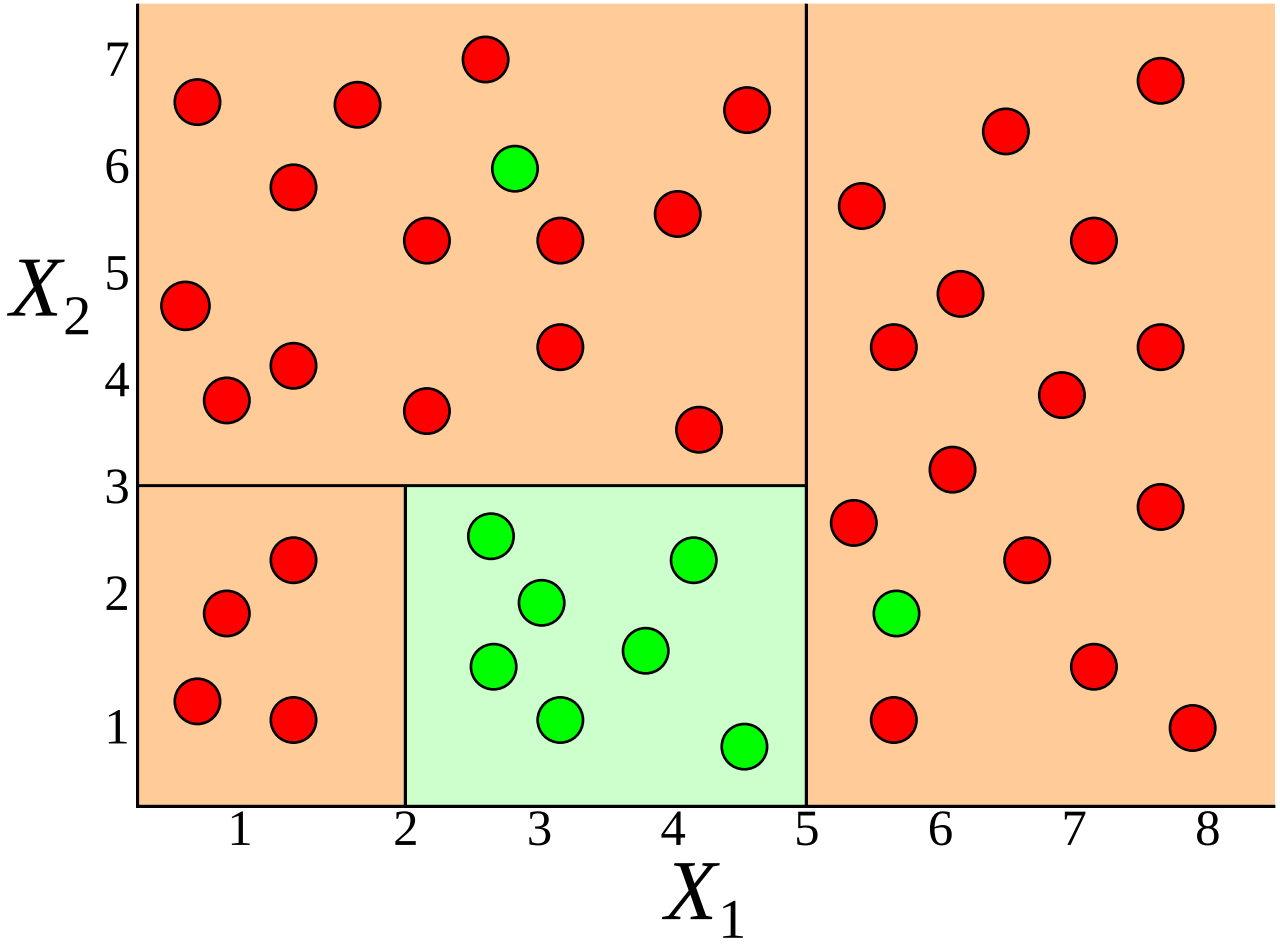
Imputeermethoden

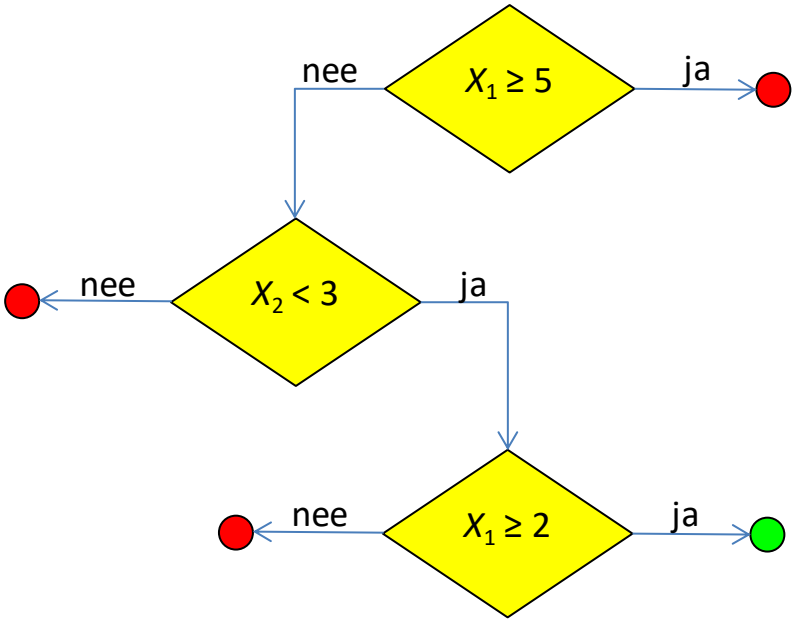
- Meervoudig (MICE, Amelia)
- Enkelvoudig
  - **Lineaire regressie**
  - **Neuraal Netwerk**
  - **Random Forest**

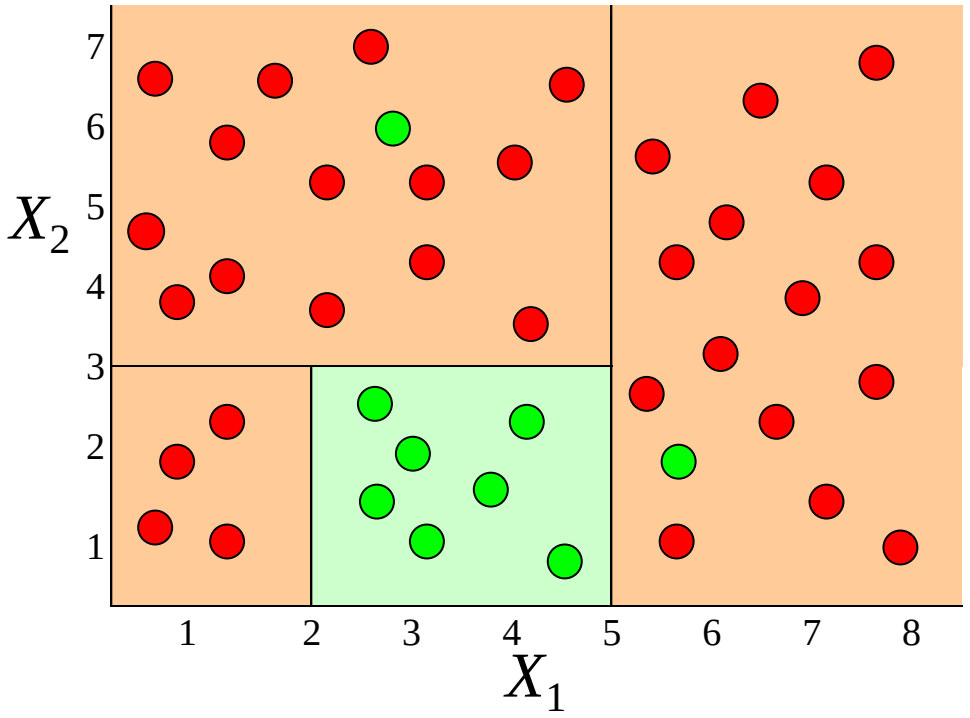
→ Machinaal Leren

X_1	X_2	Y
6,1	3,1	**A**
3,2	5,4	**A**
2,9	1,9	**B**
1,6	6,6	**A**
5,5	1,8	**B**
2,5	2,6	**B**
7,9	1,0	**A**
...	...	...




7
6
5
4
3
2
1
X_2
1
2
3
4
5
6
7
8
X_1





Per knooppunt: Vind welke waarde van welke predictor de set in de twee meest homogene delen splitst

Van zwakke naar sterke voorspeller

Leo Breiman [1993 en 2001]

- combineer veel beslisbomen
- voeg stochastiek toe

=> Random Forest

Toegevoegde stochastiek

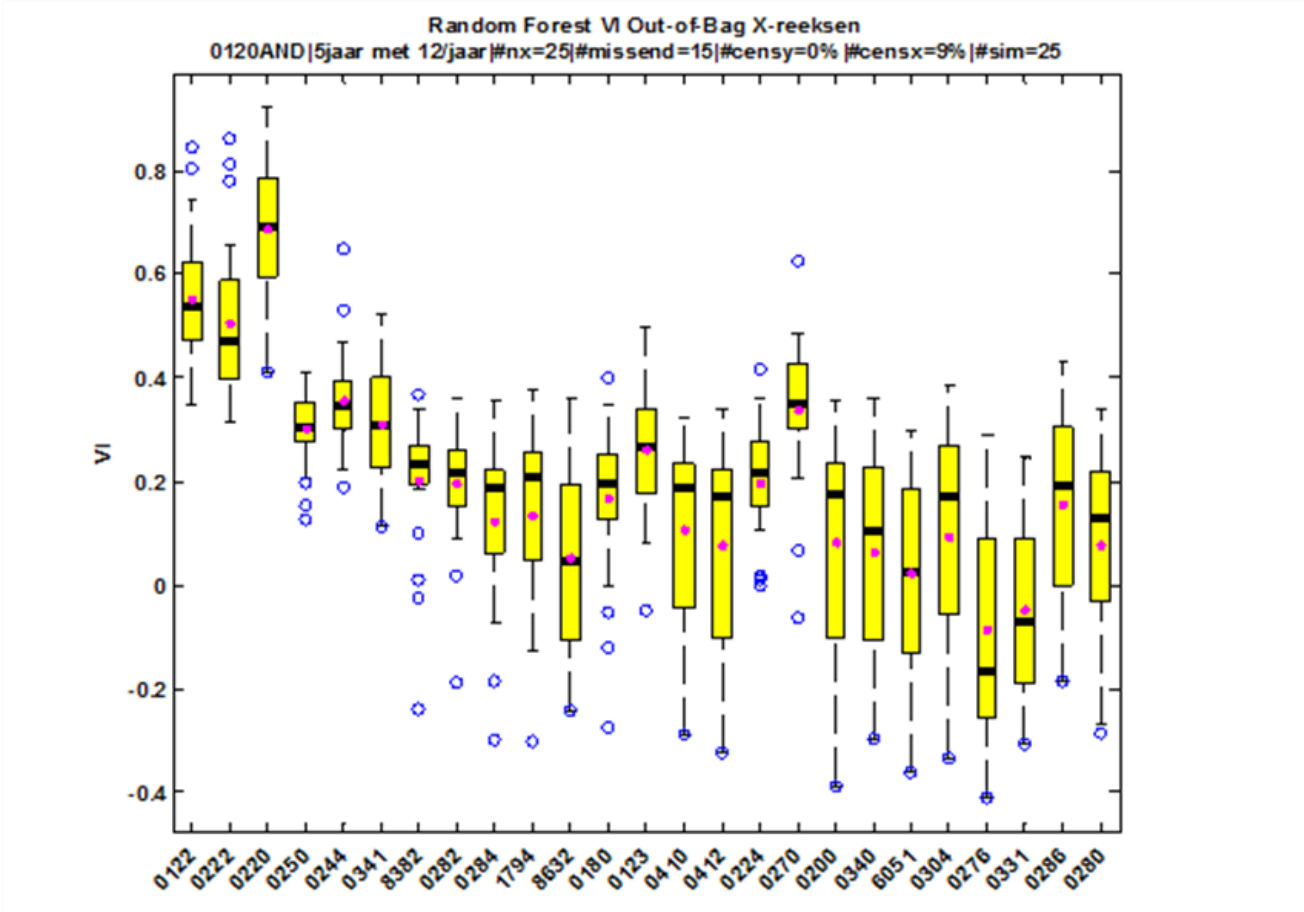
1. Stel elke beslisboom af met trainingsset die even groot is als dataset, maar getrokken mét teruglegging -> ongeveer 63% unieke records

2. Selecteer per knooppunt aselect $p/3$ predictoren (p is totaal aantal predictoren)

Nauwkeurigheid

- Per beslisboom wordt 37% van dataset niet gebruikt om boom af te stellen
- Dit is de Out-of-Box-set (OOB-set)
- Wordt gebruikt om nauwkeurigheid Random Forest-voorspellingen te schatten
- Geeft redelijk zuivere schatting van die nauwkeurigheid

Selectie van predictoren



Uit simulatie bleek dat predictorenselectie op basis van de VI tot betere imputaties leidt

Vergelijken imputeermethoden

- Monte Carlosimulaties
- Kunstmatige reeksen en praktijkreeksen, elk vijf jaar lang, maandwaarden (n = 60)
- 25 predictoren per reeks
- Verwijder waarde en imputeer deze, met elk van drie imputeermethoden
- Bepaal imputeerfout per imputeermethode
- Bepaal kenmerken kansverdeling imputeerfout
- Verleen scores aan imputeerprestaties per reeks

Scoringssysteem

Beschouwde kenmerken imputeerfout

- mediaan
- gemiddelde
- RMSE
- onnauwkeurigheid: $\max[|P_{2,5}| ; |P_{97,5}|]$

Vergelijking imputeerprestaties op kunstmatige processen

Proces	RF	LR	NN	RF	LR	NN	RF	LR	NN	RF	LR	NN
	mediaan			gemiddelde			RMSE			onnauwkeurigh		
1	16	30	54	43	37	20	54	45	1	55	45	0
2	9	33	58	43	42	15	65	34	0	61	38	0
3	14	30	55	42	38	20	48	47	5	47	48	5
4	12	32	56	41	39	19	65	35	0	63	37	0
5	15	33	53	41	36	23	56	43	1	57	43	0
6	12	29	59	43	37	20	57	42	1	57	43	0
7	17	29	55	39	37	24	51	46	3	50	47	2
8	13	32	55	48	31	21	57	43	1	55	44	1
9	10	34	56	43	42	16	66	34	0	66	34	1
10	11	33	56	42	44	14	65	35	0	64	35	0
11	12	31	57	42	43	14	65	35	0	65	35	0

RF: Random Forest

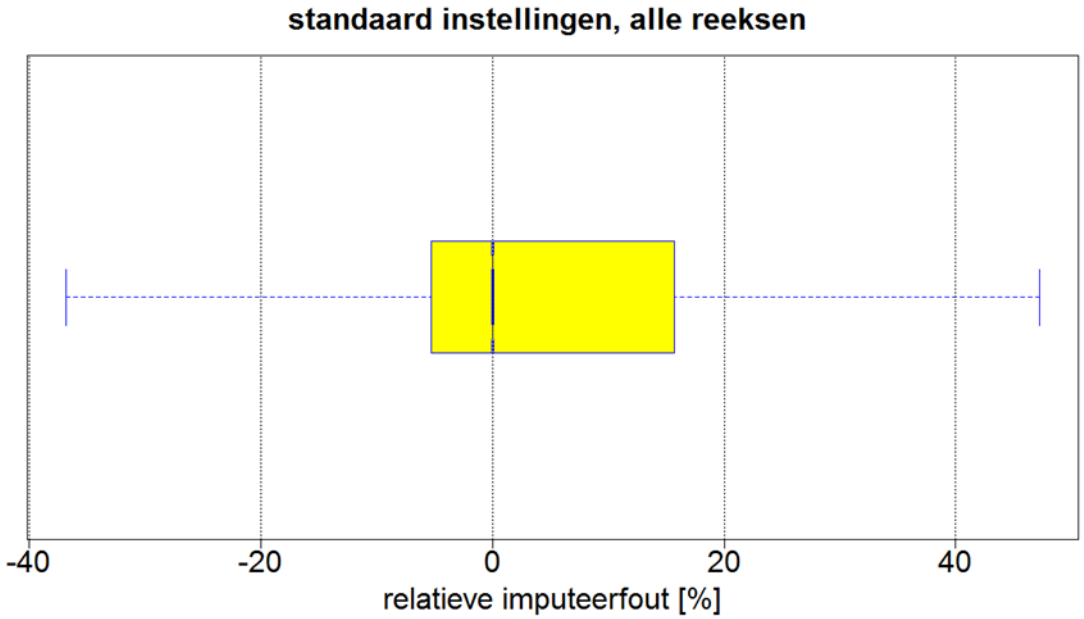
LR: Lineaire regressie

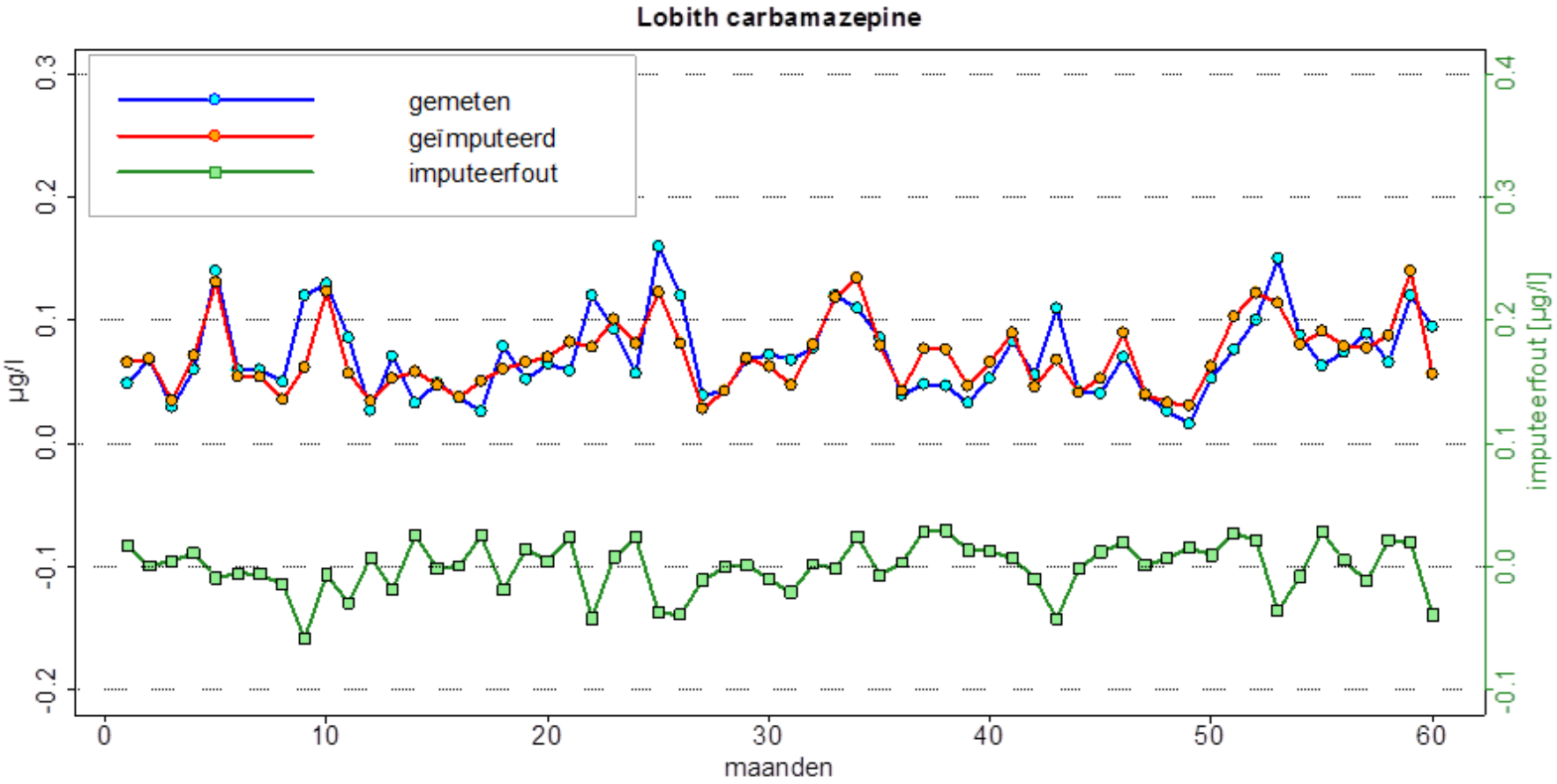
NN: Neuraal netwerk

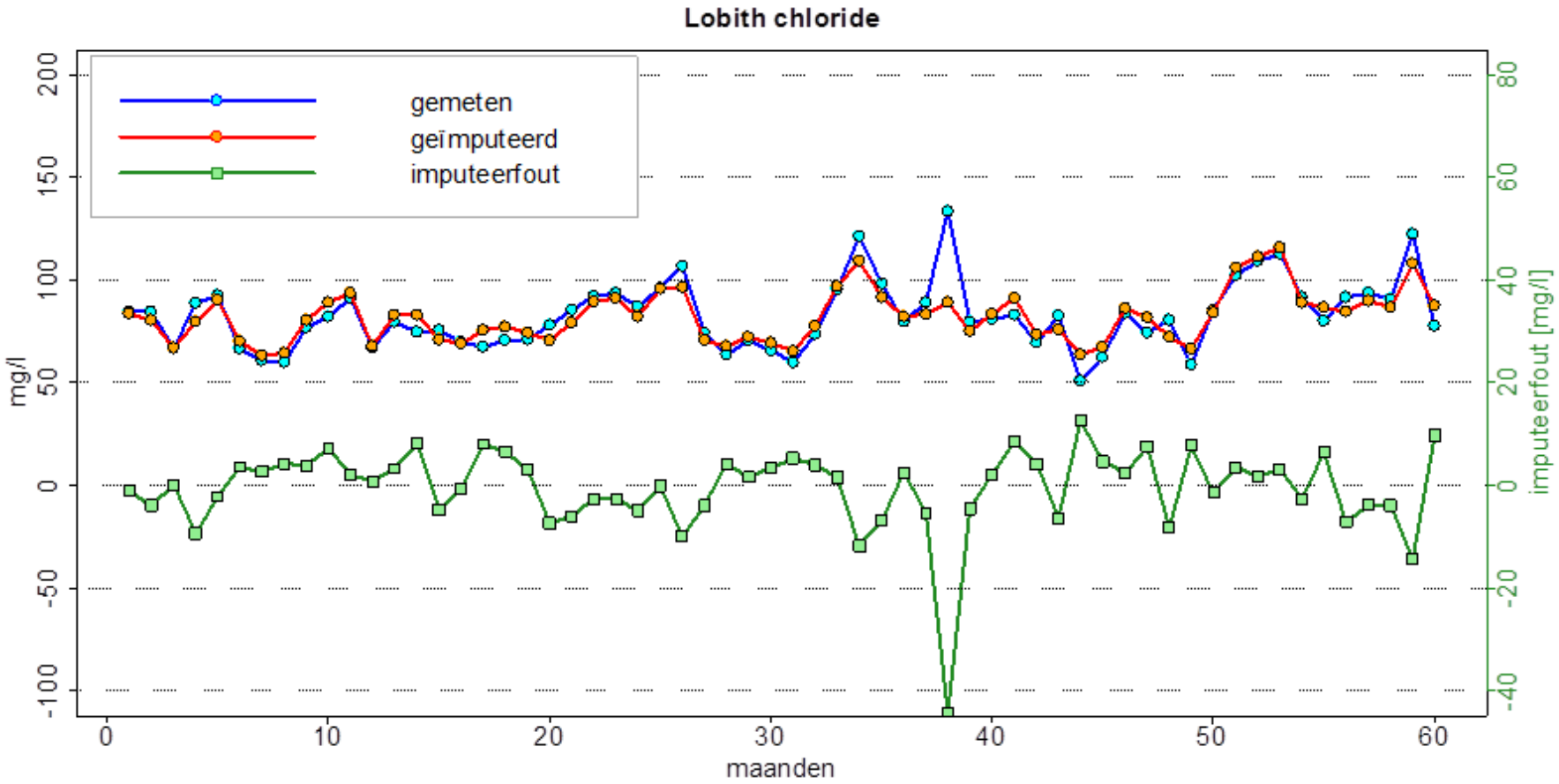
Vergelijking imputeerprestaties op 460 praktijkreeksen (RIWA-base)

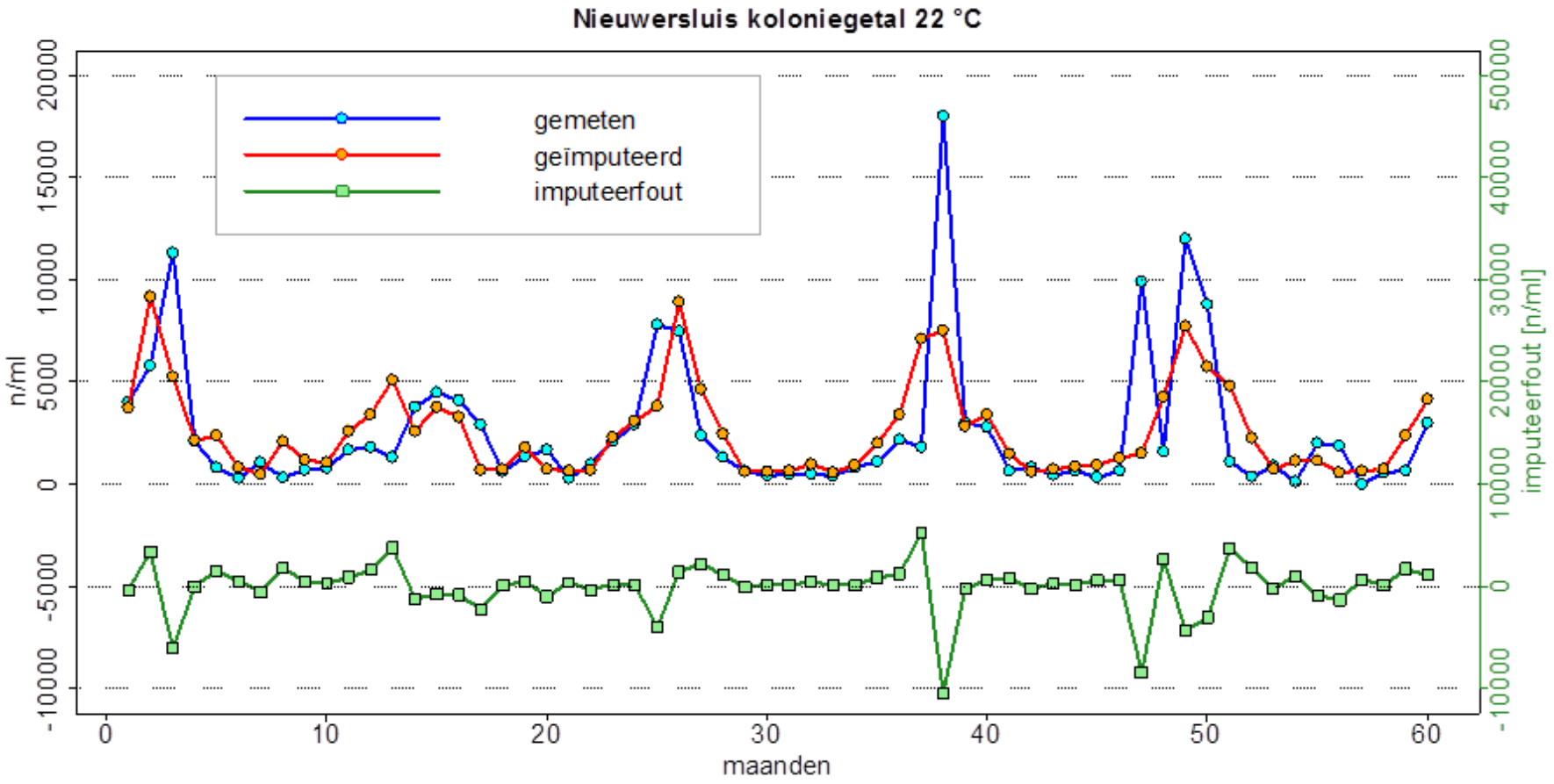
Predictorselectie	gemidd #predictor	RF	LR	NN	RF	LR	NN	RF	LR	NN	RF	LR	NN
		mediaan			gemiddelde			RMSE			onnauwkeurigheid		
25 predictoren	25	22	41	37	48	24	28	58	25	17	50	32	18
P50 VI	13,5	24	38	38	45	29	25	54	34	12	46	39	16
P25 VI	8,5	26	37	37	42	33	25	51	39	10	43	43	14
P10 VI	6,0	26	37	36	43	34	23	50	42	9	43	46	11
P5 VI	5,5	26	36	38	45	34	21	50	42	8	44	44	12

Imputeerprestaties Random Forest op 460 meetreeksen RIWA-base




Lobith carbamazepine
gemeten
geïmputeerd
imputeerfout
µg/l
imputeerfout [µg/l]
maanden
0
10
20
30
40
50
60
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4


Lobith chloride
gemeten
geïmputeerd
imputeerfout
mg/l
imputeerfout [mg/l]
maanden
200
150
100
50
0
-50
-100
80
60
40
20
0
-20
-40
0
10
20
30
40
50
60


Nieuwersluis koloniegetal 22 °C
gemeten
geïmputeerd
imputeerfout
n/ml
imputeerfout [n/ml]
maanden
-10000
-5000
0
5000
10000
15000
20000
10000
20000
30000
40000
50000
0
10
20
30
40
50
60

Conclusies

- Random Forest imputeert doorgaans beter dan lineaire regressie (en is numeriek stabieler)
- Klassieke parameters zijn beter te imputeren (o.a. minder extreme waarden / meetfouten)
- Een andere geschikte toepassing is het beoordelen van reekswaarden

Rapport beschikbaar februari 2014 (riwa.org)

Gebruikte software

- **R**
  - Package missForest
  - Package party
- **Matlab**
  - Treebagger

Vragen?