Development of Europe-Wide Models for Particle Elemental Composition Using Supervised Linear Regression and Random Forest

Supporting Information

Jie Chen[a]*, Kees de Hoogh[b,c], John Gulliver[d], Barbara Hoffmann[e], Ole Hertel[f], Matthias Ketzel[f,g], Gudrun Weinmayr[h], Mariska Bauwelinck[i], Aaron van Donkelaar[j,k], Ulla A. Hvidtfeldt[l], Richard Atkinson[m], Nicole AH Janssen[n], Randall V. Martin[k,j,o], Evangelia Samoli[p], Zorana J. Andersen[q], Bente M. Oftedal[r], Massimo Stafoggia[s,t], Tom Bellander[t], Maciej Strak[a,n], Kathrin Wolf[u], Danielle Vienneau[b,c], Bert Brunekreef[a,v], Gerard Hoek[a]

a. Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
b. Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland
c. University of Basel, Petersplatz 1, Postfach 4001, Basel, Switzerland
d. Centre for Environmental Health and Sustainability, School of Geography, Geology and the Environment, University of Leicester, University Road, Leicester, LE1 7RH, UK
e. Institute for Occupational, Social and Environmental Medicine, Centre for Health and Society, Medical Faculty, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany
f. Department of Environmental Science, Aarhus University, P.O. Box 358, Frederiksborgvej 399, 4000 Roskilde, Denmark
g. Global Centre for Clean Air Research (GCARE), Department of Civil and Environmental Engineering, University of Surrey, Guildford, GU2 7XH, UK
h. Institute of Epidemiology and Medical Biometry, Ulm University, Helmholtzstr.22, 89081, Ulm, Germany
i. Interface Demography – Department of Sociology, Vrije Universiteit Brussel, Pleinlaan 2, 1050, Brussels, Belgium.
j. Department of Physics and Atmospheric Science, Dalhousie University, B3H 4R2, Halifax, Nova Scotia, Canada
k. Department of Energy, Environmental & Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri, United States
l. Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark
m. St George's University of London, UK
n. National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, the Netherlands
o. Atomic and Molecular Physics Division, Harvard-Smithsonian Center for Astrophysics, Cambridge, 60 Garden St, Cambridge, MA 02138, US
p. Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias Str, 115 27 Athens, Greece
q. University of Copenhagen, Denmark
r. Department of Environmental Health, Norwegian Institute of Public Health, PO Box 4404 Nydalen N-0403 Oslo, Norway
s. Department of Epidemiology, Lazio Region Health Service / ASL Roma 1, Via Cristoforo Colombo, 112 – 00147, Rome, Italy
t. Institute of Environmental Medicine, Karolinska Institutet, SE-171 77, Stockholm, Sweden
u. Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Institute of Epidemiology, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany
v. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, Netherlands

*Corresponding author: j.chen1@uu.nl

Number of pages: 16
Number of tables: 4
Number of figures: 6

Table S1. Overview of potential predictor variables

Group	Predictor variable	Variable Name	Unit	Buffer size (radius in m) or point estimate	Direction of effect	Restricted to components
Altitude[a]	Altitude	ALT		Point (~90m spacing)	-	All
Road	Road length of major roads in a buffer	MJRD	meter	50, 100, 200, 300, 400, 500, 700, 1000, 2000, 5000, 10000	+	All
	Road length of all roads in a buffer	ALRD				
CTM estimate	Black carbon Aerosol Optical Depth	BCAOD		Point (~13.9km spacing)	+	Cu, Fe, K, Ni, V, Zn
	Sulphate Aerosol Optical Depth	SUAOD				Ni, S, V
	Total column SO_2	TCSO2				Ni, S, V
SAT estimate	Sulfate	SO4	µg/m^3	Point (~1.11km spacing)	+	Ni, S, V
	Organic Matter	OM				Cu, Fe, K, Ni, V, Zn
	Mineral Dust	SOIL				All
	Black Carbon	BC				Cu, Fe, K, Ni, V, Zn
Population	Population density	POP		Point (1km spacing)	+	All
Land use	Percent area of continuous urban fabric - high/ low density	RES	%	50, 100, 200, 300, 400, 500, 600, 700, 800, 1000, 1200, 1500, 1800, 2000, 2500, 3000, 3500, 4000, 5000, 6000, 7000, 8000, 10000	+	All
	Percent area of total build up	TBU			+	
	Percent area of natural lands	NAT			-	
	Percent area of industrial/ commercial lands	IND			+	
	Percent area of ports	POR			+	
	Percent area of urban green	UGR			-	
Industry	Inverse distance weighted sum emission amount of Cu within a buffer	Cu_emi	ton	2000, 4000, 10000	+	Cu
	Inverse distance weighted sum emission amount of Ni within a buffer	Ni_emi				Ni, V
	Inverse distance weighted sum emission amount of PM_{10} within a buffer	PM_{10}_emi				All
	Inverse distance weighted sum emission amount of SO_x within a buffer	SO_x_emi				Ni, S, V

	Inverse distance weighted sum emission amount of Zn within a buffer	Zn_emi				Zn
	Inverse distance weighted number of industrial sites within a buffer	industry	count	2000, 4000, 10000	+	All
	Inverse distance weighted number of industrial sites emitting metal aerosols within a buffer	metal				
	Inverse distance weighted number of industrial sites emitting Cu within a buffer	Cu				
	Inverse distance weighted number of industrial sites emitting Ni within a buffer	Ni				
	Inverse distance weighted number of industrial sites emitting PM_{10} within a buffer	PM_{10}				
	Inverse distance weighted number of industrial sites emitting SO_x within a buffer	SO_x				
	Inverse distance weighted number of industrial sites emitting Zn within a buffer	Zn				
North-south/ east-west gradient[b]	east-west gradient north-south gradient	X_coord Y_coord		Point	+/-	All

CTM = chemical transport model, SAT = satellite-model

[a] Transformed altitude is calculated as √(nalt/max(nalt)), where nalt = altitude − min(altitude).

[b] Transformed X, Y coordinates are calculated as X_coord = (X –Xmin)/(Xmax – Xmin), Y_coord = (Y –Ymin)/(Ymax – Ymin)

Table S2. Performance of $PM_{2.5}$ composition models over Europe

	Inclusion of X, Y coordinates	Component	Cu	Fe	K	Ni	S	Si	V	Zn
		No. of sites	414	413	414	402	404	400	402	413
			Five-fold Hold-Out Validation							
SLR	**One-step**	Regression-based r^2	0.47	0.48	0.58	0.57	0.76	0.50	0.63	0.41
		MSE-based R^2	0.47	0.48	0.58	0.57	0.76	0.50	0.63	0.41
	Two-step, step1	Regression-based r^2	0.44	0.46	0.50	0.51	0.76	0.46	0.60	0.42
		MSE-based R^2	0.44	0.46	0.50	0.51	0.76	0.45	0.59	0.42
	Two-step, step2	Regression-based r^2	0.48	0.48	0.59	0.56	0.79	0.46	0.63	0.41
		MSE-based R^2	0.48	0.48	0.59	0.56	0.79	0.46	0.63	0.41
RF	**One-step**	Regression-based r^2	0.60	0.60	0.82	0.74	0.91	0.62	0.85	0.68
		MSE-based R^2	0.59	0.60	0.82	0.74	0.91	0.62	0.85	0.67
	Two-step, step1	Regression-based r^2	0.59	0.59	0.79	0.74	0.90	0.60	0.84	0.68
		MSE-based R^2	0.58	0.59	0.78	0.74	0.90	0.60	0.86	0.68
	Two-step, step2	Regression-based r^2	0.59	0.61	0.80	0.76	0.90	0.62	0.86	0.71
		MSE-based R^2	0.59	0.61	0.80	0.76	0.90	0.62	0.86	0.71

SLR = Supervised Linear Regression; RF = Random Forest; r^2 = squared Pearson correlation; MSE-based R^2 = Mean Square Error-based R^2

Table S3. Truncation frequency (%) for model predictions at 41,936 random locations

	Inclusion of X, Y coordinates		Cu	Fe	K	Ni	S	Si	V	Zn
SLR	**one-step**	exceed maximum	0	0	0	0	0	0	0	0.1
		negative	38.9	36.3	12	26.2	0	3.9	23.7	18.1
	Two-step, step1	exceed maximum	0	0	0	0	0	0	0	0.1
		negative	53.6	10	0	26.7	0	0	24.3	19.4
	Two-step, step2	exceed maximum	0	0	0	0	0	0	0	0.1
		negative	41.3	10	11.5	21.8	0	0	20.5	19.8
RF	**one-step**	exceed maximum	0	0	0	0	0	0	0	0
		negative	0	0	0	0	0	0	0	0
	Two-step, step1	exceed maximum	0	0	0	0	0	0	0	0
		negative	0	0	0	0	0	0	0	0
	Two-step, step2	exceed maximum	0	0	0	0	0	0	0	0
		negative	0	0	0	0	0	0	0	0

SLR = Supervised Linear Regression; RF = Random Forest

Table S4. Correlation of predictions at monitoring sites between Europe-wide models[a] and ESCAPE models

	Overall r^2	Average of within-area r^2	Range of within-area r^2
Cu	0.63	0.50	(0.02, 0.77)
Fe	0.55	0.47	(0.07, 0.75)
K	0.26	0.10	(0.00, 0.47)
Ni	0.69	0.22	(0.00, 0.78)
S	0.84	0.22	(0.00, 0.73)
Si	0.49	0.27	(0.00, 0.68)
V	0.73	0.27	(0.00, 0.79)
Zn	0.13	0.22	(0.00, 0.74)

SLR = two-step Supervised Linear Regression step2; ESCAPE = area-specific ESCAPE model predictions; r^2 = squared Pearson correlation

[a] We only presented correlations between ESCAPE and SLR predictions, as correlations between ESCAPE and RF predictions cannot be interpreted because RF models have “by design” perfect predictions at training sites.

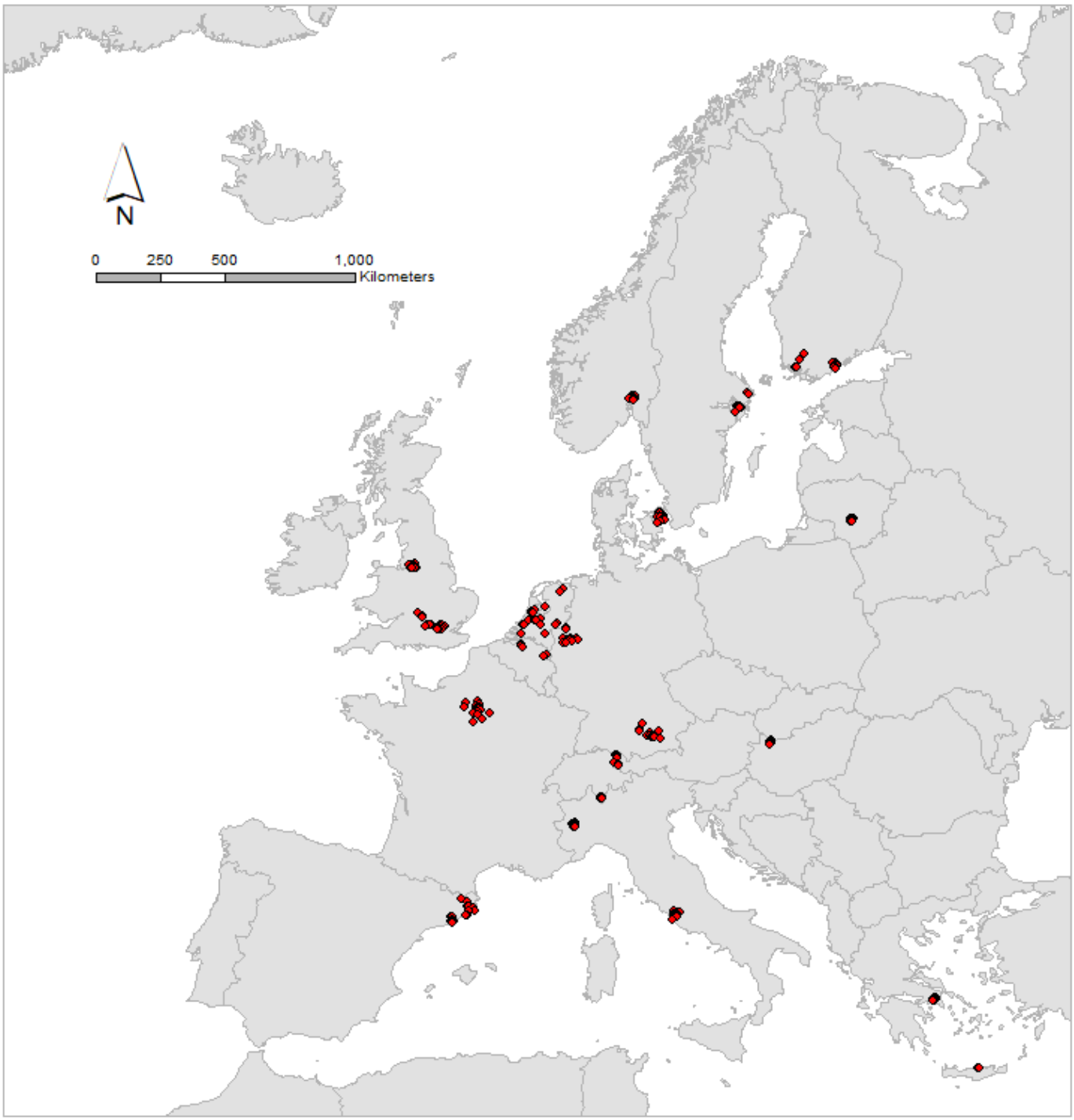


Figure S1. Distribution of 416 ESCAPE monitoring sites across 19 study areas. Each area has 20 sites (40 sites in the Netherlands/Belgium and Catalunya)

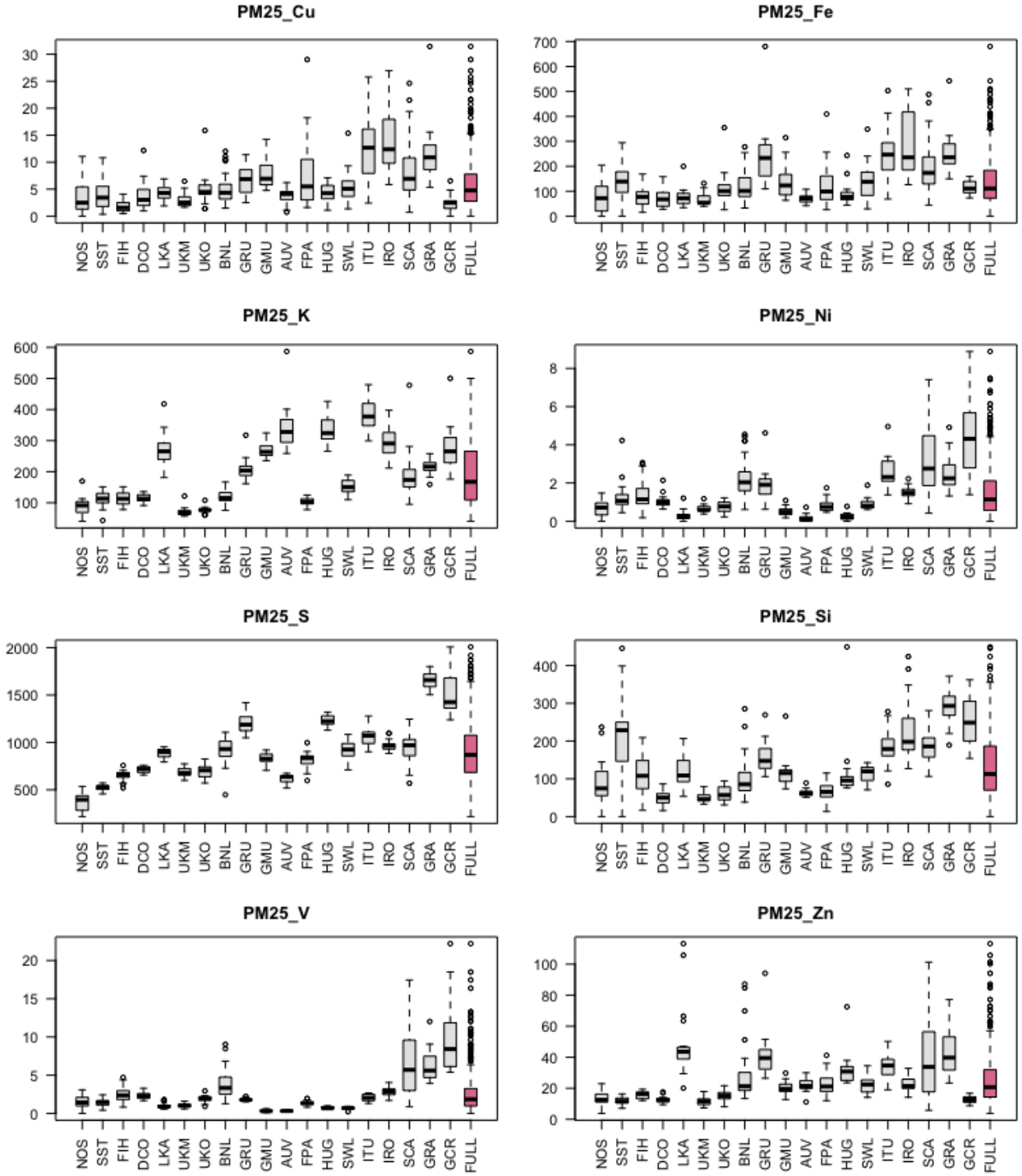


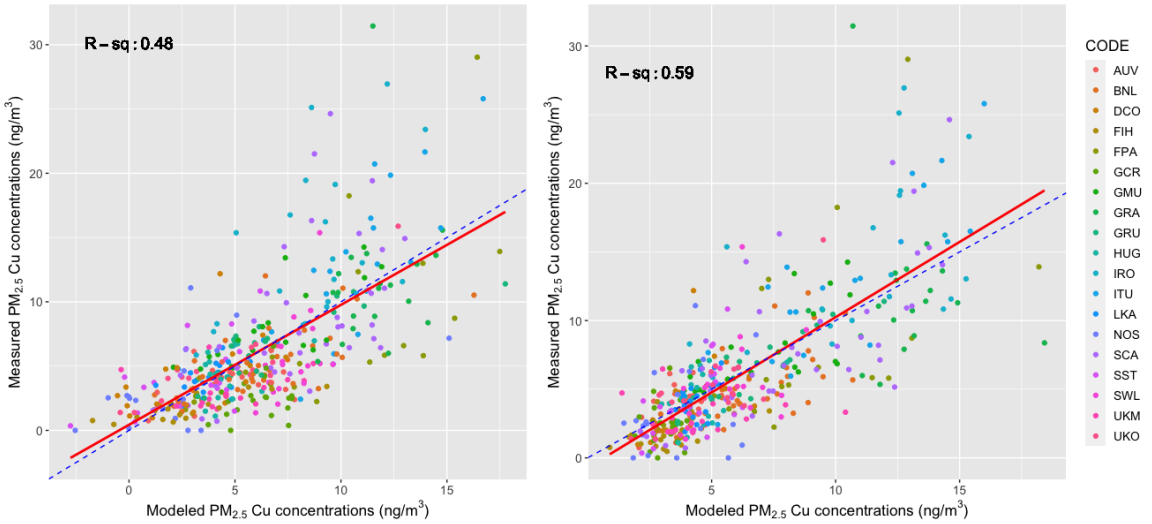
Figure S2. Boxplots of annual mean concentrations for $PM_{2.5}$ composition (ng/m^3) in individual study areas and in the full dataset (box shown in red). Individual study areas are shown from north to south. NOS: Oslo (Norway); SST: Stockholm County (Sweden); FIH: Helsinki/Turku (Finland); DCO: Copenhagen (Denmark); LKA: Kaunas (Lithuania); UKM: Manchester (United Kingdom); UKO: London/Oxford (United Kingdom); BNL: Netherlands/Belgium; GRU: Ruhr Area (Germany); GMU: Munich/Augsburg (Germany); AUV: Vorarlberg (Austria); FPA: Paris (France); HUG: Gyor (Hungary); SWL: Lugano (Switzerland); ITU: Turin (Italy); IRO: Rome (Italy); SCA: Catalunya (Spain); GRA: Athens (Greece); GCR: Heraklion (Greece).

Figure S3. Scatter plots of the stacked predictions at 5 held-out sites versus measurements, obtained from 5-fold hold-out validation analyses

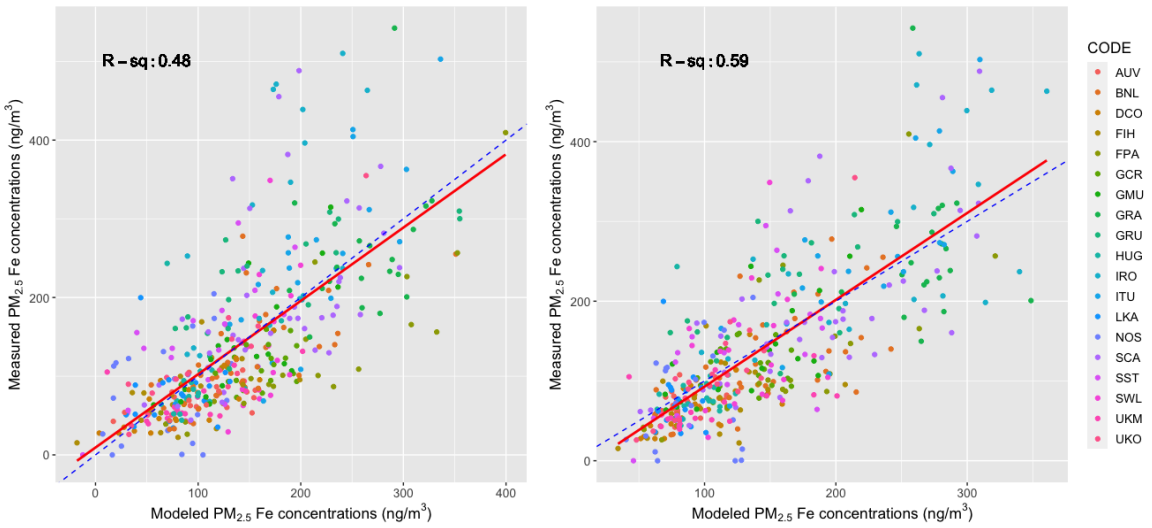
SLR = Supervised Linear Regression; RF = Random Forest

AUV: Vorarlberg (Austria); BNL: Netherlands/Belgium; DCO: Copenhagen (Denmark); FIH: Helsinki/Turku (Finland); FPA: Paris (France); GCR: Heraklion (Greece); GMU: Munich/Augsburg (Germany); GRA: Athens (Greece); GRU: Ruhr Area (Germany); HUG: Gyor (Hungary); IRO: Rome (Italy); ITU: Turin (Italy); LKA: Kaunas (Lithuania); NOS: Oslo (Norway); SCA: Catalunya (Spain); SST: Stockholm County (Sweden); SWL: Lugano (Switzerland); UKM: Manchester (United Kingdom); UKO: London/Oxford (United Kingdom)

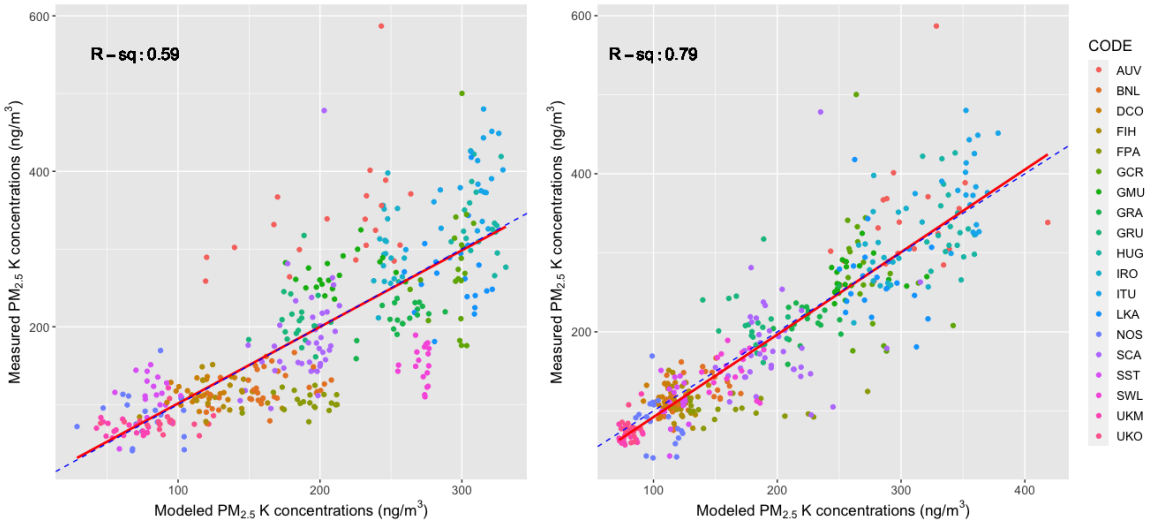
(1) $PM_{2.5}$ Cu (Left two-step SLR step2, Right two-step RF step 1)



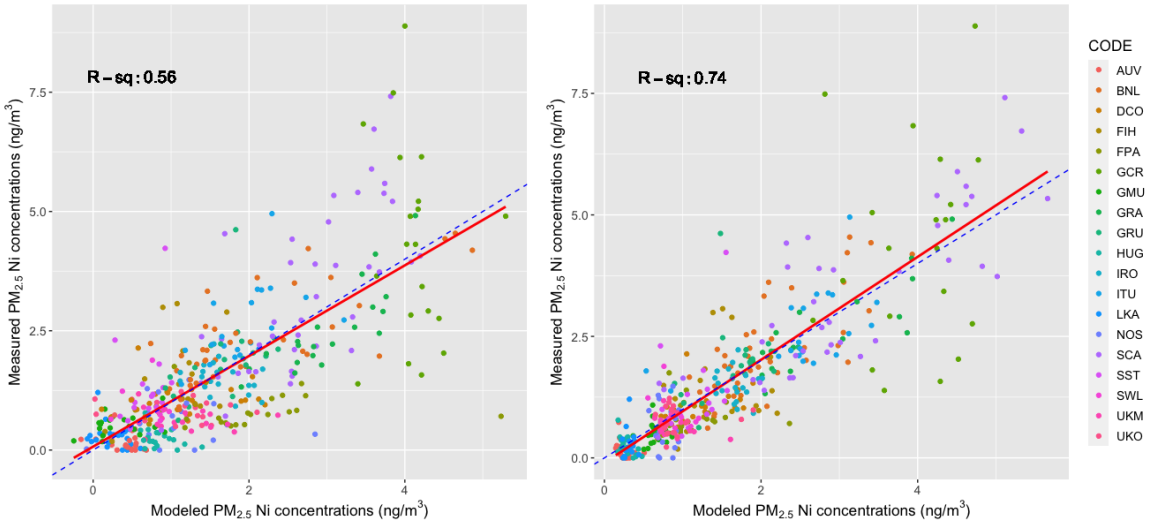
(2) $PM_{2.5}$ Fe (Left two-step SLR step2, Right two-step RF step 1)



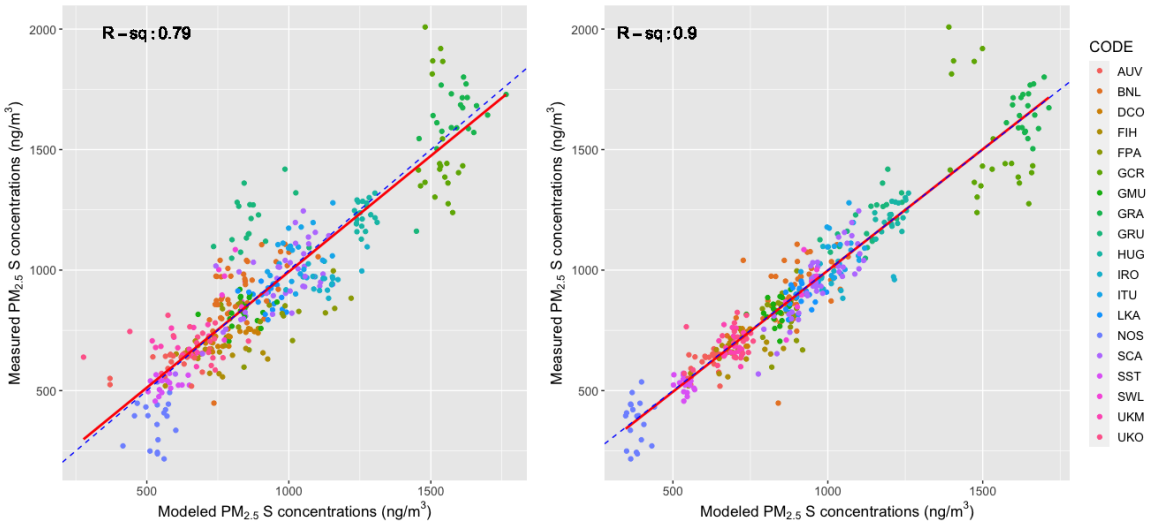
(3) $PM_{2.5}$ K (Left two-step SLR step2, Right two-step RF step 1)



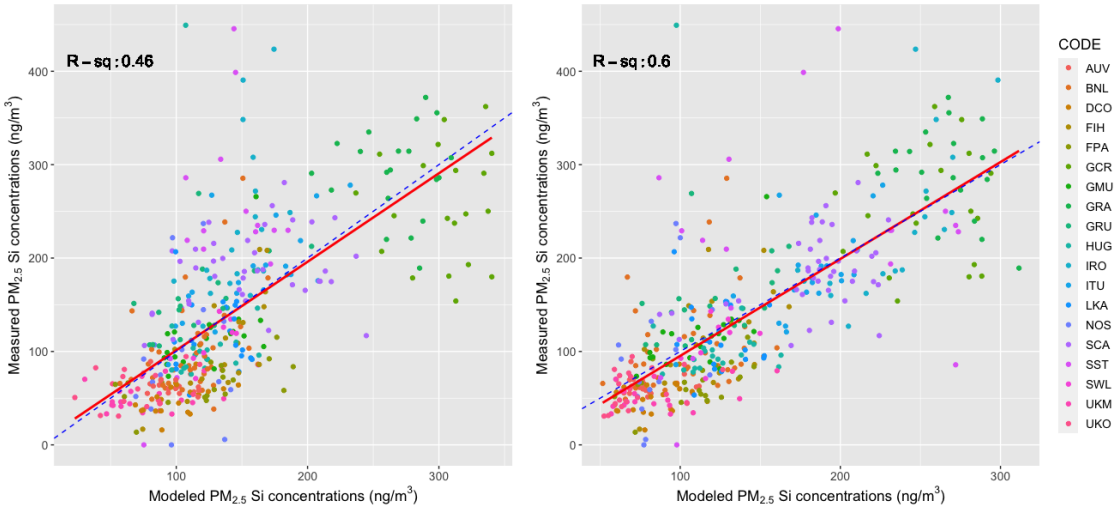
(4) $PM_{2.5}$ Ni (Left two-step SLR step2, Right two-step RF step 1)



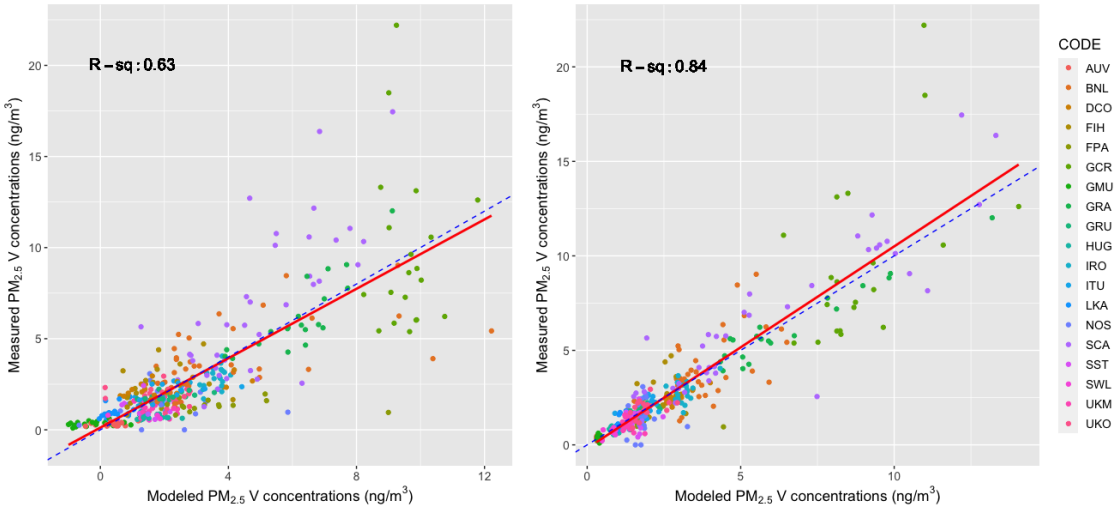
(5) $PM_{2.5}$ S (Left two-step SLR step2, Right two-step RF step 1)



(6) $PM_{2.5}$ Si (Left two-step SLR step2, Right two-step RF step 1)



(7) $PM_{2.5}$ V (Left two-step SLR step2, Right two-step RF step 1)



(8) $PM_{2.5}$ Zn (Left two-step SLR step2, Right two-step RF step 1)

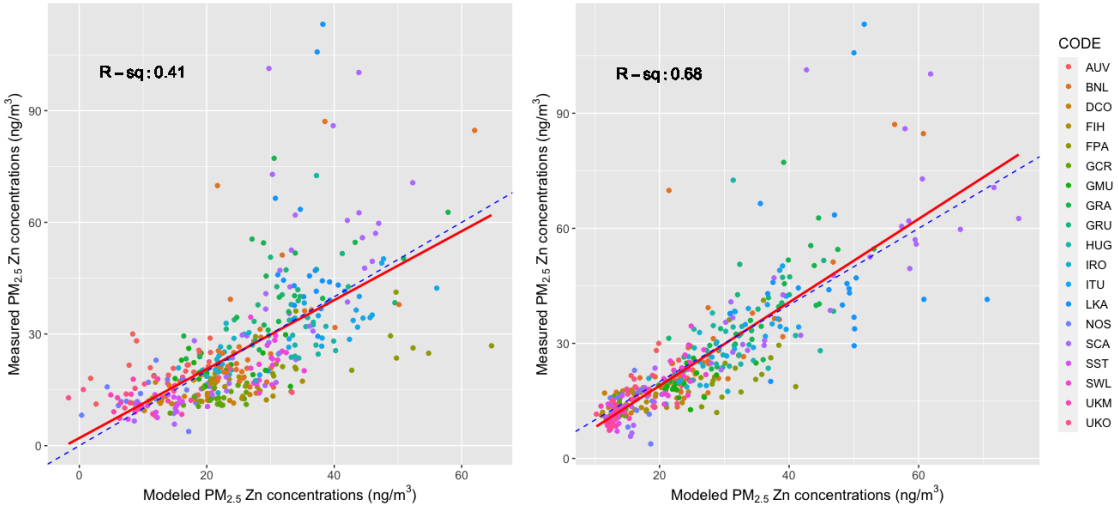
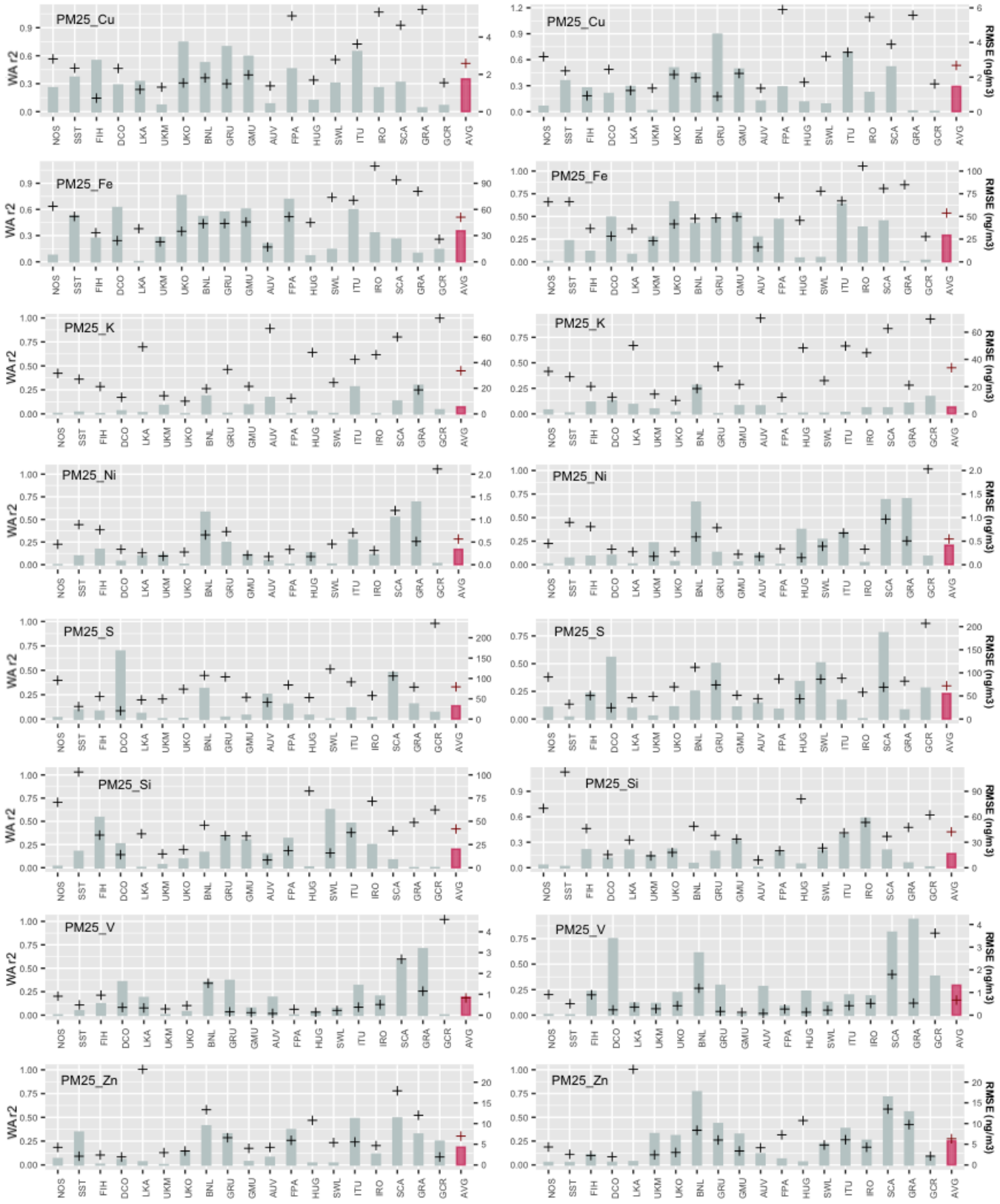


Figure S4. Within-area r^2s (bars and scale on left) and Root-Mean-Square Errors (RMSEs) (plus signs and scale on right) of $PM_{2.5}$ composition models evaluated by five-fold hold-out-validation

(a) **One-step Supervised Linear Regression** **One-step Random Forest**

(b) **Two-step Supervised Linear Regression, step1** **Two-step Random Forest, step1**

(c) Two-step Supervised Linear Regression, step2 **Two-step Random Forest, step2**

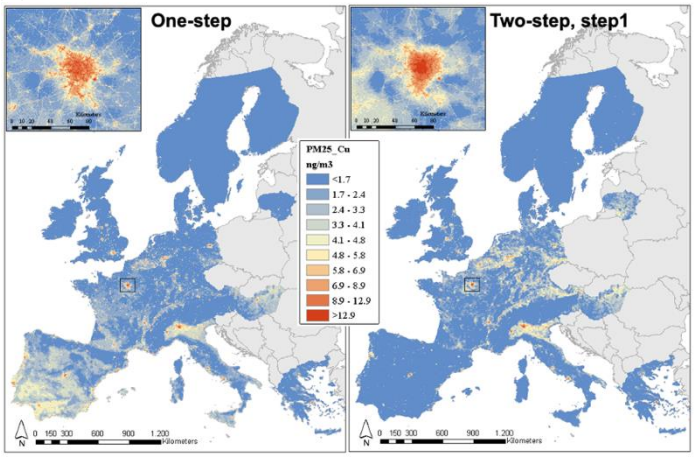


Study areas are shown from north to south. NOS: Oslo (Norway); SST: Stockholm County (Sweden); FIH: Helsinki/Turku (Finland); DCO: Copenhagen (Denmark); LKA: Kaunas (Lithuania); UKM: Manchester (United Kingdom); UKO: London/Oxford (United Kingdom); BNL: Netherlands/Belgium; GRU: Ruhr Area (Germany); GMU: Munich/Augsburg (Germany); AUV: Vorarlberg (Austria); FPA: Paris (France); HUG: Gyor (Hungary); SWL: Lugano (Switzerland); ITU: Turin (Italy); IRO: Rome (Italy); SCA: Catalunya (Spain); GRA: Athens (Greece); GCR: Heraklion (Greece); AVG = average.

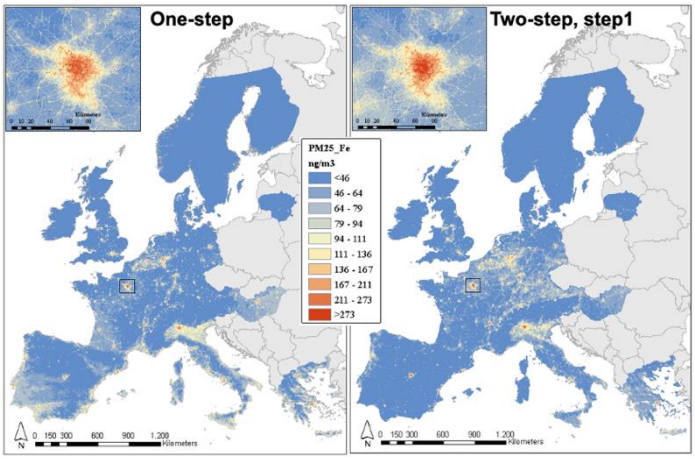
Figure S5. Maps of $PM_{2.5}$ components

(a) Supervised Linear Regression models

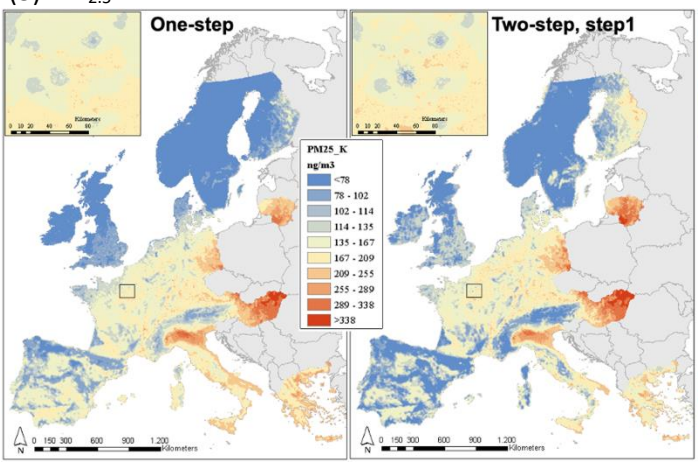
(1) $PM_{2.5}$ Cu



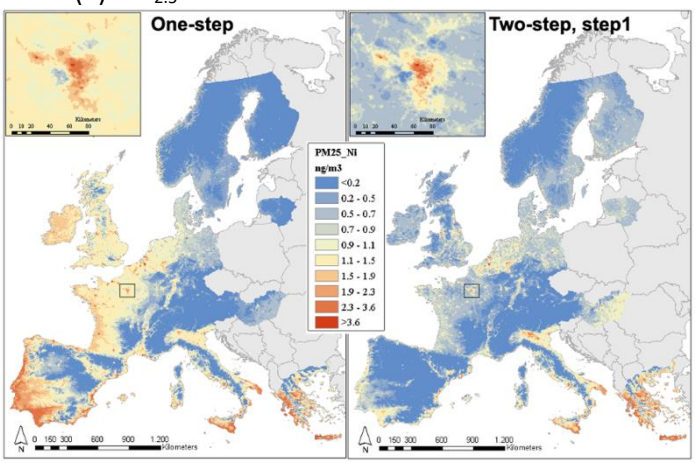
(2) $PM_{2.5}$ Fe



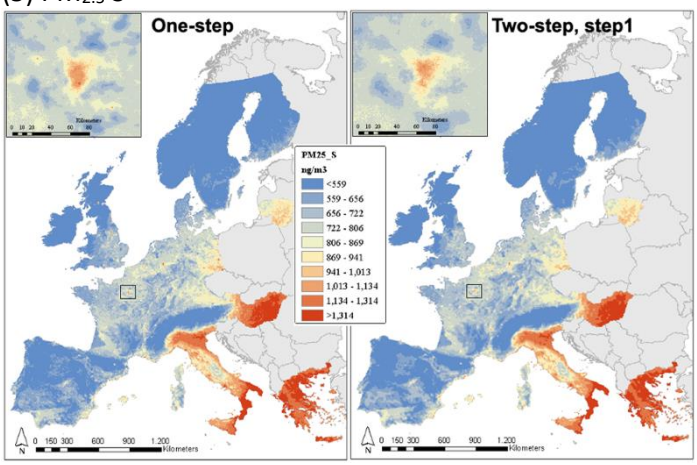
(3) $PM_{2.5}$ K



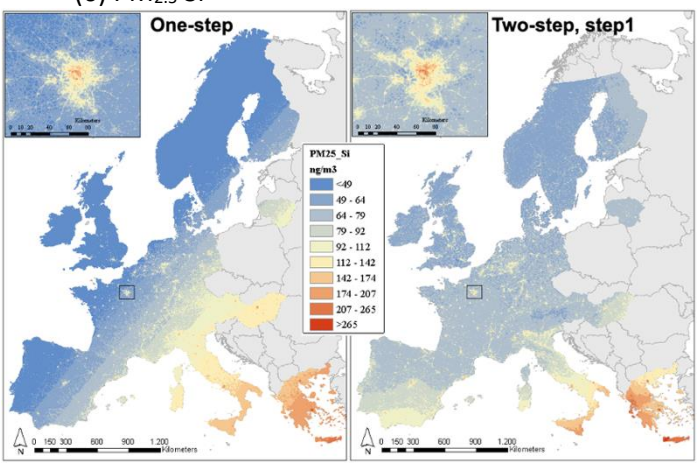
(4) $PM_{2.5}$ Ni



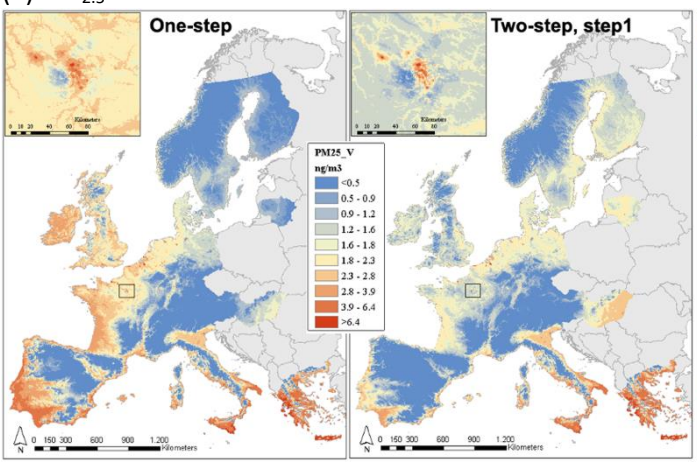
(5) $PM_{2.5}$ S



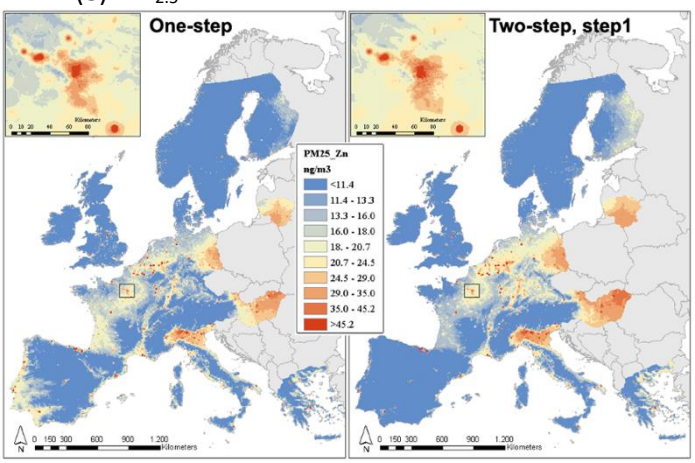
(6) $PM_{2.5}$ Si



(7) $PM_{2.5}$ V

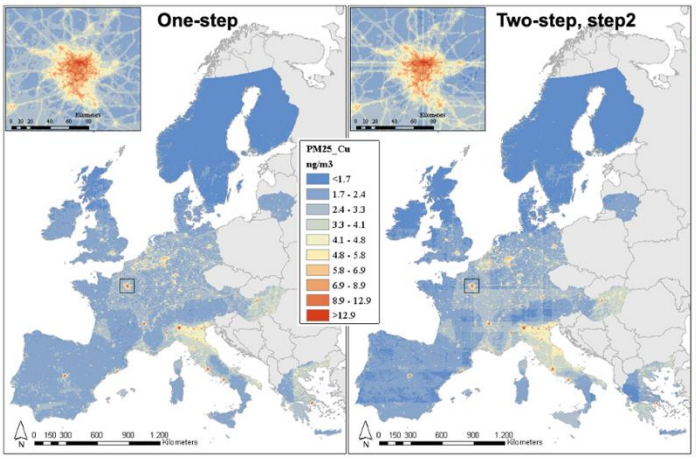


(8) $PM_{2.5}$ Zn

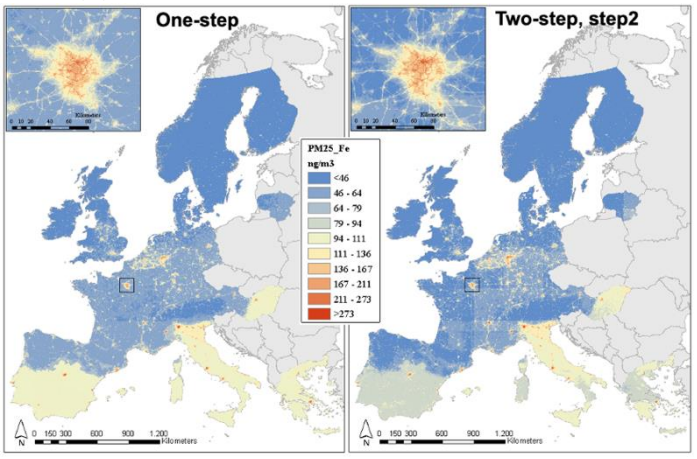


(b) Random Forest models

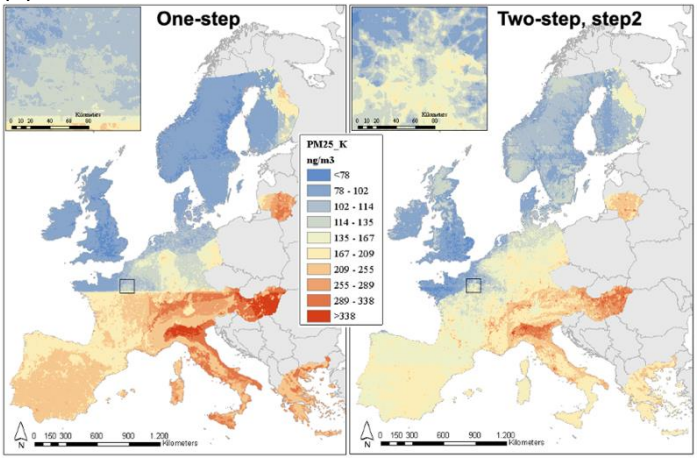
(1) $PM_{2.5}$ Cu



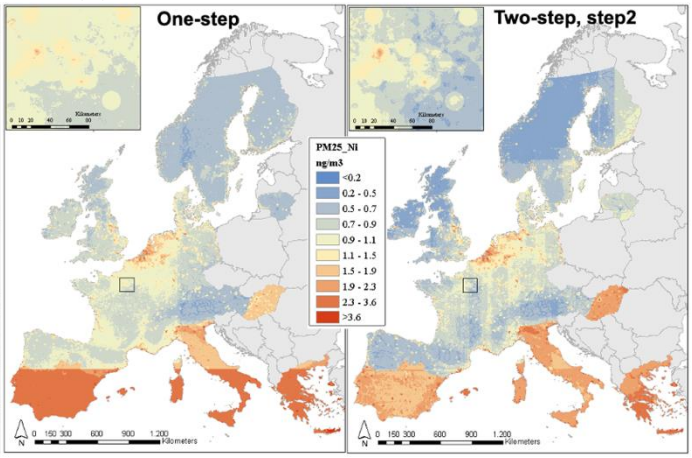
(2) $PM_{2.5}$ Fe



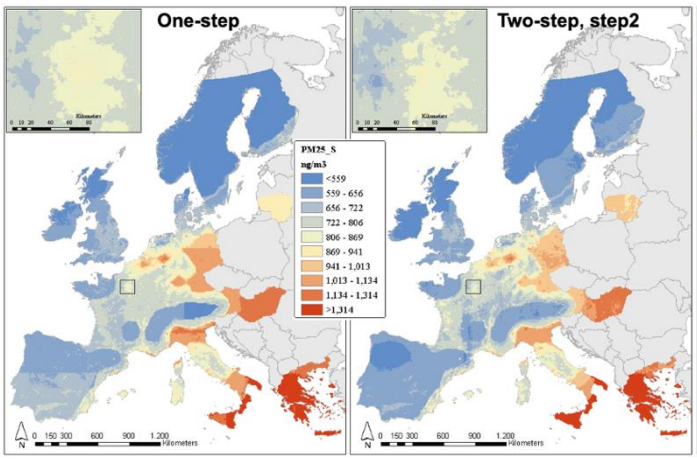
(3) $PM_{2.5}$ K



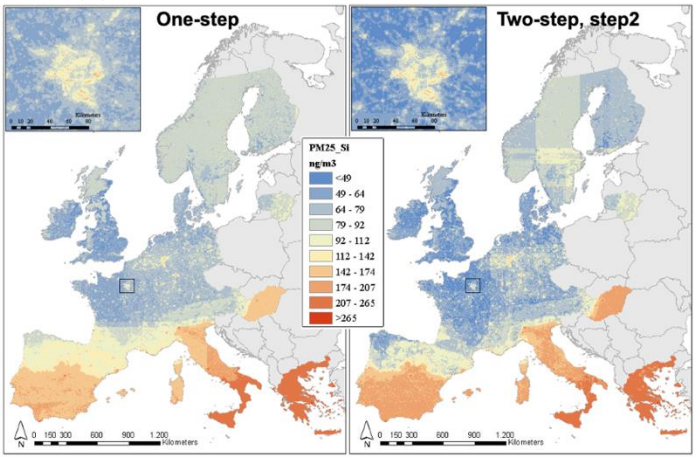
(4) $PM_{2.5}$ Ni



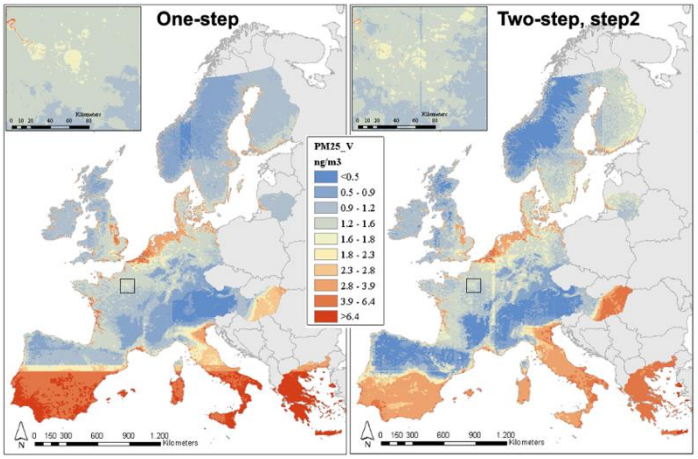
(5) $PM_{2.5}$ S



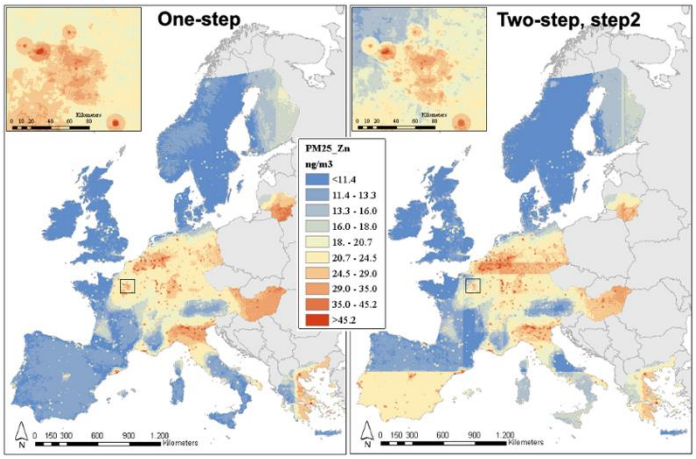
(6) $PM_{2.5}$ Si



(7) $PM_{2.5}$ V



(8) $PM_{2.5}$ Zn



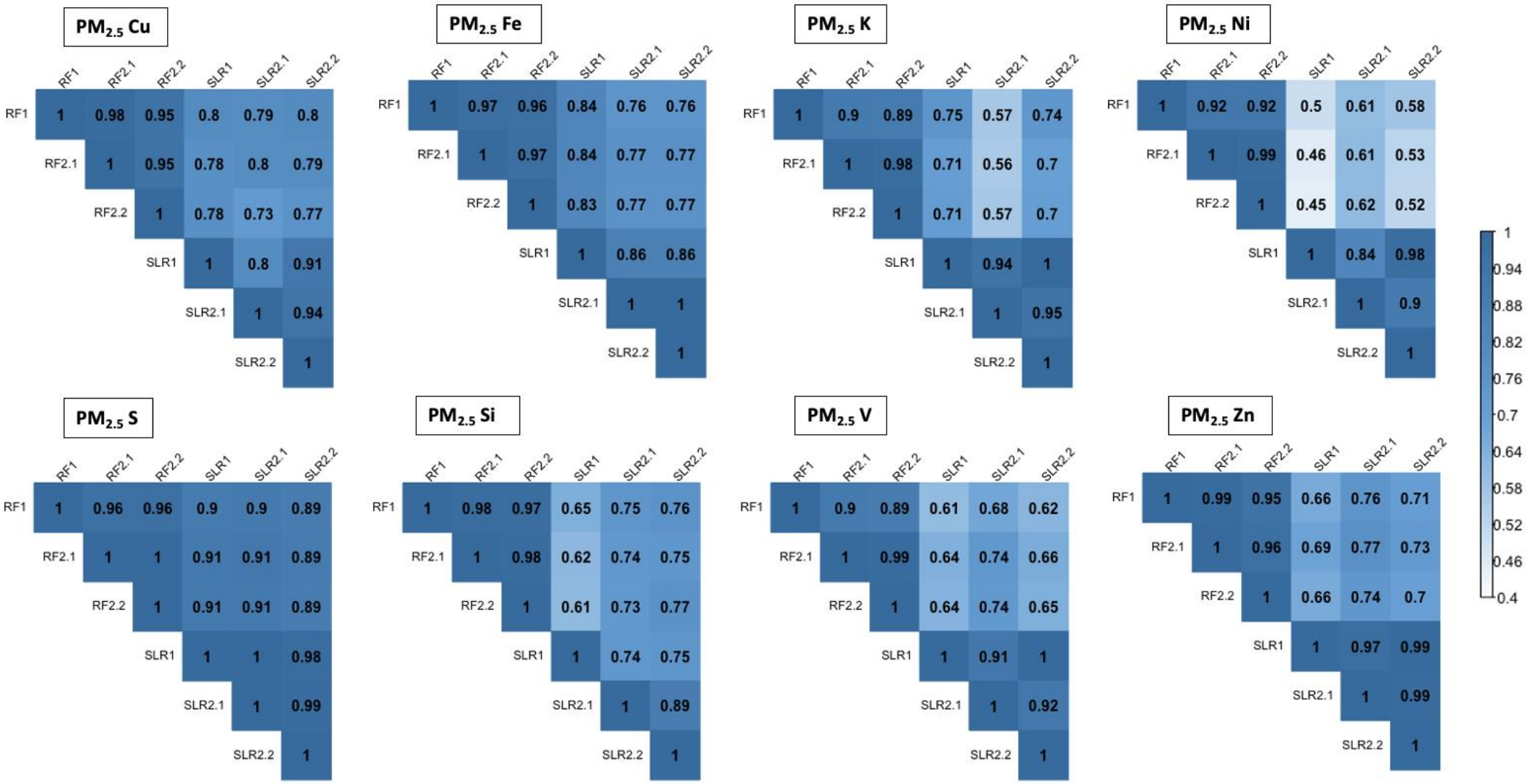


Figure S6. Pearson correlation between model predictions at random locations across ELAPSE countries (N=27,411)

SLR1 = One-step Supervised Linear Regression; SLR2.1 = Two-step Supervised Linear Regression, step1; SLR2.2 = Two-step Supervised Linear Regression, step2; RF1 = One-step Random Forest; RF2.1 = Two-step Random Forest, step1; RF2.2 = Two-step Random Forest, step2.