Data harmonization of enviromental variables: from simple to general solutions

O. Baume, J.O. Skøien, F. Carrée,
G. Heuvelink, E. Pebesma

ENVIRONMENTAL SCIENCES GROUP
WAGENINGEN UR

Context

- Historical Databases at a regional or national level:
  - Standards, devices, data processes are local
- Existence of a common target variable:
  - Air quality, soil quality
  - Gather a database from several measurement campaigns
- Need to merge implies:
  - Biases to appear
  - Definition of a common reference

Background

- Harmonize data in the context of automatic mapping (INTAMAP FP6 project)
  - Pre-processing step before actual statistical analysis and modelling

- Literature:
  - Fassó *et al., 2007* : developed of spatio-temporal model
  - 2 networks
  - Specific to the problem of air quality

Objective

Objective

- Simple and general method

Objective

- Simple and general method
- Geostatistical formalism

Objective

- Simple and general method
- Geostatistical formalism

Help decision making across boundaries!

Objective

- Simple and general method
- Geostatistical formalism

Help decision making across boundaries!

Illustrations

- Assessment of radioactivity exposure

Objective

- Simple and general method
- Geostatistical formalism

Help decision making across boundaries!

Illustrations

- Assessment of radioactivity exposure
- Harmonization of soil CN ratio

Material and methods

- An univoque function transforms measurements into the same target variable (state variable)
  - For network *i* :

$$Z(\mathbf{S}_i) = F_i^{-1}(Y_i(\mathbf{S}_i))$$

- We propose to use the simplest relationship:
  - One bias per network
  - Only additive biases are envolved

$$Z(\mathbf{S}_i) = Y_i(\mathbf{S}_i) - b_i$$

- Estimation by least-squares

Material and methods

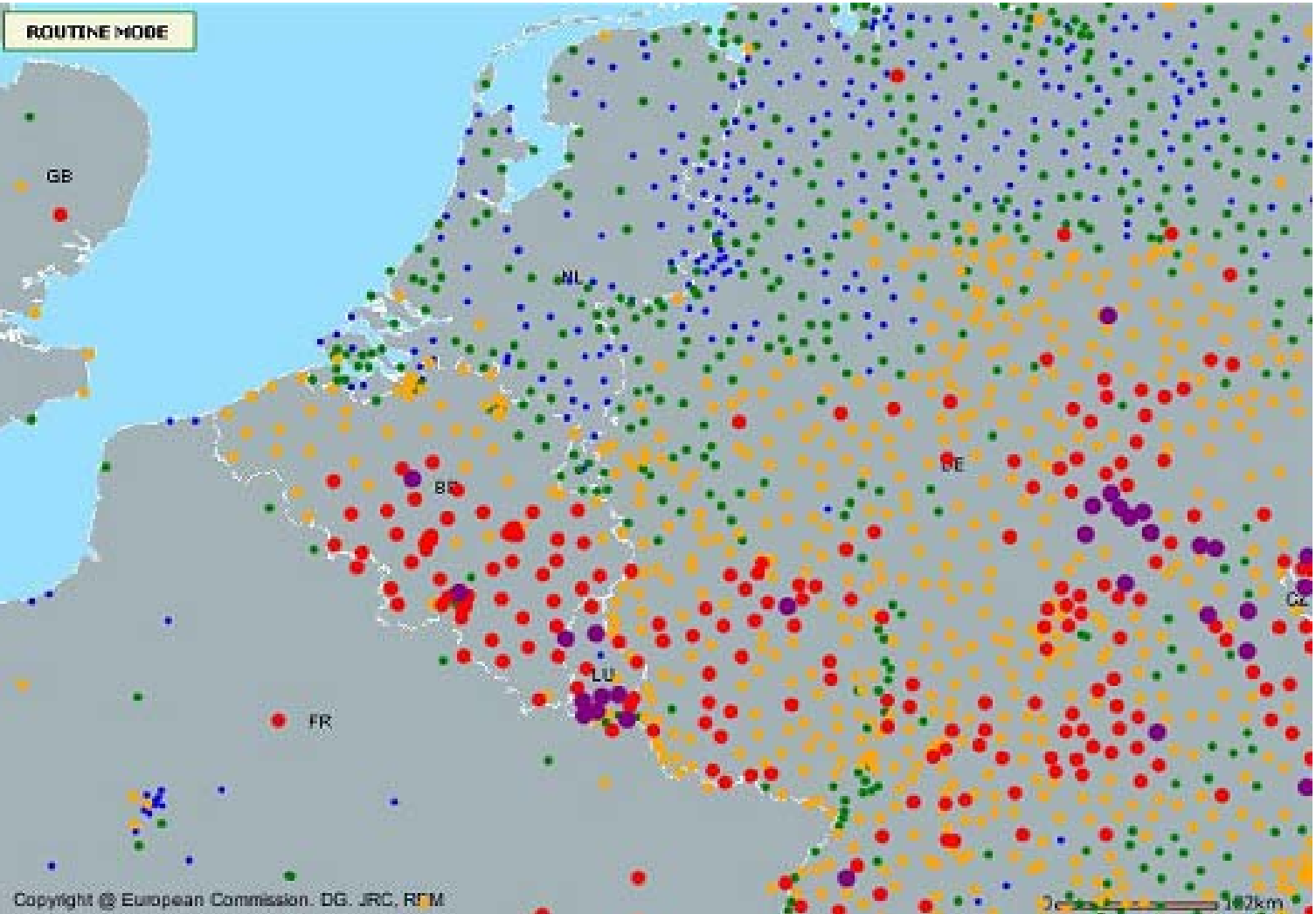
- The target variable $Z(\mathbf{S}_i)$ can also be related to covariates
  - Radioactivity: elevation, soil type
  - Ratio CN of soil cover: pH, tree type

- An extra condition is imposed for a harmonization reference

- We can optionnally include prior information on the biases or on the covariates:
  - Allows to combine several methods of estimation

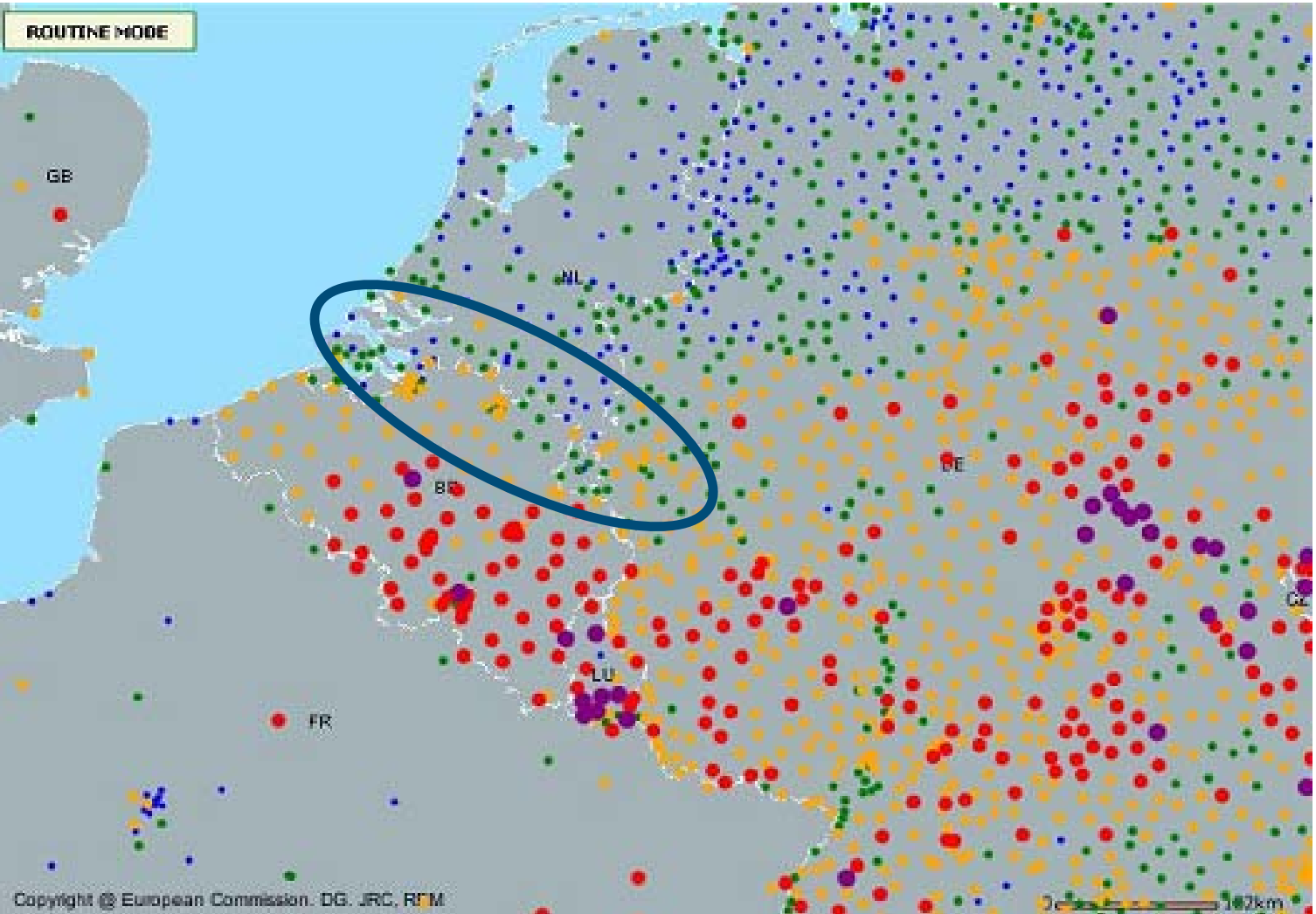
Results: Radioactivity exposure

- EURDEP database: National upload of Gamma dose



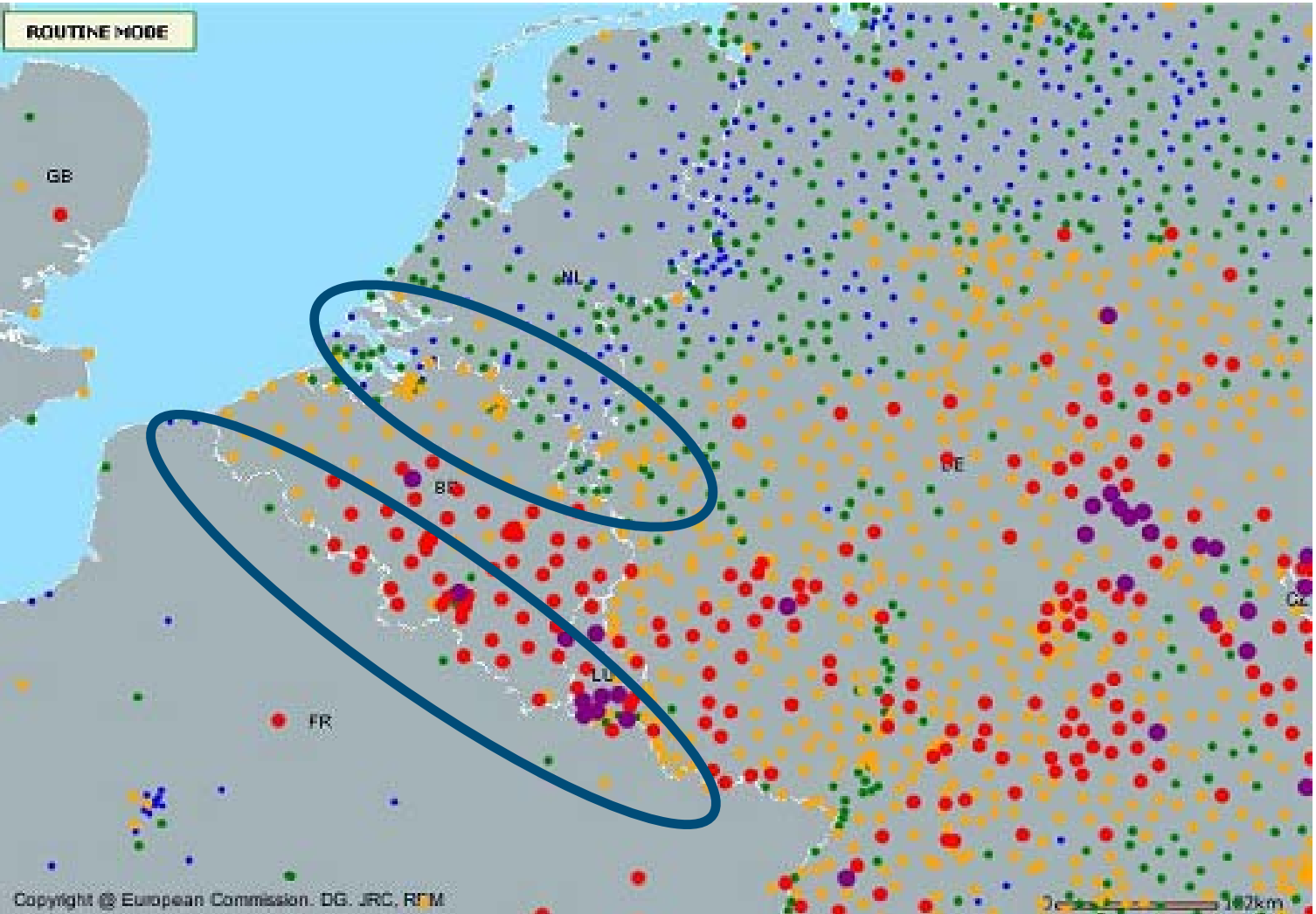
Results: Radioactivity exposure

- EURDEP database: National upload of Gamma dose



Results: Radioactivity exposure

- EURDEP database: National upload of Gamma dose



Results: Radioactivity exposure

- Introduction of prior information:
  - Bias estimates by Skoien *et al.*, 2009

Country	without prior	*Standard deviation*	Prior values	*Standard deviation*	Posterior with prior	*Standard deviation*
Austria (AT)	-18.31	*2.04*	-15.00	*3.24*	-18.29	*1.96*
Belgium (BE)	10.76	*2.66*	14.15	*2.95*	12.25	*1.93*
Switzerland (CH)	10.33	*2.98*	11.23	*2.55*	10.36	*1.93*
Czech Republic (CZ)	-0.63	*2.87*	2.18	*3.95*	0.10	*2.27*
Germany (DE)	-5.88	*1.43*	-3.52	*2.60*	-5.20	*1.28*
Italy (IT)	-2.81	*4.41*	-18.97	*12.14*	-6.78	*3.76*
Luxembourg (LU)	16.02	*4.08*	17.96	*5.49*	16.18	*3.15*
Netherlands (NL)	-9.49	*2.09*	-8.05	*2.85*	-8.62	*1.64*

Results: Radioactivity exposure

- Introduction of prior information:
  - Bias estimates by Skoien *et al.*, 2009

Country	without prior	*Standard deviation*	Prior values	*Standard deviation*	Posterior with prior	*Standard deviation*
Austria (AT)	-18.31	*2.04*	-15.00	*3.24*	-18.29	*1.96*
Belgium (BE)	10.76	*2.66*	14.15	*2.95*	12.25	*1.93*
Switzerland (CH)	10.33	*2.98*	11.23	*2.55*	10.36	*1.93*
Czech Republic (CZ)	-0.63	*2.87*	2.18	*3.95*	0.10	*2.27*
Germany (DE)	-5.88	*1.43*	-3.52	*2.60*	-5.20	*1.28*
Italy (IT)	-2.81	*4.41*	-18.97	*12.14*	-6.78	*3.76*
Luxembourg (LU)	16.02	*4.08*	17.96	*5.49*	16.18	*3.15*
Netherlands (NL)	-9.49	*2.09*	-8.05	*2.85*	-8.62	*1.64*

Results: Radioactivity exposure

- Introduction of prior information:
  - Bias estimates by Skoien *et al.*, 2009

Country	without prior	*Standard deviation*	Prior values	*Standard deviation*	Posterior with prior	*Standard deviation*
Austria (AT)	-18.31	*2.04*	-15.00	*3.24*	-18.29	*1.96*
Belgium (BE)	10.76	*2.66*	14.15	*2.95*	12.25	*1.93*
Switzerland (CH)	10.33	*2.98*	11.23	*2.55*	10.36	*1.93*
Czech Republic (CZ)	-0.63	*2.87*	2.18	*3.95*	0.10	*2.27*
Germany (DE)	-5.88	*1.43*	-3.52	*2.60*	-5.20	*1.28*
Italy (IT)	-2.81	*4.41*	-18.97	*12.14*	-6.78	*3.76*
Luxembourg (LU)	16.02	*4.08*	17.96	*5.49*	16.18	*3.15*
Netherlands (NL)	-9.49	*2.09*	-8.05	*2.85*	-8.62	*1.64*

Results: Radioactivity exposure

- Introduction of prior information:
  - Bias estimates by Skoien *et al.*, 2009

Country	without prior	*Standard deviation*	Prior values	*Standard deviation*	Posterior with prior	*Standard deviation*
Austria (AT)	-18.31	*2.04*	-15.00	*3.24*	-18.29	*1.96*
Belgium (BE)	10.76	*2.66*	14.15	*2.95*	12.25	*1.93*
Switzerland (CH)	10.33	*2.98*	11.23	*2.55*	10.36	*1.93*
Czech Republic (CZ)	-0.63	*2.87*	2.18	*3.95*	0.10	*2.27*
Germany (DE)	-5.88	*1.43*	-3.52	*2.60*	-5.20	*1.28*
Italy (IT)	-2.81	*4.41*	-18.97	*12.14*	-6.78	*3.76*
Luxembourg (LU)	16.02	*4.08*	17.96	*5.49*	16.18	*3.15*
Netherlands (NL)	-9.49	*2.09*	-8.05	*2.85*	-8.62	*1.64*

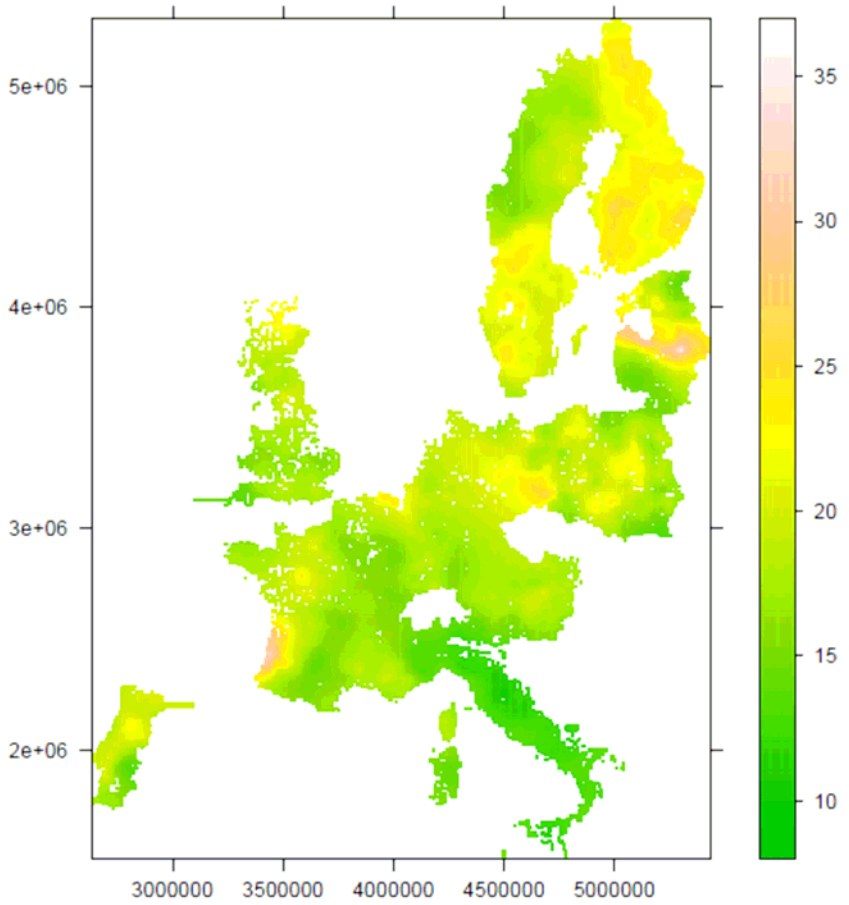
Results: Soil quality

- Biases

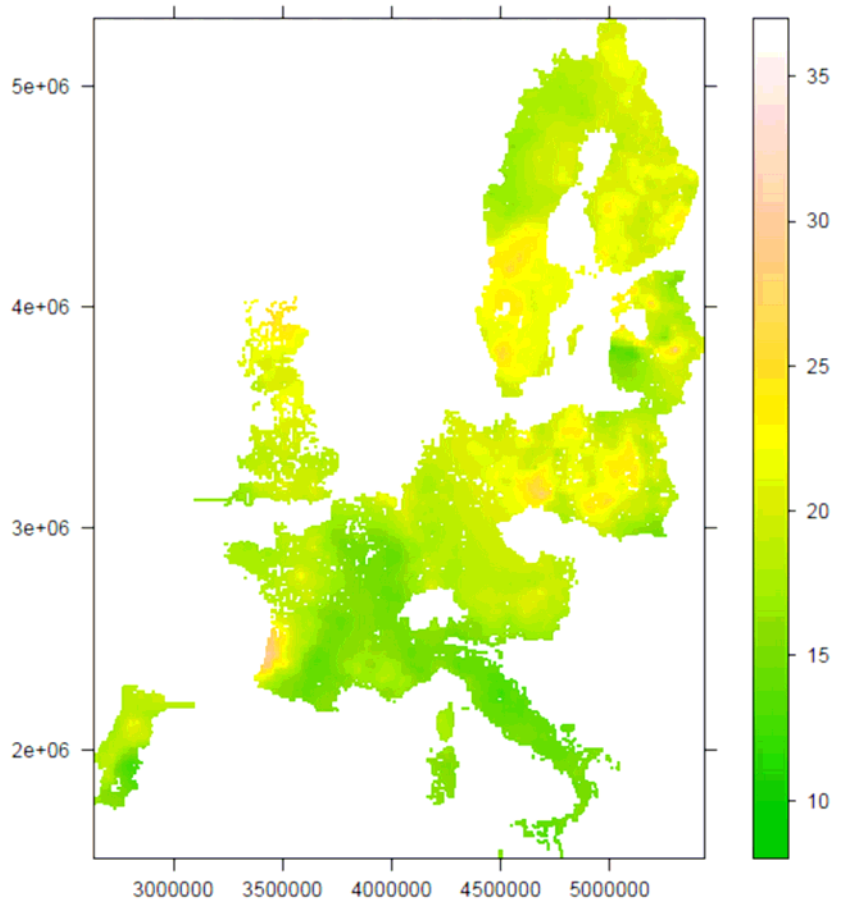
Country (lab. code)	Bias estimate	*Standard deviation*
Austria	-1.2	*0.8*
Belgium	2.6	*2.0*
Estonia	-1.5	*0.9*
Finland	2.4	*0.5*
France	0.9	*0.5*
Germany (401-Lab01)	-2.1	*1.2*
Germany (404-FHE)	0.7	*1.2*
Germany (408-lab_0)	-1.2	*1.6*
Germany (410-lab_g)	0.5	*1.2*
Germany (411-1)	-0.7	*1.6*
Germany (412-Lab_4)	1.5	*2.3*
Italy	-2.0	*0.6*
Latvia	7.5	*1*
Lithuania	-2.8	*1.1*
Poland	-1.7	*0.5*
Portugal	1.3	*0.8*
Slovenia	-0.7	*1.5*
Sweden	-1.0	*0.5*
Great-Britain	-2.4	*0.6*

Results: Soil quality

- Heterogeneous



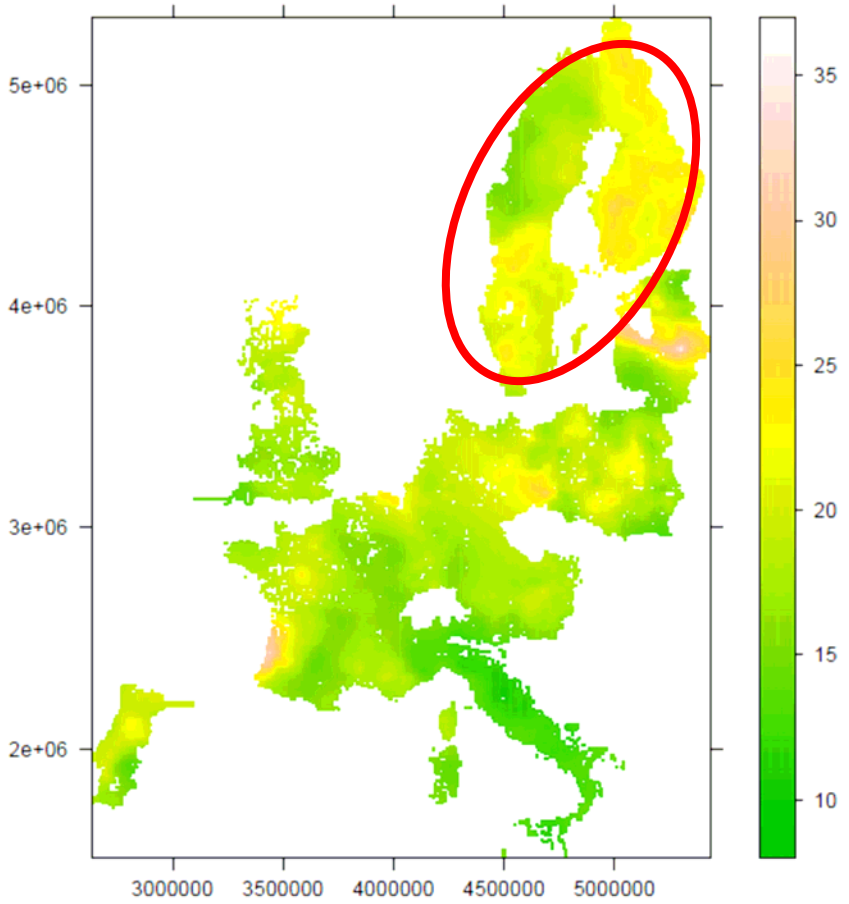
- Harmonized



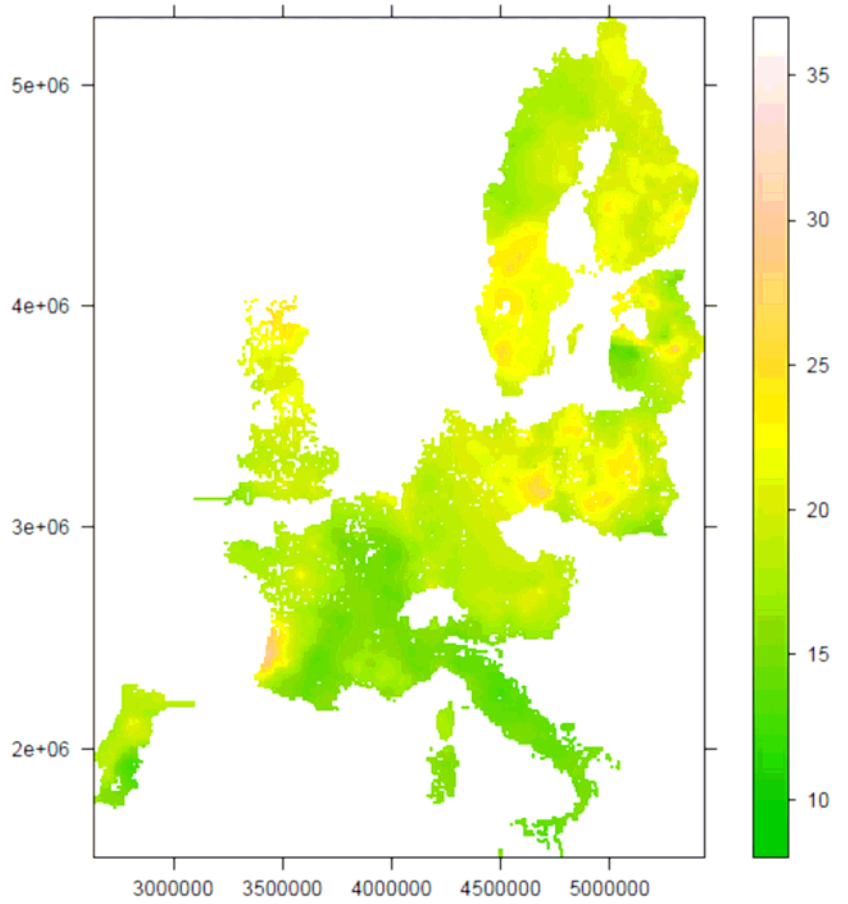
ENVIRONMENTAL SCIENCES GROUP
WAGENINGEN UR

Results: Soil quality

- Heterogeneous

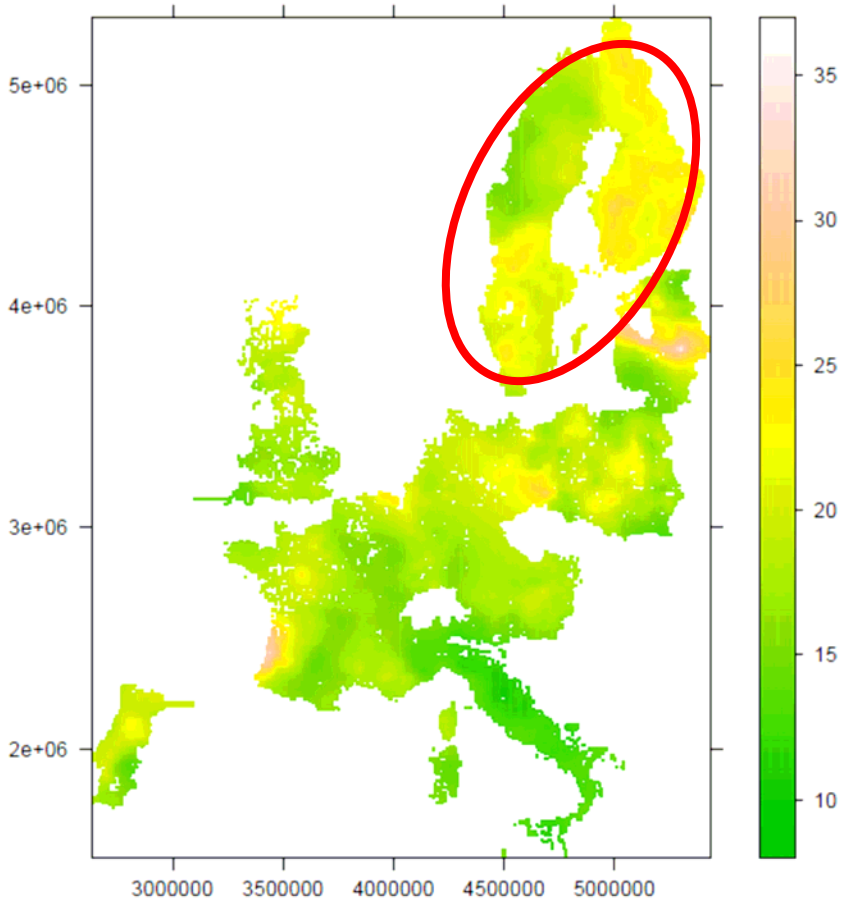


- Harmonized

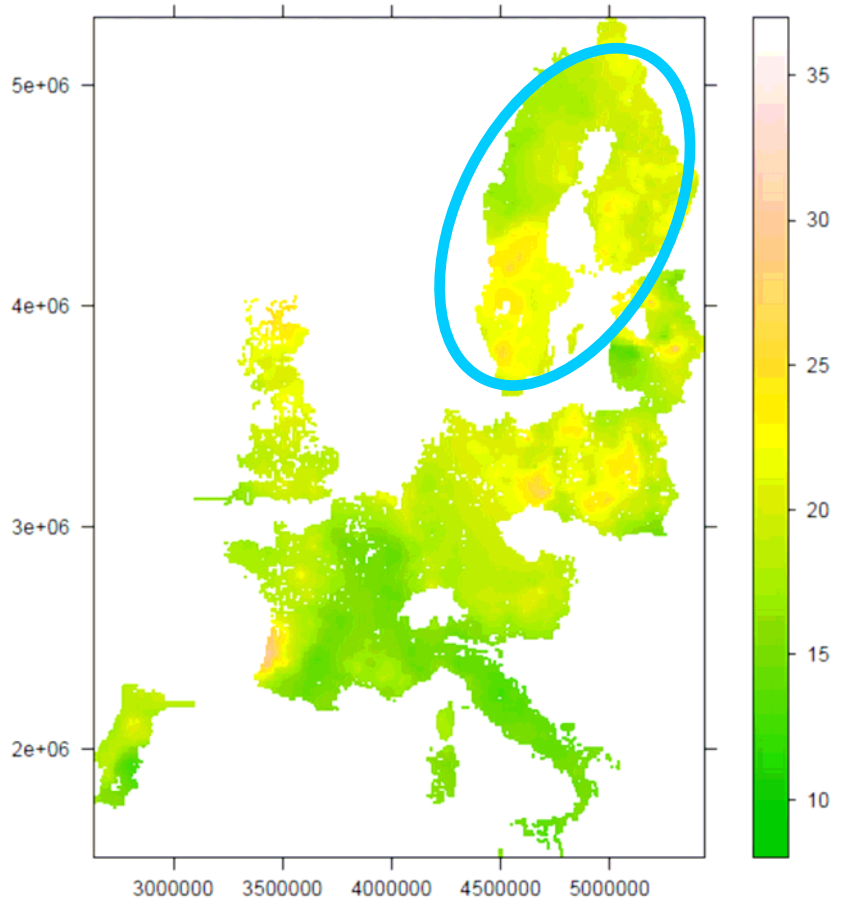


Results: Soil quality

- Heterogeneous



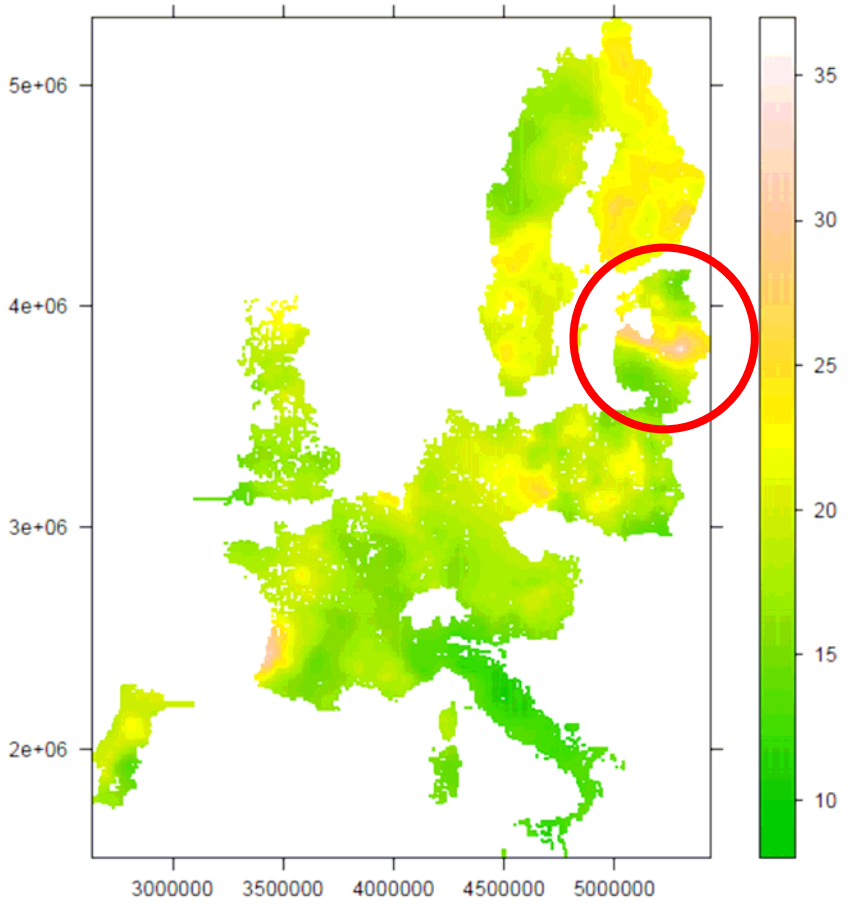
- Harmonized



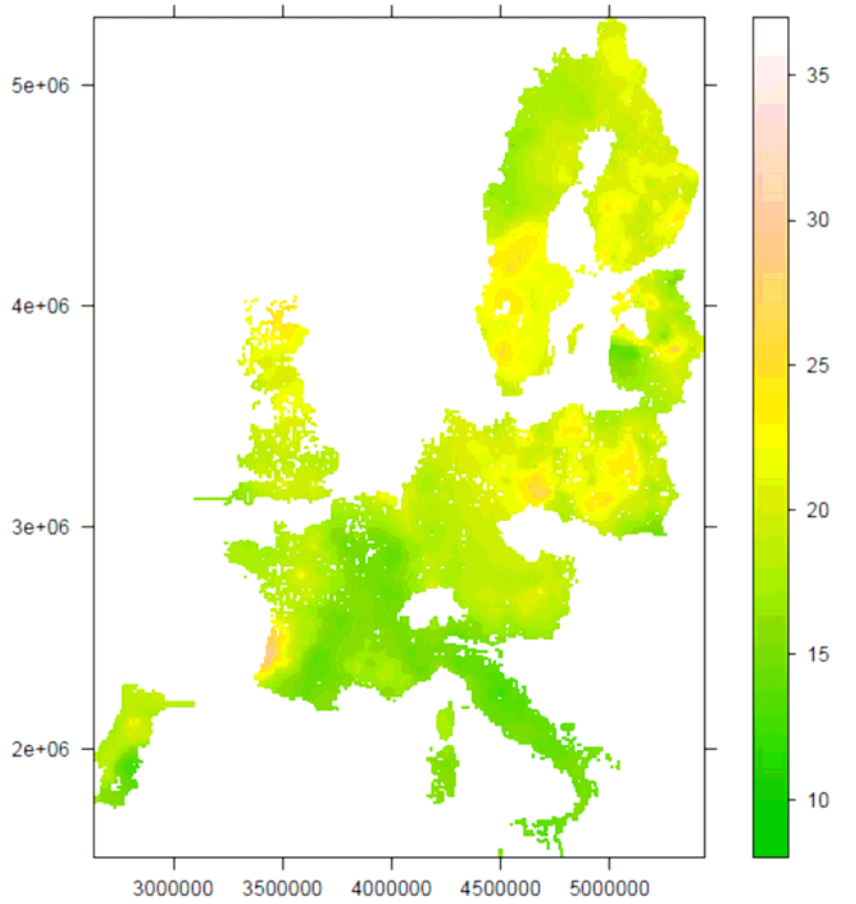
ENVIRONMENTAL SCIENCES GROUP
WAGENINGEN UR

Results: Soil quality

- Heterogeneous



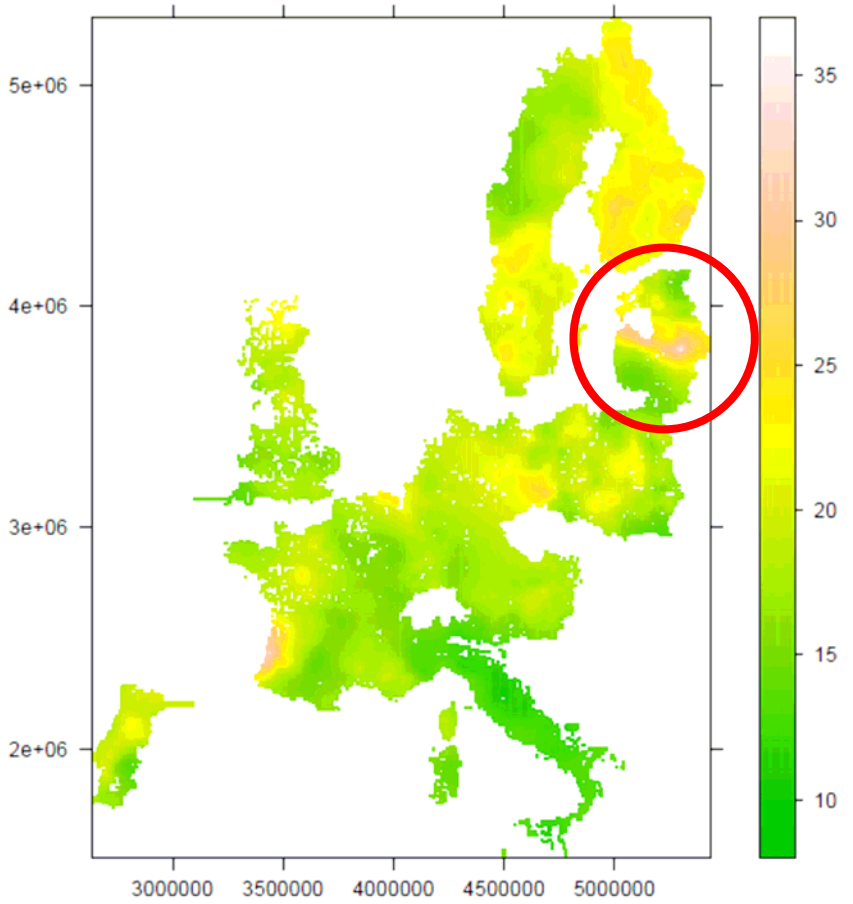
- Harmonized



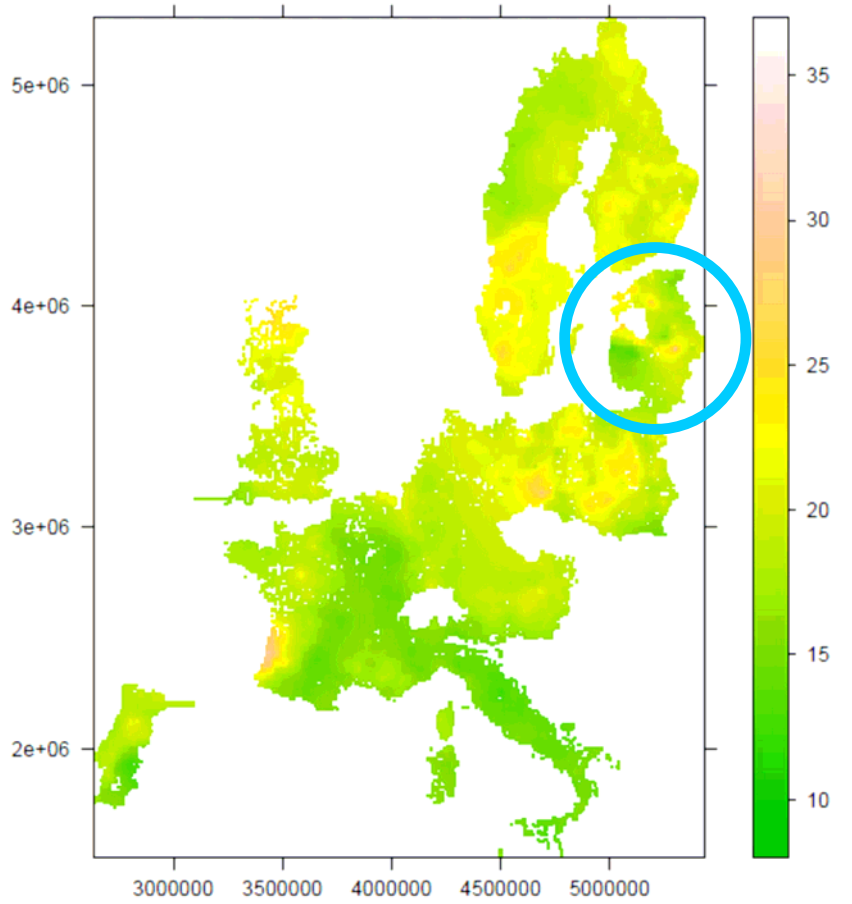
ENVIRONMENTAL SCIENCES GROUP
WAGENINGEN UR

Results: Soil quality

- Heterogeneous



- Harmonized



ENVIRONMENTAL SCIENCES GROUP
WAGENINGEN UR

Discussion

- Accuracy of the estimates of the biases
  - Optimization of the estimation process

Discussion

- Accuracy of the estimates of the biases
  - Optimization of the estimation process
- What about other factors of biases?
  - Need for metadata

Discussion

- Accuracy of the estimates of the biases
  - Optimization of the estimation process
- What about other factors of biases?
  - Need for metadata
- Should more complex relationships be envisaged?
  - Use both additive and multiplicative biases, more general calibration relationships?

Discussion

- Flexibility of Harmonization procedures:
  - Integration of prior information in a Bayesian setting

Discussion

- Flexibility of Harmonization procedures:
  - Integration of prior information in a Bayesian setting
- Simple and general?
  - Our solution works for 2 applications and the whole of European data

Discussion

- Flexibility of Harmonization procedures:
  - Integration of prior information in a Bayesian setting
- Simple and general?
  - Our solution works for 2 applications and the whole of European data
- Harmonization and decision making?

Data harmonization of enviromental variables: from simple to general solutions

Thank you for your attention.

ENVIRONMENTAL SCIENCES GROUP
WAGENINGEN UR