Supplementary Information

Land use regression models for Ultrafine Particles in six European areas

Erik van Nunen[1]*, Roel Vermeulen[1], Ming-Yi Tsai[2,3,4], Nicole Probst-Hensch[2,3], Alex Ineichen[2,3], Mark Davey[2,3], Medea Imboden[2,3], Regina Ducret-Stich[2,3], Alessio Naccarati[5], Daniela Raffaele[5], Andrea Ranzi[6], Cristiana Ivaldi[7], Claudia Galassi[8], Mark Nieuwenhuijsen[9,10,11], Ariadna Curto[9,10,11], David Donaire-Gonzalez[9,10,11], Marta Cirach[9,10,11], Leda Chatzi[12], Mariza Kampouri[12], Jelle Vlaanderen[1], Kees Meliefste[1], Daan Buijtenhuijs[1], Bert Brunekreef[1], David Morley[13], Paolo Vineis[5,13] John Gulliver[13], Gerard Hoek[1]

[1] Institute for Risk Assessment Sciences (IRAS), division of Environmental Epidemiology (EEPI), Utrecht University, Utrecht, the Netherlands
[2] Swiss Tropical and Public Health (TPH) Institute, University of Basel, Basel, Switzerland
[3] University of Basel, Basel, Switzerland
[4] Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA USA
[5] Human Genetics Foundation, Turin, Italy
[6] Environmental Health Reference Centre, Regional Agency for Prevention, Environment and Energy of Emilia-Romagna, Modena, Italy
[7] ARPA Piemonte, Turin, Italy
[8] Unit of Cancer Epidemiology, Citta' della Salute e della Scienza University Hospital and Centre for Cancer Prevention, Turin, Italy
[9] ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
[10] Department of Experimental and Health Sciences, Pompeu Fabra University (UPF), Barcelona, Spain
[11] CIBER Epidemiologia y Salud Pública (CIBERESP), Barcelona, Spain

[12] Department of Social Medicine, University of Crete, Heraklion, Greece
[13] MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, St Mary's Campus, London, United Kingdom

* Corresponding author
Institute for Risk Assessment Sciences (IRAS), division of Environmental Epidemiology (EEPI), Utrecht University, Yalelaan 2, 3584 CM Utrecht, the Netherlands; Tel +31 30 253 9474; Fax +31 30 253 9499; e-mail address: e.vannunen@uu.nl

For sumbission to ES&T:

Pages:	35
Figures:	7
Tables:	9

Supplement 1: Study areas and site distributions

All study areas covered in the EXPOsOMICS short-term campaign are presented in the map below along with the distribution of monitoring sites per area.

Figure S1; Map of Europe highlighting study areas.

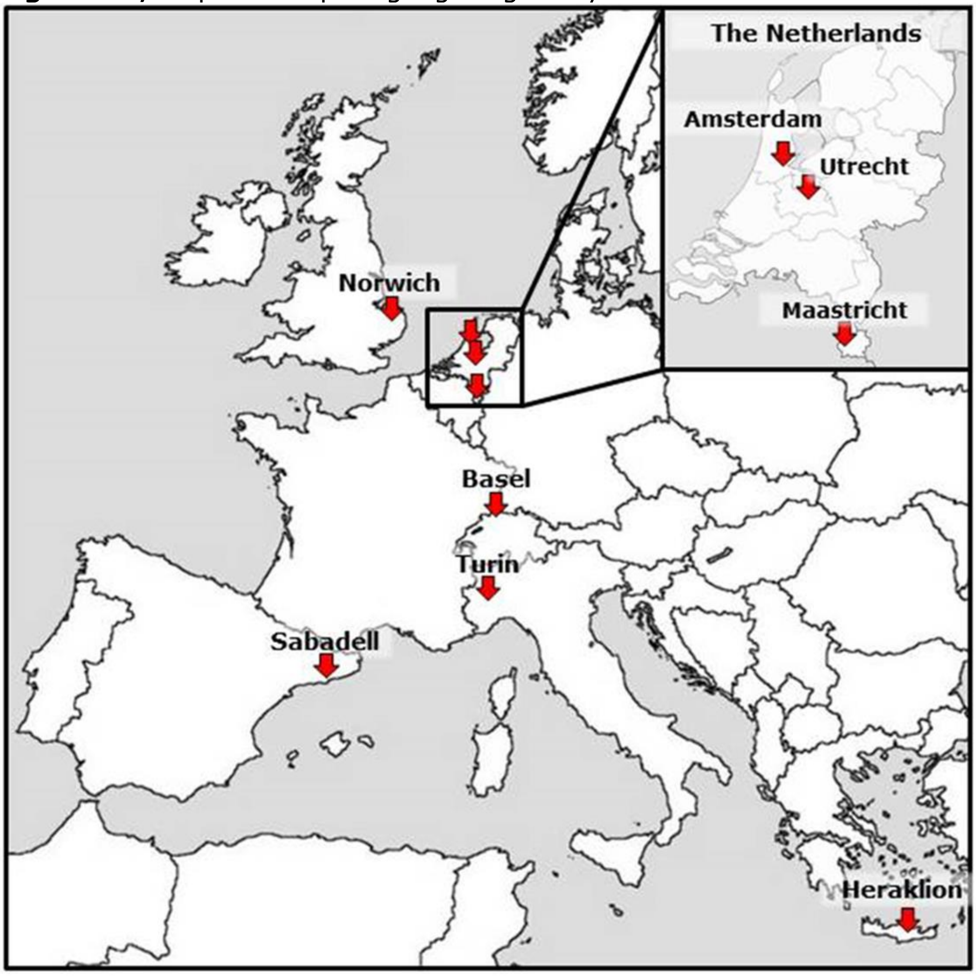
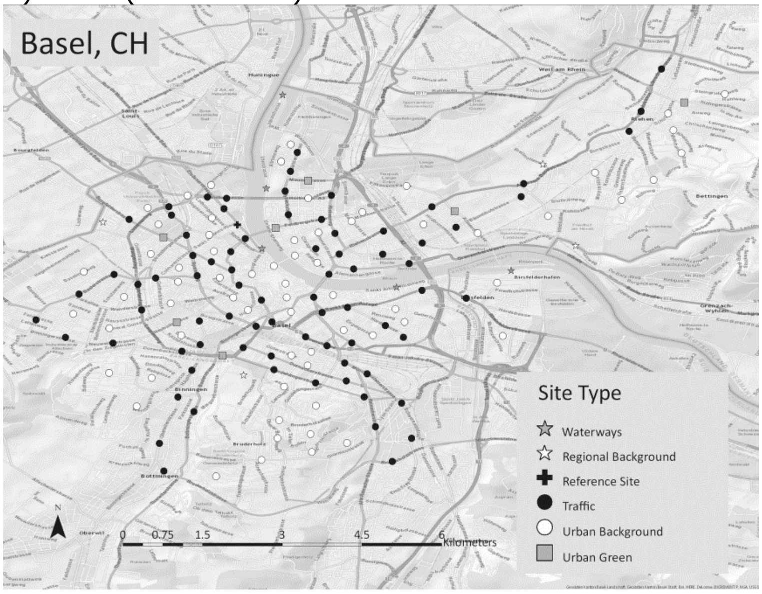
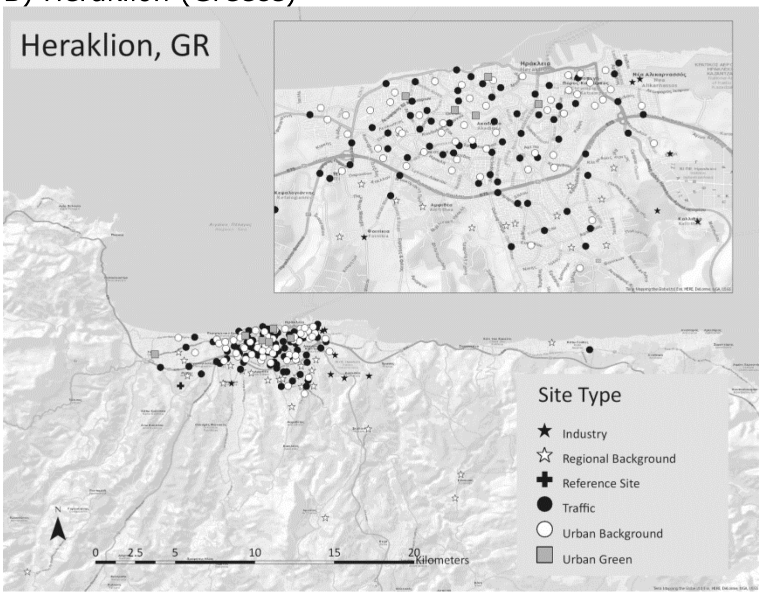


Figure S2; Detailed study area maps and distribution of monitoring sites.

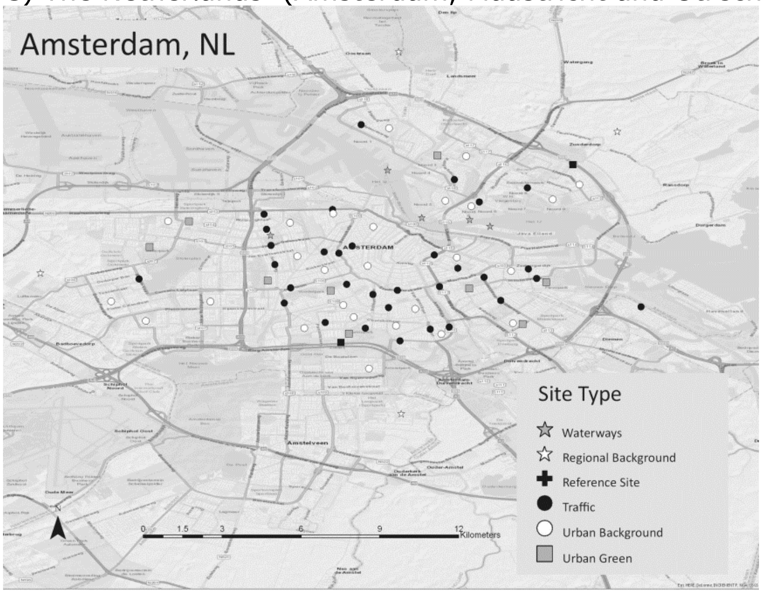
A) Basel (Switzerland)

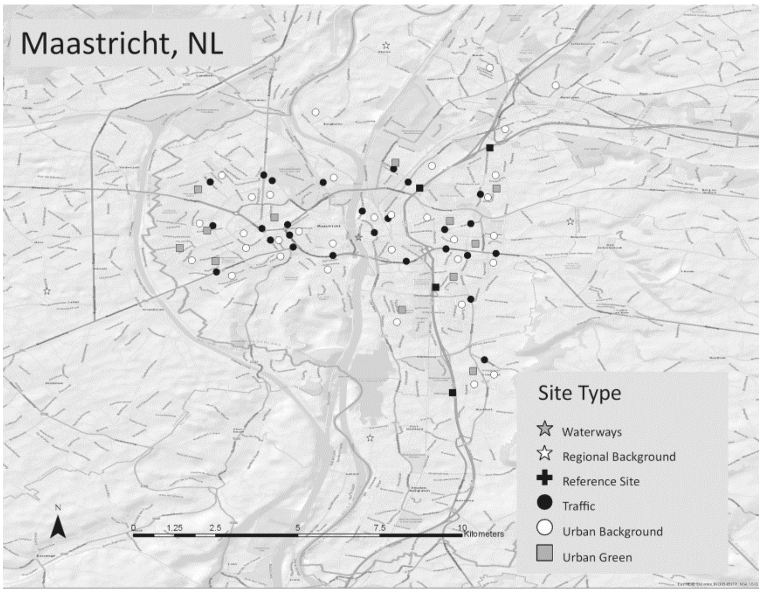


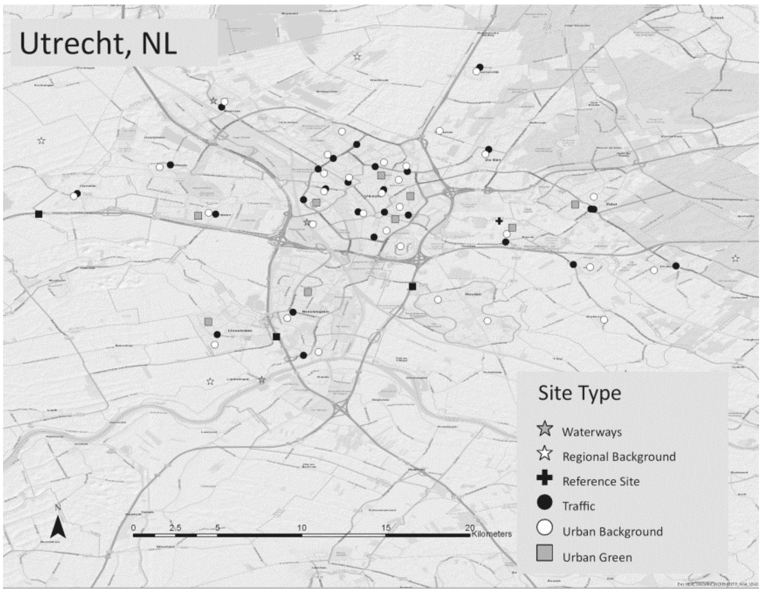
B) Heraklion (Greece)



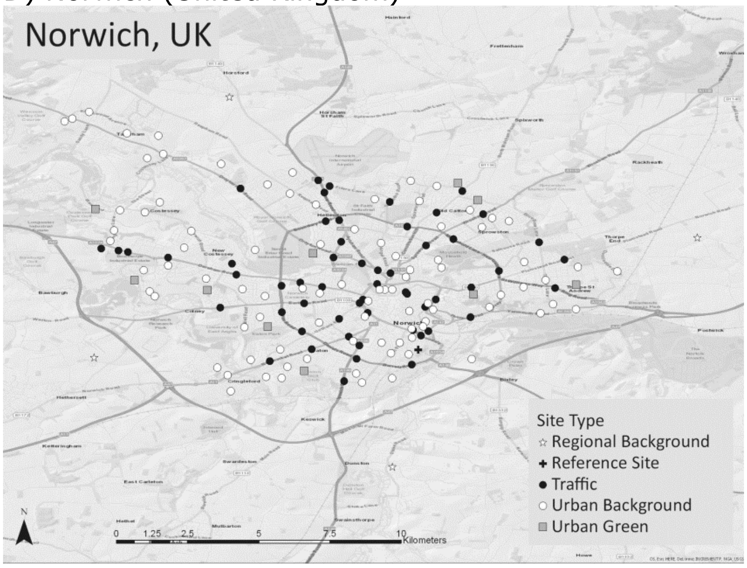
C) The Netherlands (Amsterdam, Maastricht and Utrecht)



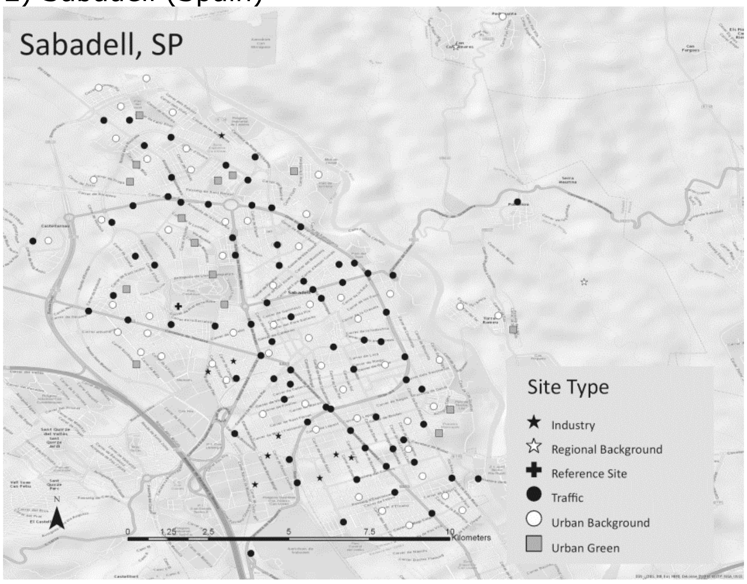




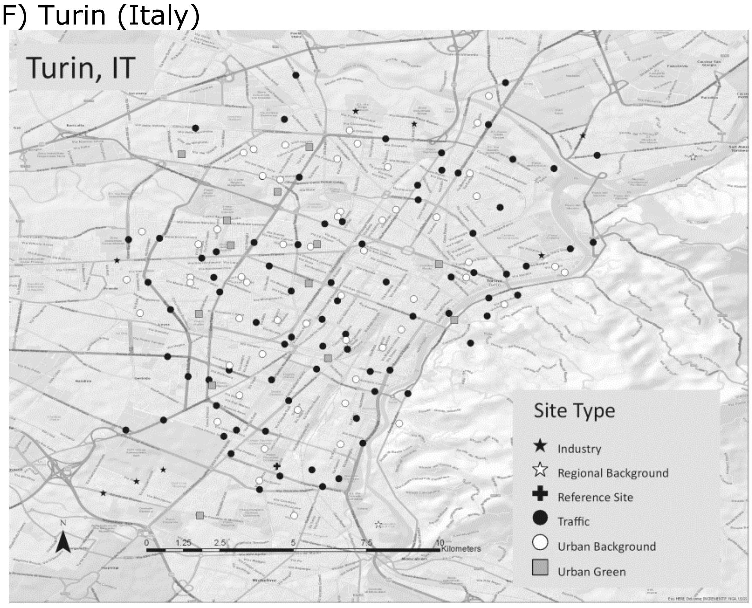
D) Norwich (United Kingdom)



E) Sabadell (Spain)



F) Turin (Italy)



Supplement 2: Co-location of UFP monitors in study areas.

In each study area two instruments were used, one for sampling the 160 (or 240) sites and one at the reference site. To evaluate consistency in UFP levels, devices were co-located in each area during the short-term monitoring campaign regularly for at least 180 minutes per comparison. In all study areas, the CPC 3007 (TSI Inc., Tennessee, USA) was used for monitoring the 160 (or 240) sites. In the Netherlands and Heraklion, another CPC3007 instrument was used at the reference site, while in the other four areas the MiniDiSC (Testo AG, Lenzkirch, Germany) was used at the reference site. Table S1 and Figure S1 present the results of the comparisons, expressed as the ratio of the UFP measurements with the instrument used at the sampling site and the instrument at the reference site. In the Netherlands, Norwich and Sabadell, the ratios of two instruments were close to unity. In Turin, the CPC used at the short-term sites gave about 30% lower readings than the MiniDiSC used at the reference site. We did not correct the measurements for these modest differences, as the reference site measurements is used only to correct for temporal variation using difference of the reference site measurement in a specific 30-minute period and the overall average. In Heraklion, the monitoring site CPC gave higher UFP readings than the reference site CPC with large variation. No trend over time was present. We did not correct the inconsistent comparisons, leading to added uncertainty of the correction for temporal variation.

Table S1; Agreement between monitoring devices

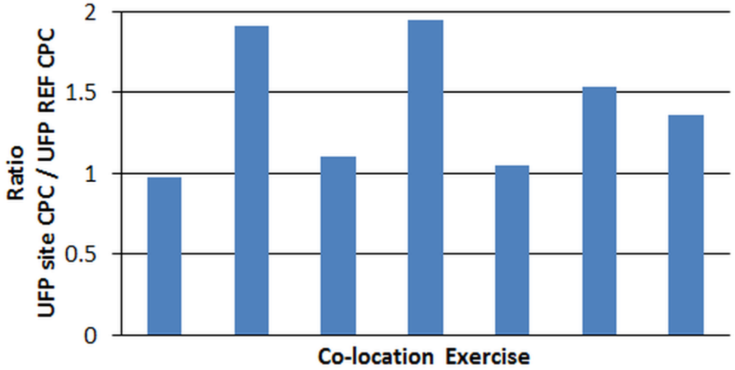
Study area	**Comparison**	**Mean (SD) ratio**[a]
Heraklion	CPC – CPC	1.41 (0.40)
Netherlands	CPC – CPC[b]	1.09 (0.16)
Norwich	CPC – Minidisc	1.02 (0.14)
Sabadell	CPC – Minidisc	0.86 (0.11)
Turin	CPC – Minidisc	0.73 (0.06)

[a] UFP instrument at monitoring sites / reference site
[b] Three comparisons of ref site CPC and a Minidisc resulted in a mean ratio of ~1

Figure S3; Co-location performed per study area, presented by average ratio per exercise (min. 180 minutes)

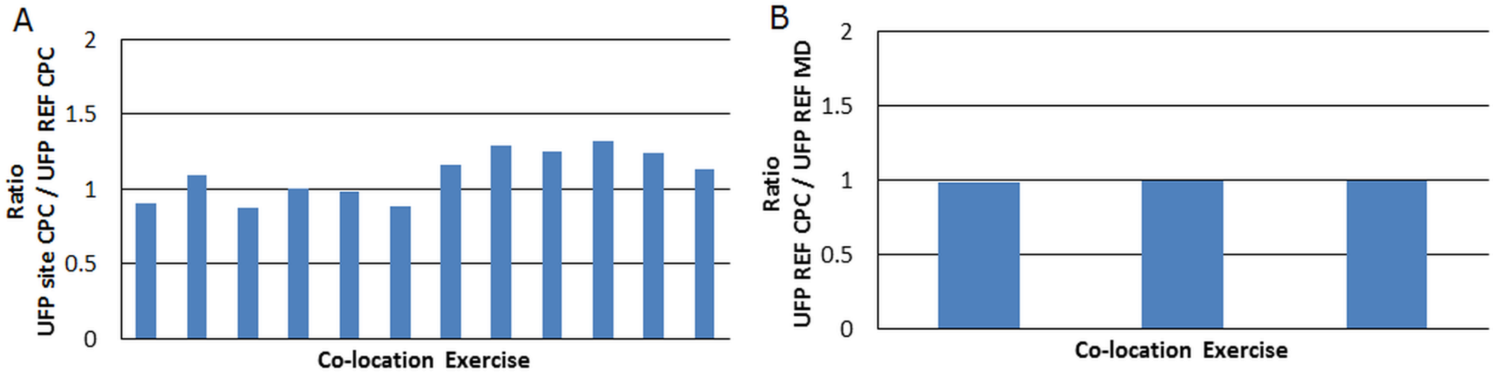
Heraklion

Average Ratio between reported UFP by the CPC from the mobile campaign and CPC at the reference site



The Netherlands

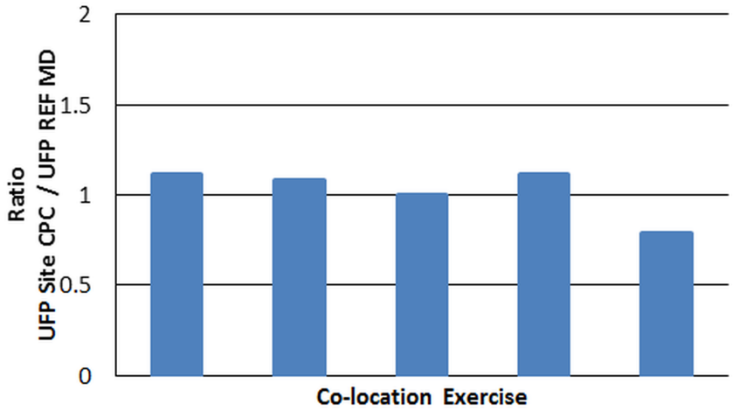
Average Ratio between reported UFP by the CPC from the mobile campaign and CPC at the reference site (A)



Average Ratio between reported UFP by the CPC from the reference site and MiniDiSC (MD) at the reference site (B)

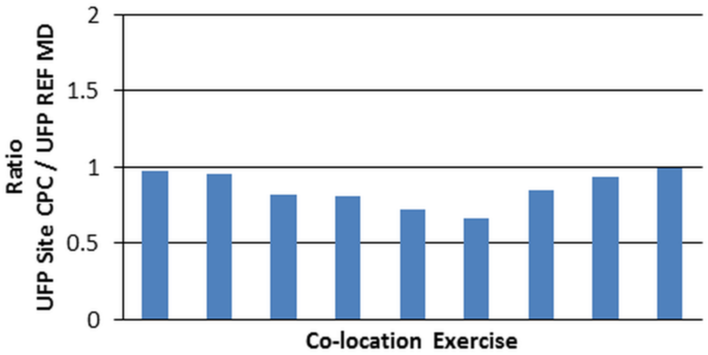
Norwich

Average Ratio between reported UFP by the CPC from the mobile campaign and MiniDiSC (MD) at the reference site

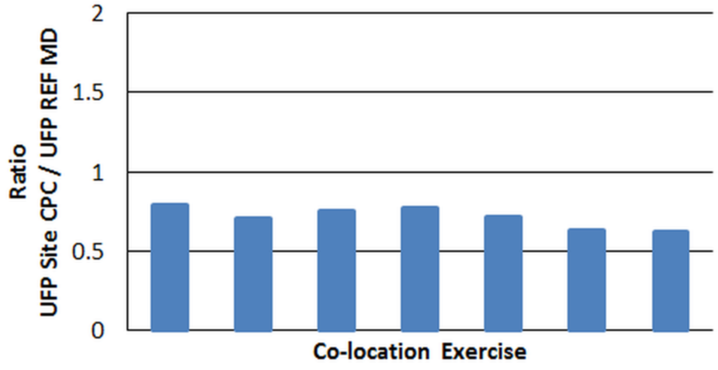


Sabadell

Average Ratio between reported UFP by the CPC from the mobile campaign and MiniDiSC (MD) at the reference site



Turin
Average Ratio between reported UFP by the CPC from the mobile campaign and MiniDiSC (MD) at the reference site

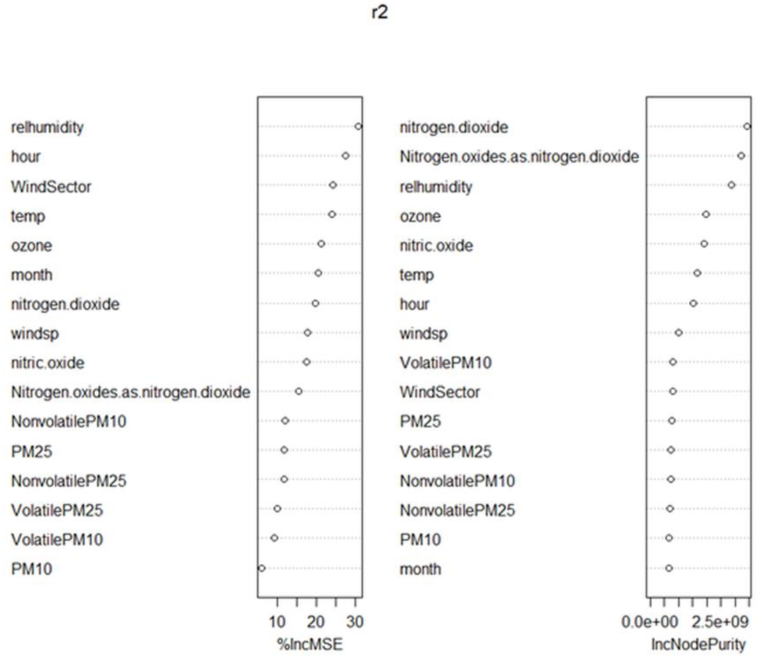


Supplement 3: Regression models applied to calculate missing Reference Site UFP observations.

Missing 30 minute reference site UFP measurements arose in all study areas due to cleaning of spurious UFP readings, removal of device reported error messages and mismatches in monitoring times. These missing 30-minute values were imputed per area by applying regression models built on routine air pollution and meteorological data, using available 30-minute reference site UFP observations as dependent variable. The default was to use linear regression models. When model R^2 exceeded 50%, regression models were accepted for prediction of reference UFP concentrations. Only in Torino, the model R^2 of linear models exceeded 50%. In Norwich linear regression models achieved an R^2 just below 50%. As the number of missing data was appreciable, a random forest model was then applied which achieved a R^2 of 50%. In the other three areas, the R^2 of linear models was below 6% and no further modelling was attempted.

Norwich
Reference site UFP observations were missing for 78/480 site measurements (16.5%).
Routine and Meteorological data were applied in a Random Forest regression model, explaining 50.2% in UFP variation for 402 available observations.

Figure S4; Output Random Forrest Regression



Turin
Reference site UFP observations were missing for 313/480 measurements (65.2%).
Routine and Meteorological data were applied in a linear regression model, explaining 62.1% in UFP variation for 167 available observations.
312/313 reference site UFP values could be imputed by applying the model below, routine NOx was missing for one 30 minute interval.

Reference UFP (cm^{-3}) = 3.860e+05 + 9.800e+01 * Routine NOx (μg/m3) - 1.391e+03 * Hour - 3.597e+00 * Barometric Pressure (0.1hPa) - 8.701e+01 * Relative Humidity (%)

Basel
Reference site UFP observations were missing for 36/480 measurements (7.5%).
Routine and Meteorological data could not explain 30-minute UFP observations for the 444 available observations in a regression model, where the highest observed R^2 for a single predictor reached 0.3%.

Heraklion
Reference site UFP observations were missing for 85/480 measurements (17.7%).
Routine and Meteorological data could not explain 30-minute UFP observations for the 395 available observations in a regression model, where the highest observed R^2 for a single predictor did not exceed 2%.

The Netherlands
Reference site UFP observations were missing for 48/723 measurements (6.6%).
Routine and Meteorological data could not explain 30-minute UFP observations for the 675 available observations in a regression model, where the highest observed R^2 for a single predictor reached 5.7%.

Sabadell
Reference site UFP observations were missing for 31/480 measurements (6.5%).
Routine and Meteorological data could not explain 30-minute UFP observations for the 449 available observations in a regression model, where the highest observed R^2 for a single predictor reached 3.8%.

Supplement 4: GIS predictors for Land Use Regression Modelling in the EXPOsOMICS study.

- Starting point are the GIS predictors previously applied in the MUSIC study (Montagne et al. 2015). For predictor deletions and additions ESCAPE predictors were also evaluated (Eeftens, Beelen, et al. 2012). Airport was added as buffer rather than distance, which is difficult to define given the large area an airport covers.
  Restaurant density was added as number of amenities in a buffer radius given that restaurant data consisted of both spot and polygon data.
- Buffer sizes have been adapted to the sizes for which there were sufficient numbers of non-zero values expected.

TableS2; overview GIS predictors

Predictor Variable	**Variable Name**	**Units**	**Direction**	**Buffer sizes (m)**
SPATIAL PREDICTORS				
CORINE land use predictors				
Industry	INDUSTRY	m^2	+	100, 300, 500, 1000, 5000
Port	PORT	m^2	+	100, 300, 500, 1000, 5000
Airport	AIRPORT	m^2	+	1000, 5000
Urban Green	URBGREEN	m^2	-	100, 300, 500, 1000, 5000
Semi-natural and forested areas	NATURAL	m^2	-	100, 300, 500, 1000, 5000
Low density residential land	LDRES	m^2	+	100, 300, 500, 1000, 5000
High density residential land	HDRES	m^2	+	100, 300, 500, 1000, 5000
Sum of low and high density residential land	HDLDRES	m^2	+	100, 300, 500, 1000, 5000
Sum of URBGREEN & NATURAL	UGNL	m^2	-	100, 300, 500, 1000, 5000
Other spatial predictors				
Population data	POPEEA	m^2	+	100, 300, 500, 1000, 5000
Household density	HHOLD	Number	+	100, 300, 500, 1000, 5000
Traffic intensity on nearest road	TRAFNEAR	Veh. day^{-1}	+	

Inverse distance to nearest road	DISTINVNEAR1	m^{-1}	+	
Product of traffic intensity on nearest road and inverse distance to the nearest road	INTINVDIST	Veh. $day^{-1}m^{-1}$	+	
Traffic intensity on nearest major road	TRAFMAJOR	Veh. day^{-1}	+	
Inverse distance to nearest major road	DISTINVMAJOR1	m^{-1}	+	
Product of traffic intensity in nearest major road and inverse of distance to nearest major road	INTMAJORINVDIST	Veh. $day^{-1}m^{-1}$	+	
Total traffic load of major roads in a buffer (sum of (traffic intensity*length of all segments))	TRAFMAJORLOAD	Veh. $day^{-1}m$	+	50, 100, 300, 500, 1000
Traffic total load of roads in a buffer (sum of (traffic intensity * length of all segments))	TRAFLOAD	Veh. $day^{-1}m$	+	50, 100, 300, 500, 1000
Heavy-duty traffic intensity om nearest road	HEAVYTRAFNEAR	Veh. day^{-1}	+	
Product of heavy-duty traffic intensity on nearest road and inverse distance to the nearest road	HEAVYINTINVDIST	Veh. $day^{-1}m^{-1}$	+	
Heavy-duty traffic intensity om nearest major road	HEAVYTRAFMAJOR	Veh. day^{-1}	+	
Total heavy-duty traffic load of all major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments)	HEAVYTRAFMAJORLOAD	Veh. $day^{-1}m$	+	50, 100, 300, 500, 1000
Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments)	HEAVYTRAFLOAD	Veh. $day^{-1}m$	+	50, 100, 300, 500, 1000
Road length of all roads in a buffer	ROADLENGTH	m	+	50, 100, 300, 500, 1000
Road length of all major roads in a buffer	MAJORROADLENGTH	m	+	50, 100, 300, 500, 1000
Restaurants [a] [b]	RESTAURANT	Number	+	100, 300, 500, 1000, 5000
Altitude above sea level	SQRALT	m	-	Square root altitude
Distance to major point source	DISTPOINT	m	-	If applicable to the area

[a] Generated from the Overpass Turbo Web application, selecting amenities marked as “Restaurant”, “Fast_food”, “Pub” or “Cafe” in OpenStreetMap. It is plausible that restaurant density is underreported in all areas, since owners should actively report their facility and pay a fee to be in the OpenStreetMap database. Local researchers evaluated plausibility of restaurant representation and decided on the use of this data.

[b] Data not collected for Heraklion, coverage of amenities was low and differential among neighborhoods

Supplement 5: Local and combined Land Use Regression (LUR) models developed within the EXPOsOMICS study.

- 10 local models / 10 combined models on pooled data were developed, each built on 90% of the site measurements (Model R2 is shown), and subsequently validated on the other 10% (Holdout Validation not shown).
  Model structures per model per area are presented below;
- A predictor was used when there was at least 10% representation over monitoring sites (90^{th} percentile differed from 0).
  In addition for EU models, a predictor had to be represented in 3 or more (≥50%) of study areas;
- Predictor coefficients presented are multiplied by the spread in the specific predictor, calculated as the 90^{th} – 10^{th} percentile, expressing the proportional change in UFP for an increase between the 10^{th} and 90^{th} percentile of the predictor;
- Predictor categories are presented in the first column;
  (NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry, RE = Restaurants, PT = Port, AI = Airport, GR = Greenspace).

Table S3; Model structures per study area

A) Basel

		PREDICTOR 10^{th} – 90^{th} percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		5332	6589	5784	5697	6013	6635	4922	7014	6487	6381
NT	Traffic intensity on nearest road	**17038**	4306*	6261	6909	5264	6639	7095	5157	6278	4275	6397
NT	Road length of all major roads in a buffer of 50m	**100**	1544			1655			1373		1522	
PP	Sum of low and high density residential land in a buffer of 500m	**356207**					2548		3017	2029	2050	1817
PP	Sum of low and high density residential land in a buffer of 1000m	**1234324**	2775	2149	2380	2398		2027				
RE	Number of restaurants in a buffer of 100m	**2**	2570	2761	3026	2404	2769	3416	2008	2808	3078	1400
RE	Number of restaurants in a buffer of 1000m	**97**										2831
Model R^2			**29.3%**	**26.2%**	**32.3%**	**30.4%**	**28.0%**	**33.8%**	**27.9%**	**27.0%**	**31.3%**	**30.9%**

NT = Nearby Traffic, PP = Population, RE = Restaurants
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor

B) Heraklion

		PREDICTOR 10^{th} – 90^{th} percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		4709	1493	3873	1065	6255	3258	2534	2821	1759	1701
NT	Traffic intensity on nearest road	**14154**	6830*	6851	8437	5770	6466	6429	2961	6163	6727	7154
NT	Road length of all roads in a buffer of 50m	**234**		3851					2539	3958	3532	2710
NT	Product of traffic intensity on nearest road and inverse distance to the nearest road	**1630**				2756			6562			
DT	Road length of all roads in a buffer of 500m	**17125**				2872						
PP	Population land use in a buffer of 100m	**1292**				3417		2455				
PP	Population land use in a buffer of 300m	**11270**	3299	3342	3250							2329
IN	Industry within a buffer of 5000m	**3201922**	4861	4665	4934	5306		6561	3407		4585	5222
AI	Airport within a buffer of 5000m	**2830488**					3903			3575		
PT	Port within a buffer of 1000m	**306569**					3173		3715	2788		
GR	Urban green + semi-natural and forested areas within a buffer of 500m	**112393**	-1938			-2570						
GR	Urban green + semi-natural and forested areas within a buffer of 1000m	**379034**		-2917								
Model R^2			**32.6%**	**38.9%**	**35.0%**	**40.8%**	**33.3%**	**33.7%**	**44.9%**	**35.2%**	**34.7%**	**35.8%**

NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry, AI = Airport, PT = Port, GR = Greenspace
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor

C) The Netherlands

		PREDICTOR $10^{th} - 90^{th}$ percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		8045	8315	7828	8568	8355	8333	7938	7663	6800	8314
NT	Traffic intensity on nearest road	**14909**	2339*	2408		4434	3865		2277	2650	3389	
NT	Heavy traffic intensity on nearest road	**416**			1562			1504				1825
NT	Traffic total load of roads in a buffer of 50m (sum of(traffic intensity * length of all segments))	**2197210**	3621	3975	3588			3559	3518	3592		3001
NT	Total heavy-duty traffic load of all roads in a buffer of 50m (sum of(heavy-duty traffic intensity * length of all segments)	**91509**				860						
NT	Road length of all major roads in a buffer of 50m	**174**	2571		3490	2096	4066	2995	3327		4183	3382
NT	Road length of all major roads in a buffer of 100m	**389**		2336						2168		
DT	Traffic total load of roads in a buffer of 300m (sum of(traffic intensity * length of all segments))	**34730902**				2431	3161				3091	
PP	Household density in a buffer of 1000m	**21417**	4646					3221	4955			
PP	Sum of low and high density residential land in a buffer of 5000m	**34317497**									2780	
PP	Population land use in a buffer of 5000m	**459495**		2609	3554					4995		3721
IN	Industry within a buffer of 300m	**6299**										172
IN	Industry within a buffer of 500m	**108639**	1405	734	775	1100	834	1031		794		
PT	Port within a buffer of 5000m	**8495354**		2413	2404	3296	3101	2404			2319	
Model R^2			**45.8%**	**45.9%**	**49.5%**	**49.7%**	**48.5%**	**47.0%**	**50.9%**	**45.0%**	**49.2%**	**47.4%**

NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry, PT = Port

* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor

D) Norwich

		PREDICTOR 10th – 90th percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		2350	2523	217	3017	2547	2603	2795	1495	2192	2501
NT	Traffic intensity on nearest road	**14483**	7155*	4482	6766	5115	4931	4171	4873		2911	5367
NT	Traffic total load of roads in a buffer of 50m (sum of(traffic intensity * length of all segments))	**1904994**		2875		2720		3174	2673	3757		
NT	Road length of all roads in a buffer of 50m	**155**			2361							
NT	Road length of all major roads in a buffer of 50m	**99**					1903					1950
NT	Product of traffic intensity on nearest road and inverse distance to the nearest road	**2224**								2637	3194	
PP	Population land use in a buffer of 5000m	**90942**	2893	2604	3184	2273	3394	2290	2343	3861	3444	3440
IN	Industry within a buffer of 500m	**154676**				3544						2582
IN	Industry within a buffer of 1000m	**501510**	2912	3019	3278		2373	3765	3360	3443	3786	
AI	Airport within a buffer of 5000m	**2388905**	1978	2009		1818		2260	2007			
Model R^2			**43.1%**	**38.7%**	**39.9%**	**40.6%**	**35.9%**	**42.6%**	**37.4%**	**37.1%**	**37.6%**	**39.9%**

NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry, AI = Airport

* All presented coefficients are Model coefficients multiplied by the value of the 10th-90th percentile of the predictor

E) Sabadell

		PREDICTOR 10^{th} – 90^{th} percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		9036	9194	9142	9321	9874	9982	9189	9222	5524	9284
NT	Traffic intensity on nearest road	**33758**	5225*	6208	5936	5212	6869	6056	5823	5894	4610	6613
DT	Road length of all major roads in a buffer of 1000m	**7724**									3395	
IN	Industry within a buffer of 300m	**12517**	386	353	368	404	328	301	435	362		438
IN	Industry within a buffer of 5000m	**6563930**									2678	
RE	Number of restaurants in a buffer of 100m	**5**				2244					3310	
RE	Number of restaurants in a buffer of 1000m	**182**	8417	7719	7774	7216	6374	6391	7658	7920	5274	7501
Model R^2			**25.5%**	**27.4%**	**26.9%**	**29.4%**	**30.1%**	**30.6%**	**27.7%**	**27.3%**	**30.2%**	**27.6%**

NT = Nearby Traffic, DT = Distant Traffic, IN = Industry, RE = Restaurants
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor

F) Turin

		PREDICTOR 10^{th} – 90^{th} percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		6413	7126	6740	8764	8812	7691	7014	7467	8518	6860
NT	Traffic total load of roads in a buffer of 50m (sum of(traffic intensity * length of all segments))	**3368797**	8840*	8304	5912	8712	9035	8937	8557	8264	9038	8190
NT	Road length of all major roads in a buffer of 50m	**100**			2427							
PP	Sum of low and high density residential land in a buffer of 100m	**12848**	2642	2386	2324	1961	1711	2144	2390	2212	1854	2460
IN	Industry within a buffer of 1000m	**461740**	1284	1231	1092				1085	1198		1413
GR	Urban green within a buffer of 1000m	**391625**				-1717						
Model R^2			**41.3%**	**38.3%**	**41.9%**	**41.1%**	**41.7%**	**37.9%**	**42.8%**	**39.5%**	**39.9%**	**39.0%**

NT = Nearby Traffic, PP = Population, IN = Industry, GR = Greenspace
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor

Table S4; Model structures of combines area models
COMBINED AREA MODEL

		PREDICTOR 10^{th} – 90^{th} percentile	UFP Model1	UFP Model2	UFP Model3	UFP Model4	UFP Model5	UFP Model6	UFP Model7	UFP Model8	UFP Model9	UFP Model10
	Intercept		6584	6450	6650	7134	5780	7268	6840	6815	6597	6602
NT	Traffic intensity on nearest road	**21097**	4883*	5289	5002	4685	5173	4504	5062	5133	4421	5183
NT	Road length of all roads in a buffer of 50m	**215**					1085					
NT	Road length of all major roads in a buffer of 100m	**266**	2160	1632	1734	2246	1496	2139	1759	1875	2036	1861
DT	Traffic total load of roads in a buffer of 1000m (sum of(traffic intensity * length of all segments))	**465493528**	1396	1862	1396	931	1396	1396	1396	1396	1396	1396
PP	Sum of low and high density residential land in a buffer of 1000m	**1989948**	925						1361		1165	
PP	Sum of low and high density residential land in a buffer of 5000m	**32357316**	1860	2403	2259	1597	2685	1796	1389	2240	1449	2354
PP	Population land use in a buffer of 300m	**5512**				2683		2773				
PP	Population land use in a buffer of 1000m	**45496**								1689		
PP	Population land use in a buffer of 5000m	**448077**	2039	1812	2589		1941		2103		2168	2233
IN	Industry within a buffer of 5000m	**7101377**	959	1321	831	1349	1129	1023	824	1122	1058	1150
Model R^2‡			**34.6%**	**34.1%**	**32.8%**	**34.8%**	**35.3%**	**32.6%**	**33.9%**	**33.1%**	**32.1%**	**33.6%**
Random area effects:												
	Basel		-247	50	-132	212	-38	164	-7	218	-52	-160
	Heraklion		-362	-320	-284	-879	-303	-810	-270	-653	-329	-280
	The Netherlands		414	402	433	566	362	366	403	220	285	259
	Norwich		-763	-654	-1066	-437	-667	-338	-987	-978	-1095	-856
	Sabadell		1510	1079	2033	1338	1278	1601	1471	1130	1651	1461
	Turin		-552	-557	-984	-799	-632	-983	-610	63	-460	-424

NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor
‡ Model R^2 based development and HV in linear regression, prior to introduction of random intercept

Supplement 6; Robustness analysis of predicted UFP concentrations

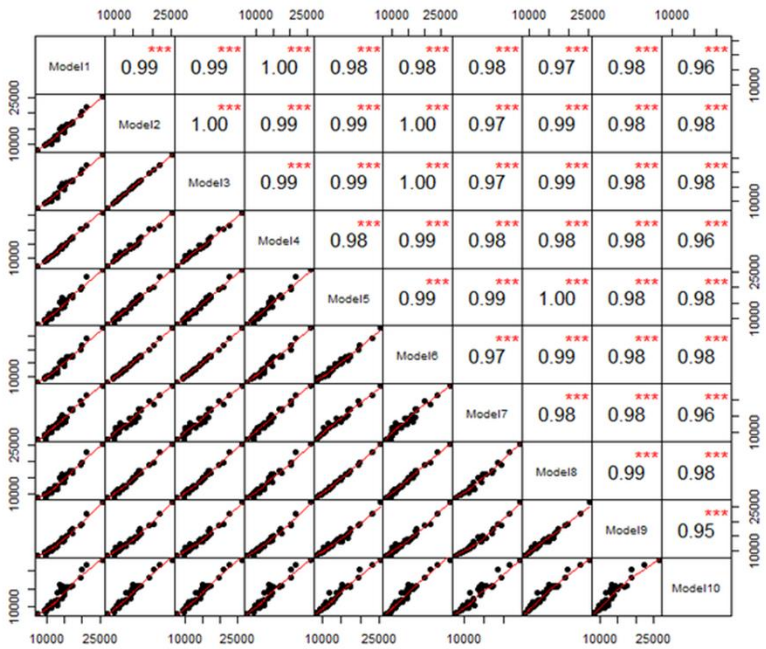
A 10-fold Holdout Validation (HV) approach was applied at model development, reducing potential model sensitivity to a single predictor in local/combined models. This approach allowed variation between models per area, aiming to generate more precise exposure predictions in epidemiological studies. 10 unique local/combined models were built that differed in intercept, predictors and/or coefficients. Consistency in UFP predictions was analyzed to assess model robustness.

1) Robustness of Local models

Model robustness of Local models was tested on external sites from each study area. These were home address locations visited in a Personal Exposure Monitoring campaign (Basel N=48, Heraklion N=50, the Netherlands N=42, Norwich N=31, Sabadell N=42, Turin N=44), also performed within the framework of EXPOsOMICS. Tables below show both a plot and a Pearson correlation coefficient for predicted UFP concentrations from all models per area tested against each other.

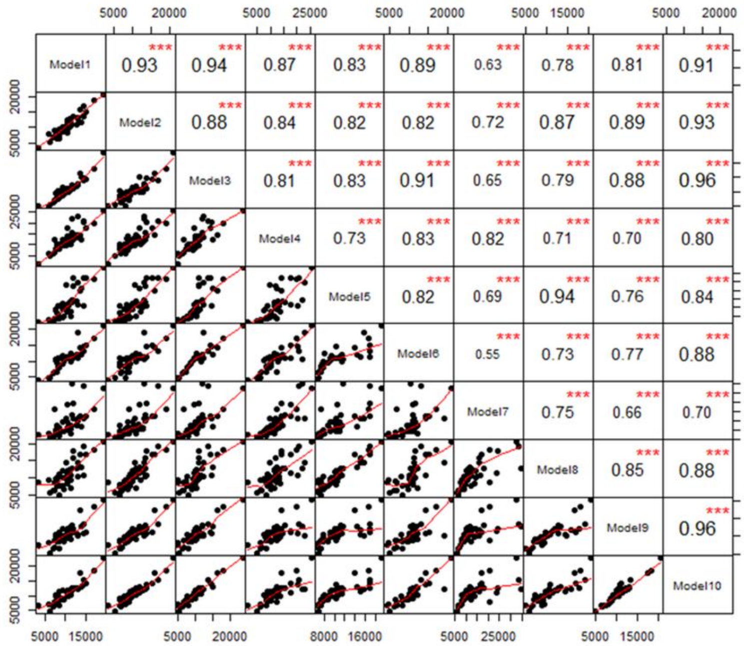
Figure S5; Correlation matrix and Pearson Correlation Coefficients of predicted UFP levels over 10 models per area

A) Basel



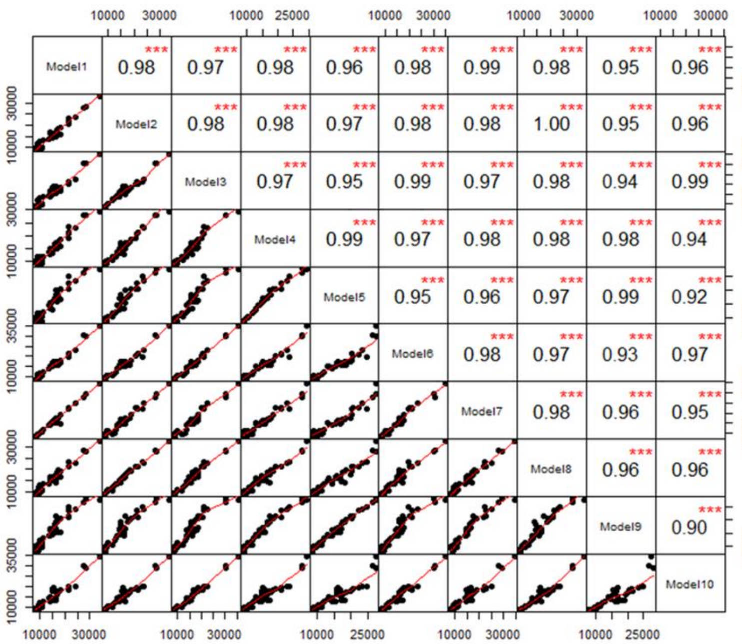
Red lines represent the best fit lines; *** = p-value < 0.001

B) Heraklion



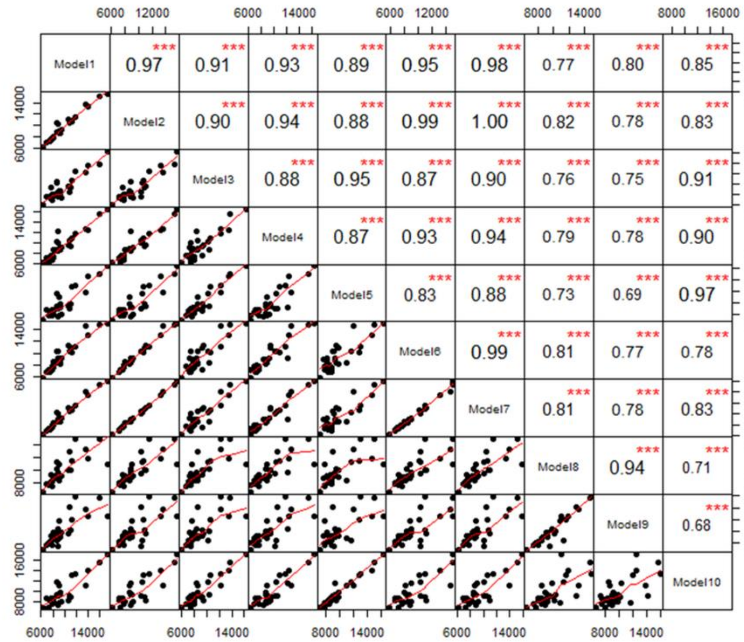
Red lines represent the best fit lines; *** = p-value < 0.001

C) The Netherlands



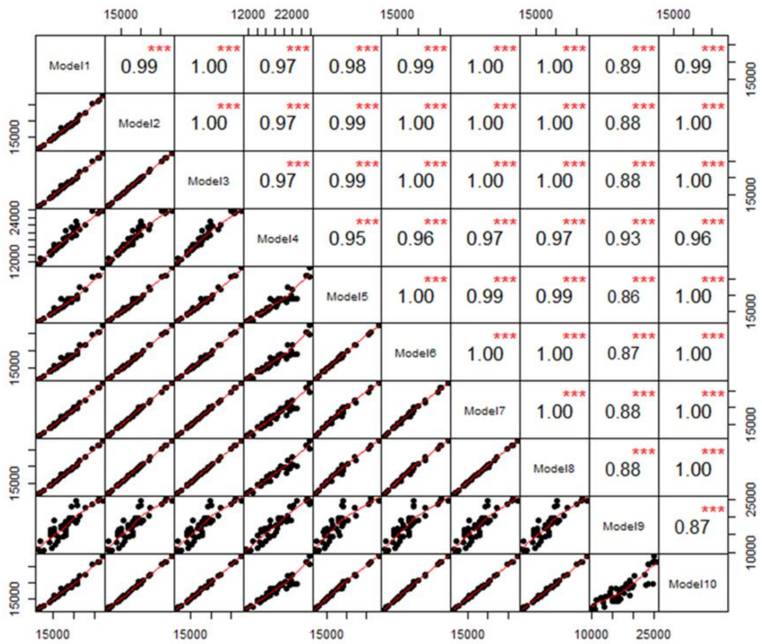
Red lines represent the best fit lines; *** = p-value < 0.001

D) Norwich



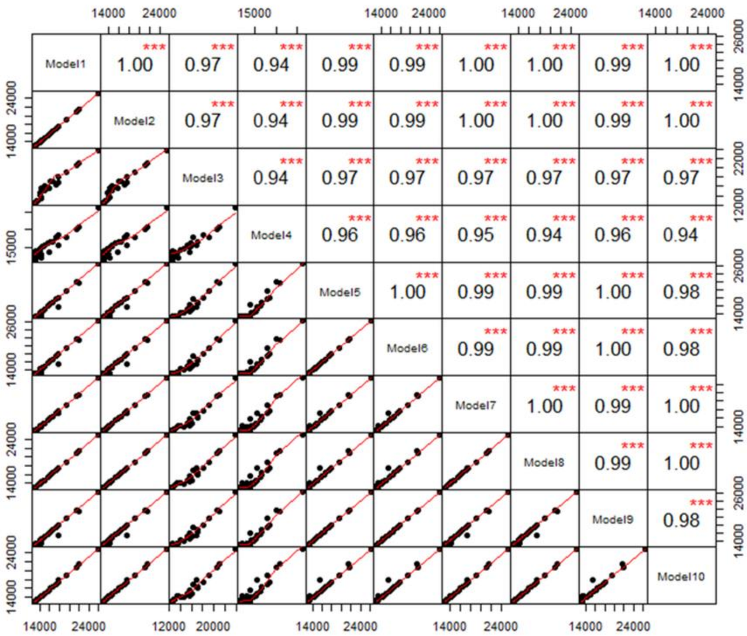
Red lines represent the best fit lines; *** = p-value < 0.001

E) Sabadell



Red lines represent the best fit lines; *** = p-value < 0.001

F) Turin

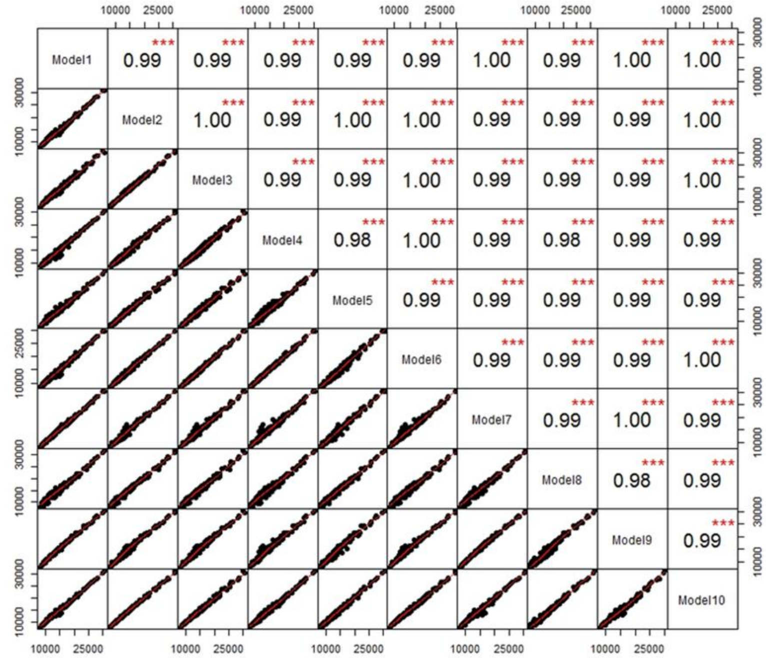


Red lines represent the best fit lines; *** = p-value < 0.001

2. Robustness of Combined area Models

Model robustness of Combined area models was tested on the same external sites, now pooled over all study areas (N=257). Again, a plot and a Pearson correlation coefficient are presented for predicted UFP concentrations from all models tested against each other.

Figure S6; Correlation matrix and Pearson Correlation Coefficients of predicted UFP levels over the 10 Combined area models



Red lines represent the best fit lines; *** = p-value < 0.001

Supplement 7; Full area models, built upon 100% of the short-term monitoring sites

- Full area models per area and on pooled data were developed on all short-term monitoring sites. Model structures per area are presented below;
- A predictor was used when there was at least 10% representation over monitoring sites (90th percentile differed from 0). In addition for EU models, a predictor had to be represented in 3 or more (≥50%) of study areas;
- Predictor coefficients presented are multiplied by the spread in the specific predictor, calculated as the 90th – 10th percentile, expressing the proportional change in UFP for an increase between the 10th and 90th percentile of the predictor;
- Predictor categories are presented in the first column; (NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry, RE = Restaurants, AI = Airport).

Table S5; Model structures of models based on 100% of the sites per area

	FULL AREA model Based on 100% of the short-term sites per area	**UFP** (90th-10th percentile) **BASEL**	**UFP** (90th-10th percentile) **HERAKLION**	**UFP** (90th-10th percentile **NETHERLANDS**	**UFP** (90th-10th percentile) **NORWICH**	**UFP** (90th-10th percentile) **SABADELL**	**UFP** (90th-10th percentile) **TURIN**	**UFP** (90th-10th percentile) **COMBINED AREA***
	Intercept	**6561**	**3714**	**7785**	**2587**	**9304**	**7083**	**6598**
NT	Traffic intensity on nearest road	**6432*** (17038)	**4993** (14154)	**2499** (14909)	**4488** (14483)	**5922** (33758)		**4931** (21097)
NT	Traffic total load of roads in a buffer of 50m (sum of(traffic intensity * length of all segments))			**3456** (2.197e6)	**2881** (1.905e6)		**8585** (3.369e6)	
NT	Road length of all major roads in a buffer of 50m			**2874** (174.3)				
NT	Road length of all major roads in a buffer of 100m							**1923** (266)
NT	Product of traffic intensity on nearest road and inverse distance to the nearest road		**3284** (1802)					
DT	Traffic total load of roads in a buffer of 1000m (sum of(traffic intensity * length of all segments))							**1419** (4.655e8)
PP	Population land use in a buffer of 300m		**2919** (11270)					**908** (5512)
PP	Population land use in a buffer of 1000m							**1606** (45496)
PP	Population land use in a buffer of 5000m			**4710** (459495)	**2574** (90942)			
PP	Sum of low and high density residential land in a buffer of 100m						**2361** (12848)	

PP	Sum of low and high density residential land in a buffer of 1000m	**2106** (1.234e6)						
PP	Sum of low and high density residential land in a buffer of 5000m							**2228** (3.236e7)
IN	Industry within a buffer of 300m					**361** (12517)		
IN	Industry within a buffer of 500m			**841** (108639)				
IN	Industry within a buffer of 1000m				**3147** (501510)		**1121** (461740)	
IN	Industry within a buffer of 5000m		**5365** (3.202e6)					**1010** (7.101e7)
RE	Number of restaurants in a buffer of 100m	**2678** (2)						
RE	Number of restaurants in a buffer of 1000m					**7702** (182.3)		
AI	Airport within a buffer of 5000m				**1682** (2.389e6)			
Model R^2 ‡		**28.3%**	**35.4%**	**46.9%**	**39.1%**	**27.2%**	**39.1%**	**33.7%‡**

The combined area model has a random intercept per area:

-174 Basel
-260 Heraklion
+350 the Netherlands
-965 Norwich
+1653 Sabadell
-604 Turin

NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry, AI = Airport, RE = Restaurant
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor
‡ Model R^2 based development and HV in linear regression, prior to introduction of random intercept

Supplement 8; Linear Mixed-Effect Models on Combined area LUR models

LUR models for Ultrafine Particles (UFP) were developed at a large scale using adjusted average UFP concentrations at 1043 monitoring sites from six study areas combined, based on two or three reference site corrected 30 minute observations from different seasons, sampled according a standardized protocol. In model building, a Linear Regression model was developed, selecting predictors in a supervised stepwise selection procedure. The final Linear Regression models on combined data was analyzed for differences in background UFP between study areas, evaluating the introduction of a random intercept per study area in a Linear Mixed-Effect (LME) model. Furthermore, differences in predictor effects (due to e.g. traffic fleet compositions, housing characteristics or typical industrial influences) were evaluated per area, evaluating application of a random slope for predictors in the model.

Models were evaluated for model fit on all monitoring sites, as well as model performance on 83 independent home outdoor locations in Netherlands (N=42) and Basel (N=41), where repeated reference site corrected 24h outdoor measurements were available.

1) Introduction of Random Intercepts:

First, a single Linear Regression model was developed on all monitoring sites, and next difference in background UFP levels between study areas was evaluated in a LME model with a random intercept by area. This was performed by using predictors from the Linear Regression model and recalculating coefficients and significance levels, resulting in considerable changes in coefficients for INDUSTRY_5000 (-45%), TRAFLOAD_1000 (-31%) and HDLDRES_5000 (+49%). Significance levels did not exceed 0.10 after LME application, not leading to exclusion of predictors from LME model (see Table1).

Table S6; Predictor coefficients and significance levels in the regular Linear Regression model and the Linear Mixed-Effect Model with Random Intercept by area.

Regular linear regression model

	Intercept	TRAFNEAR	POPEEA_1000	INDUSTRY_5000	MAJORROADLENGTH_100	TRAFLOAD_1000
Coefficient	6281	0.231	0.035	2.622e-04	7.185	4.393e-06
P-value	1.879e-25	1.833e-29	0.031	3.859e-04	4.428e-06	1.882e-06

	HDLDRES_5000	POPEEA_300
Coefficient	4.608e-05	0.186
P-value	6.148e-03	0.063

Linear Mixed-Effect Model, introducing Random Intercept by area.

fixed effects

	Intercept	TRAFNEAR	POPEEA_1000	INDUSTRY_5000	MAJORROADLENGTH_100	TRAFLOAD_1000
Coefficient	6598	0.234	0.035	1.422e-04	7.221	3.049e-06
Pr (>ChiSq)		3.285e-31	0.040	0.091	4.329e-06	3.615e-03

	HDLDRES_5000	POPEEA_300
Coefficient	6.886e-05	0.165
Pr (>ChiSq)	5.308e-03	0.097

random effects by AREA

BAS	-173.9112
HER	-259.9689
NL	350.1003
NOR	-965.3472
SAB	1653.2410
TOR	-604.1141

When applying both models on independent sites, the Linear Regression model explained UFP variability (R^2) of 55.6% in NL, 49.6% in Basel, and 52.3% over pooled areas. The LME model with a random intercept predicted UFP variation of 56.4% in NL, 51.3% in Basel and 53.8% over pooled sets in UFP variation (Summary in Table 3). Based on these findings, LME is preferred over Linear Regression, since model performance increases when difference in background UFP levels by area are taken into account.

2) Evaluating Random Intercepts versus Random Intercepts and Random Slopes:

On top of previous findings, differences in predictor effects per area might add precision in UFP predictions per area. For this reason, alternately random slope for a single predictor was added to the LME model (Figures S1-S7 at the end of this document) and model fit on test sites was analyzed against the normal LME using an ANOVA.

Table S7; Analysis of model fit for the Linear Mixed-Effect Model with Random Intercept only against this model with an added random slope for 1 predictor.

	DF	AIC	BIC	LogLik	deviance	Chisq	Chi DF	Pr(>Chisq)	
Random Intercept only	10	2515	2564	-1247	2495				
Random Intercept + random slope for:									
TRAFNEAR	12	2503	2562	-1239	2479	15.9	2	0.00034	*
POPEEA_5000	12	2513	2572	-1244	2489	5.8	2	0.055	.
INDUSTRY_5000	12	2506	2565	-1241	2482	13.2	2	0.0014	*
MAJORROADLENGTH_100	12	2512	2571	-1244	2488	6.57	2	0.037	*
TRAFLOAD_1000	12	2518	2577	-1247	2494	0.61	2	0.74	
HDLDRES_5000	12	2520	2579	-1248	2496	0	2	1	
POPEEA_300	12	2519	2578	-1247	2495	0.27	2	0.87	

*=significance <0.05; .=significance <0.10

Table 2 shows that model fit was significant different on test sites when a random slope for TRAFNEAR, INDUSTRY_5000 or MAJORROADLENGTH_100 was added to the LME model. For POPEEA_5000, model fit increase was not significant.

When testing models described above, R^2 for measured against modeled UFP concentrations was determined in NL, Basel, and on pooled data. As presented in Table 3; model performance in external sites did not increase when models also had a random slope, next to a random intercepts. For slopes that gave a significantly better model fit in the test sites, performance in the external sites decreased 0.5% at random TRAFNEAR and 2.7% at random MAJORROADLENGTH_100. A drastic decrease in R^2 was observed when random INDUSTRY_5000 slopes was used.

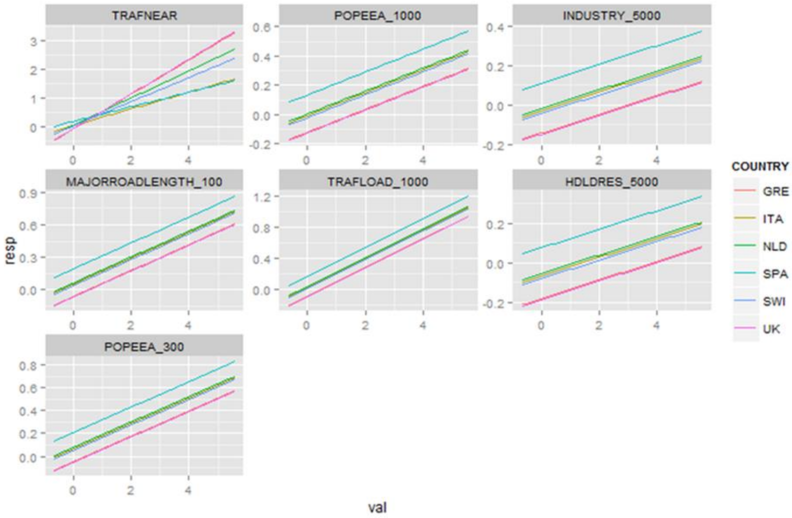
Other predictors did not show a better model fit in test data when applying a random slope; performance on external sites only increased 0.3% when a random slope for POPEEA_300 was applied.

TableS8; Model performance (R^2) for measured against modeled UFP levels at external sites from the Netherlands (NL, N=42), Basel (N=41) and in both areas pooled

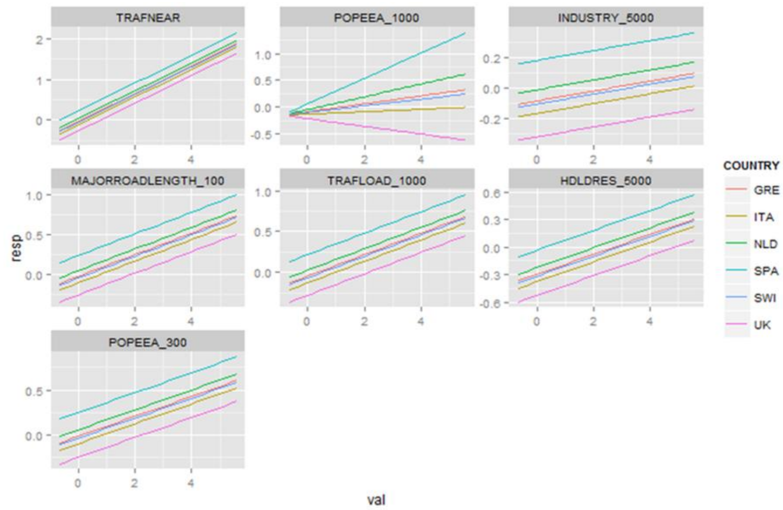
	R^2 (%) Sites NL	**R^2(%) Sites Basel**	**R^2(%) POOLED**
Linear Model	55.6	49.6	52.3
Random intercept only	56.4	51.3	0.53.8
Random intercept + random slope for:			
TRAFNEAR	55.8	52.3	53.3
POPEEA_1000	57.1	52.7	53.3
INDUSTRY_5000	31.8	9.3	6.8
MAJORROADLENGTH_100	57.6	50.7	51.1
TRAFLOAD_1000	56.6	51.1	53.7
HDLDRES_5000	57.2	49.7	53.0
POPEEA_300	56.7	51.6	54.1

Based on these findings, a LME model with random intercepts for area seems to be the best approach for predicting UFP variation at independent sites. The predictors where a significant better model fit was observed in the test sites when applying a random slope per area, did not provide a better UFP prediction in independent sites. To prevent overfitting of the model, a model that only accounts for background UFP differences is preferred, applying the LME with random intercepts per area for UFP predictions at a combined level.

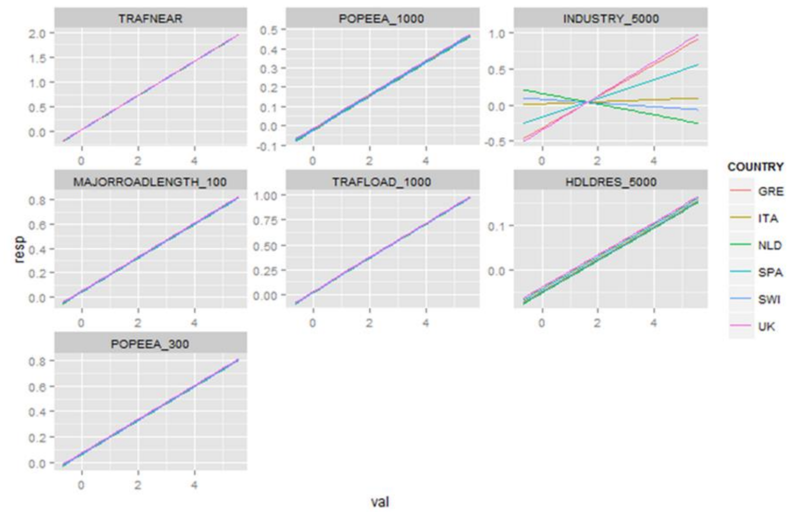
Figure S7; The effects for RANDOM SLOPES, while other predictors remain a fixed effect
A) TRAFNEAR



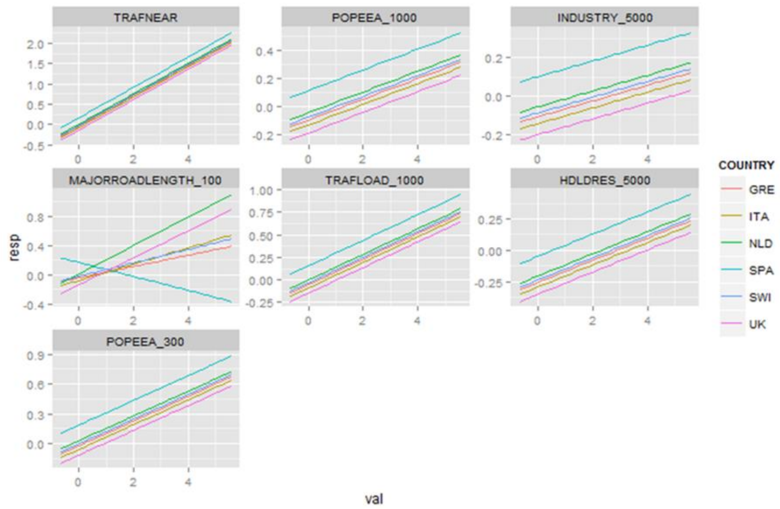
B) POPEEA_5000



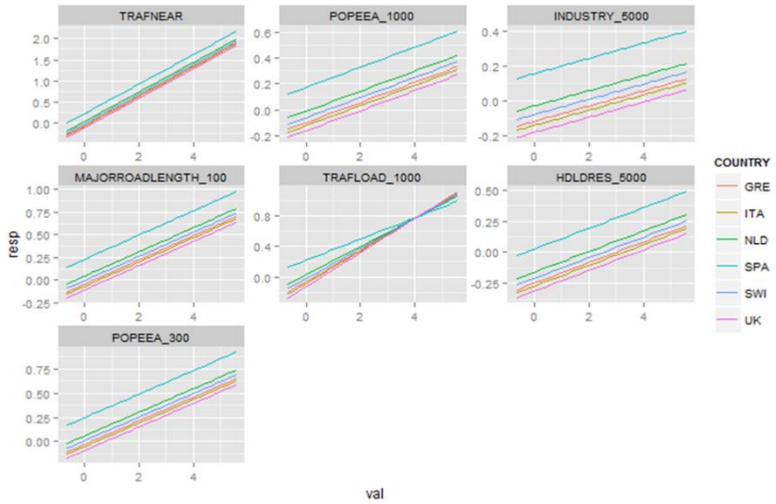
C) INDUSTRY_5000



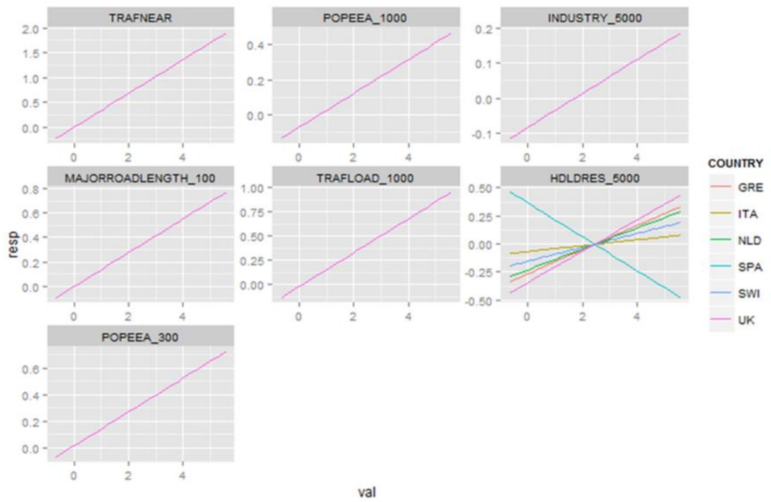
D) MAJORROADLENGTH_100



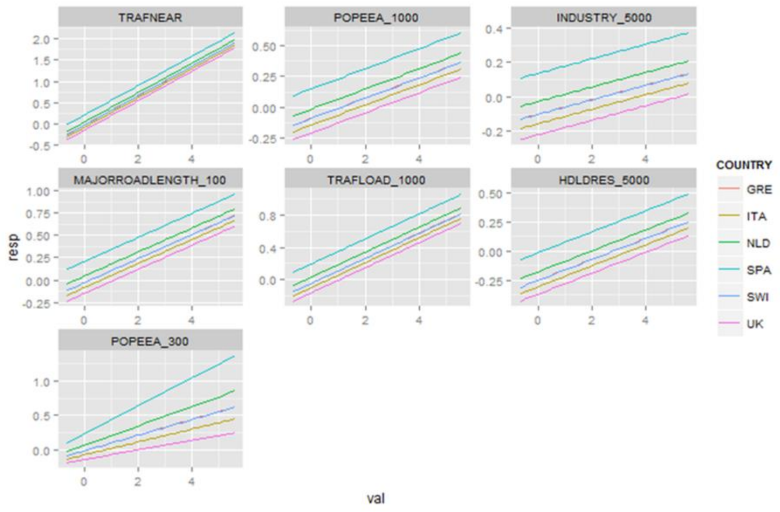
E) TRAFLOAD_1000



F) HDLDRES_5000



G) POPEEA_300


TRAFNEAR
POPEEA_1000
INDUSTRY_5000
MAJORROADLENGTH_100
TRAFLOAD_1000
HDLDRES_5000
POPEEA_300
resp
val
COUNTRY
GRE
ITA
NLD
SPA
SWI
UK

Supplement 9; Leave One Area Out LUR models

- Leave One Area Out (LOAO) models per area and were developed on all short-term monitoring sites, except one area. Model structures per excluded area are presented below;
- A predictor was used when there was at least 10% representation over monitoring sites (90^{th} percentile differed from 0). In addition for EU models, a predictor had to be represented in 3 or more (≥50%) of study areas;
- Predictor coefficients presented are multiplied by the spread in the specific predictor, calculated as the 90^{th} – 10^{th} percentile, expressing the proportional change in UFP for an increase between the 10^{th} and 90^{th} percentile of the predictor;
- Predictor categories are presented in the first column; (NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry).

Table S9; Model structures of models based on all short-term sites except one area

LOAO model Based on all short-term sites, except sites from one area		**UFP** (90th-10th percentile) **BASEL excluded**	**UFP** (90th-10th percentile) **HERAKLION excluded**	**UFP** (90th-10th percentile) **NETHERLANDS excluded**	**UFP** (90th-10th percentile) **NORWICH excluded**	**UFP** (90th-10th percentile) **SABADELL excluded**	**UFP** (90th-10th percentile) **TURIN excluded**
	Intercept	**7009**	**8506**	**4943**	**6946**	**7517**	**6018**
NT	Traffic intensity on nearest road	**4897*** (21895)	**4419** (22111)	**4943** (21735)	**4747** (21097)	**3876** (18881)	**4871** (17698)
NT	Traffic total load of roads in a buffer of 50m (sum of(traffic intensity * length of all segments))					**1941** (2.559e6)	
NT	Road length of all roads in a buffer of 50m						**1066** (201.6)
NT	Road length of all major roads in a buffer of 50m		**2139** (100.0)	**1350** (99.8)		**1382** (100.1)	
NT	Road length of all major roads in a buffer of 100m	**2155** (258.5)			**1961** (266.3)		**1428** (272.1)
DT	Traffic total load of roads in a buffer of 1000m (sum of(traffic intensity * length of all segments))	**1531** (5.107e8)	**1776** (5.149e8)	**610** (5.129e8)	**1430** (4.655e8)		
DT	Road length of all roads in a buffer of 1000m			**2448** (51635)			
PP	Population land use in a buffer of 300m			**1222** (5778)	**1492** (5512)	**1431** (5443)	
PP	Population land use in a buffer of 1000m	**2190** (48736)	**3938** (39697)				**2201** (43863)
PP	Population land use in a buffer of 5000m	**2482** (463857)				**3572** (463857)	**2026** (233063)
PP	Sum of low and high density residential land in a buffer of 1000m			**1164** (1.927e6)	**2413** (1.990e6)		
IN	Industry within a buffer of 300m		**148** (13574)				
IN	Industry within a buffer of 5000m	**1470** (8.056e6)		**2991** (7.906e6)	**994** (7.101e6)		**1353** (6.49e6)

Leave One Area Out (LOAO) models; developed on all short-term sites except the sites from the area listed in the column
NT = Nearby Traffic, DT = Distant Traffic, PP = Population, IN = Industry
* All presented coefficients are Model coefficients multiplied by the value of the 10^{th}-90^{th} percentile of the predictor