Supporting Information

for

Application of stochastic models in identification and apportionment of heavy metal pollution sources in the surface soils of a large-scale region

*Yuanan Hu, Hefa Cheng**

State Key Laboratory of Organic Geochemistry

Guangzhou Institute of Geochemistry, Chinese Academy of Sciences

Guangzhou 510640, China

Environmental Science and Technology

Date: March 10, 2013

Number of Pages: 10

Number of Table: 1

Number of Figures: 3

1. General Partitioning Procedure for Construction of the Conditional Inference Trees

Step 1 Finding the covariate X_j with the strongest association to *Y*. For the given case weight *W*, the relationship between the response variable *Y* and any of the *m* covariates $(X_1, X_2, ..., X_m)$ is measured by statistical significance tests. The global hypothesis of independence is formulated as $H_0 = \bigcap_{j=1}^{m} H_0^j$ and the partial hypothesis H_0^j is defined as $H_0^j : D(Y \mid X_j) = D(Y)$. The partitioning is stopped if the global hypotheses cannot be rejected at a pre-specified nominal level. Otherwise, the covariate X_{j*} with strongest association to *Y* is selected for further split of the node.

Step 2 Finding the optimal binary split using the selected covariate X_{j*}. For each possible binary split, the goodness of the partitioning is evaluated by two-sample linear statistics measuring the discrepancy between the two disjoint subgroups $\{\mathrm{Y}_i \mid w_i > 0 \wedge X_{ji} \in A; i = 1, ..., n\}$ and $\{\mathrm{Y}_i \mid w_i > 0 \wedge X_{ji} \notin A; i = 1, ..., n\}$. The two case weights W_{left} and W_{right} are determined according to the results of the statistics.

Step 3 Repeat steps 1 and 2 with updated case weights W_{left} and W_{right}.

2. General Principle of Random Forest

Given an original training data set ϕ_n of size *n*, random forest generates multiple new learning data sets $\phi_{n'}$ of size *n'* (*n'* < *n*) to build new trees, by sampling elements from the data set ϕ_n

uniformly and with replacement. Out-of-bag samples, which are observations not included in the bootstrap sample, are used for estimating the prediction accuracy of the trees and evaluating the variable importance. For each of the bootstrap samples, random forest builds classification or regression trees. Instead of selecting the best splitting variable by optimizing certain criterion such as the Gini gain, the individual tree in the random forest performs the recursive partitioning by choosing the variable for further split at each node randomly.

3. Spatial Distribution of Heavy Metals

Figures S2a-S2d show the spatial distribution maps of Cd, Cr, Hg, and As in the study region based on the traditional Kriging interpolation of the respective heavy metal concentrations in the surface soils from the sampling sites. These four metal species were selected as examples here because they exhibit distinctly different spatial distribution patterns. The highest levels of soil Cd (Figure S2a) were found in the area surrounding Zhaoqing, which hosts a range of metalliferous mining and smelting industries. The area between Guangzhou and Zhuhai, which covers Foshan, Zhongshan, and Jiangmen, also had elevated soil Cd levels. This area is highly industrialized and urbanized, and hosts many factories producing electronic products, clothes, and other consumer products. These observations indicate that the "hot spots" of soil Cd pollution were closely related to industrial activities in the PRD, and thus the soil Cd probably came mainly from anthropogenic sources. In contrast to Cd, the occurrence of soil Cr pollution (Figure S2b) appeared not to be strongly correlated with the industrial and urban centers, which suggests that the variation of natural background might be primarily responsible for the observed spatial distribution of Cr in the surface soils of the PRD. The sources of soil Hg and As were difficult to identify based on their spatial distribution patterns (Figure S2c and S2d), as soil Hg and As concentrations were elevated in some highly industrialized and urbanized areas, while some of the much less industrialized and urbanized areas also had elevated soil Hg and As concentrations. Overall, the spatial distribution patterns could reveal, even though only qualitatively, the possible sources for some heavy metals (e.g., Cd). However, for the heavy metals without distinct natural or anthropogenic input sources, little source information could be obtained from such geochemical maps. As a result, stochastic models are needed to identify the sources of heavy metals and to apportion the contributions from natural background and human activities to the observed heavy metal levels in the surface soils.

4. Table

Table S1. The fitted parameters for the log-normal distributions of the concentrations of Cd, Cu, Zn, Pb, Cr, Ni, As, and Hg in the surface soils of the PRD and the goodness of fit obtained from FMDM analysis.

Metal	First component (background distribution)			Second component (polluted distribution)			Df*	Chisq#	*p*$
	Weight	Mean	Standard deviation	Weight	Mean	Standard deviation			
Cd	0.54	0.11	0.07	0.46	0.30	0.24	45	48.00	0.35
Cu	0.45	18.04	15.50	0.55	36.02	21.09	45	44.93	0.47
Zn	0.46	78.92	35.88	0.54	116.43	109.64	44	35.50	0.82
Pb	0.88	46.40	32.15	0.12	58.45	5.31	45	40.79	0.65
Cr	0.46	58.02	70.87	0.54	67.73	29.52	45	35.90	0.83
Ni	0.42	20.07	43.04	0.58	23.54	13.19	69	61.42	0.73
As	0.88	11.88	9.13	0.12	30.76	9.57	45	31.92	0.93
Hg	0.95	0.05	0.04	0.05	0.21	0.02	45	34.34	0.88

Notes:

* — degrees of freedom of the fitted mixture model;

# — the chi-squared goodness-of-fit statistic;

\$ — the significance level (*p*-value) for the test with the null hypothesis H_0: the estimated model is consistent with the observed distribution. Reject H_0 if $p \leq 0.1$.

5. Figures

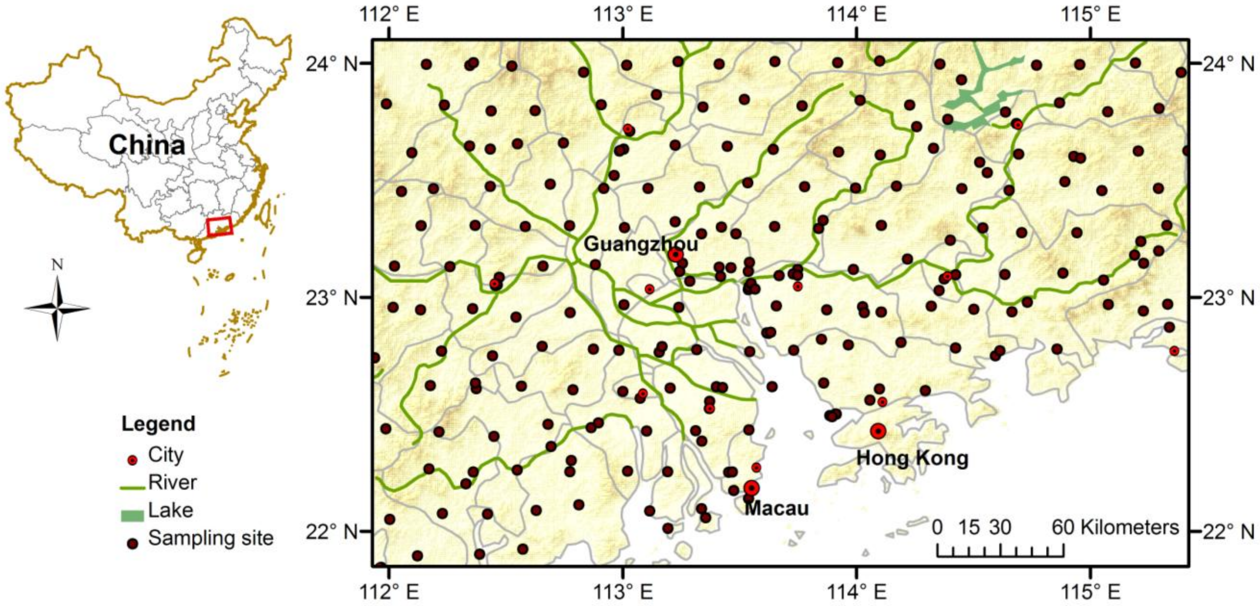
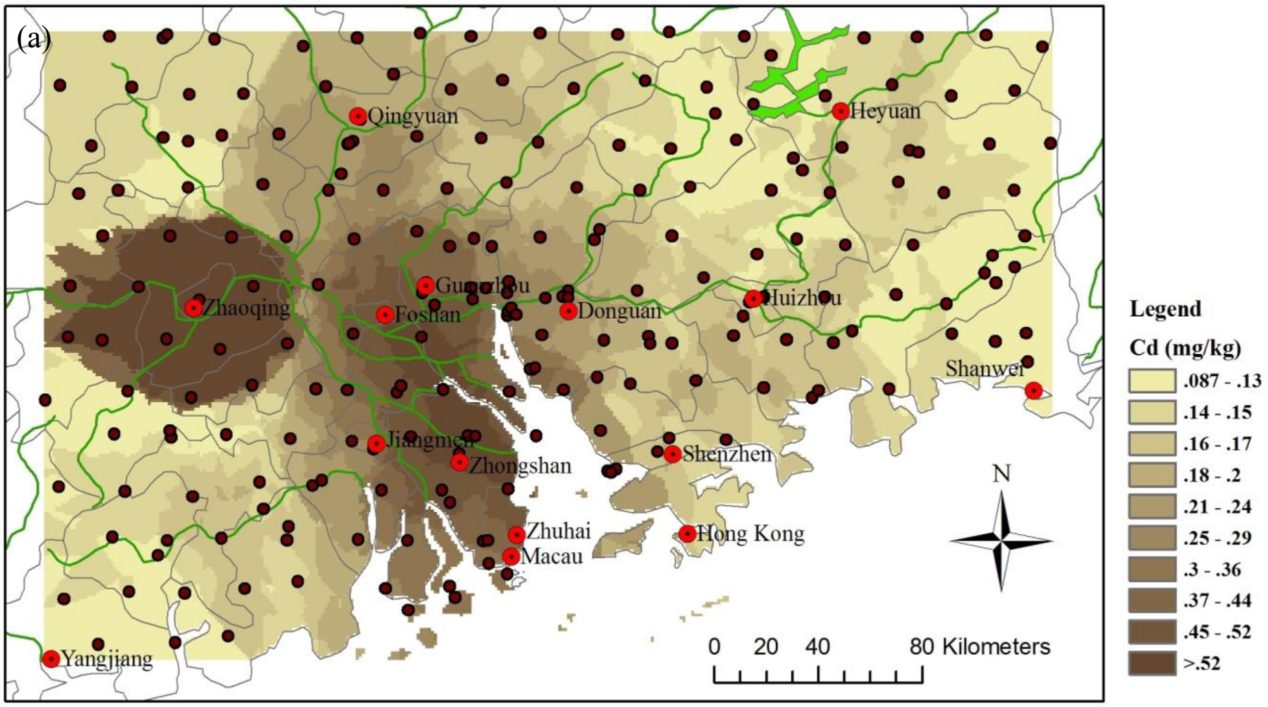
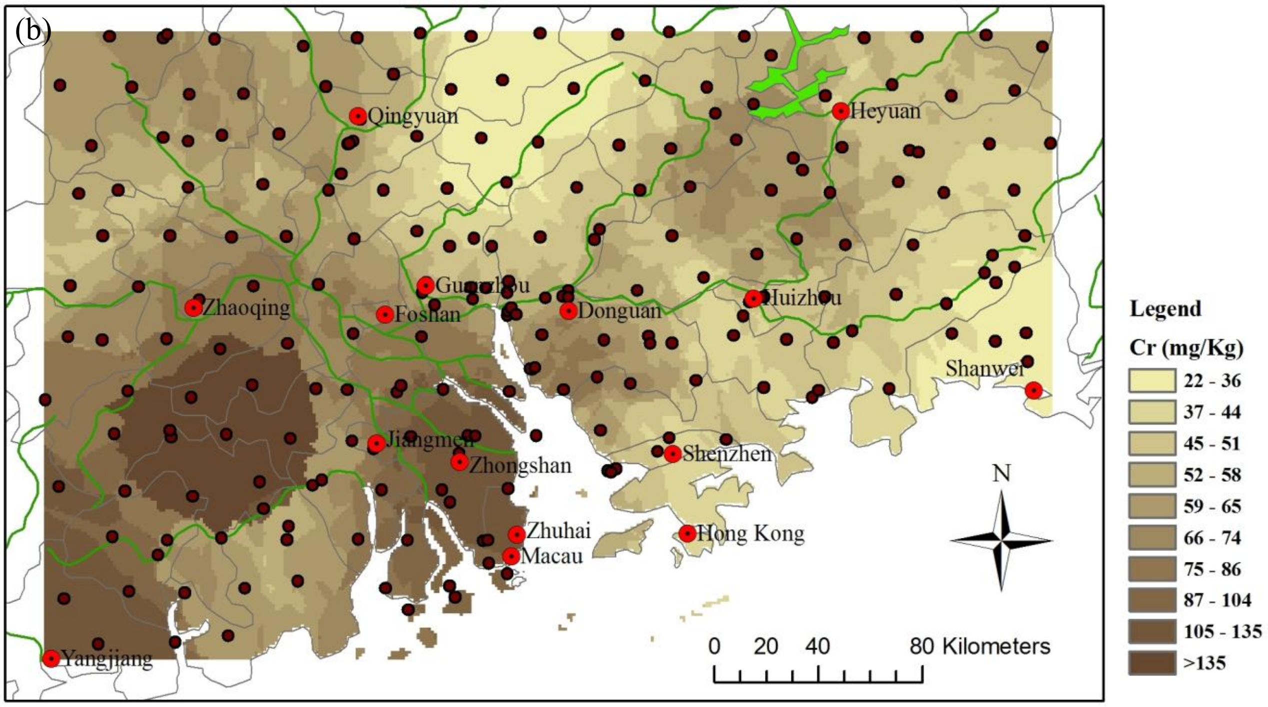


Figure S1. Map of the Pearl River Delta (PRD) in China and the locations of soil sampling sites. All samples were collected at least 30 m away from the main road to minimize the direct influence of vehicular emissions (1).


(a)
Qingyuan
Heyuan
Zhaoqing
Guangzhou
Foshan
Donguan
Huizhou
Shanwei
Jiangmen
Zhongshan
Shenzhen
Zhuhai
Macau
Hong Kong
Yangjiang
N
0 20 40 80 Kilometers
Legend
Cd (mg/kg)
.087 - .13
.14 - .15
.16 - .17
.18 - .2
.21 - .24
.25 - .29
.3 - .36
.37 - .44
.45 - .52
>.52


(b)
Qingyuan
Heyuan
Zhaoqing
Guangzhou
Foshan
Donguan
Huizhou
Shanwei
Jiangmen
Zhongshan
Shenzhen
Zhuhai
Macau
Hong Kong
Yangjiang
N
0 20 40 80 Kilometers
Legend
Cr (mg/Kg)
22 - 36
37 - 44
45 - 51
52 - 58
59 - 65
66 - 74
75 - 86
87 - 104
105 - 135
>135

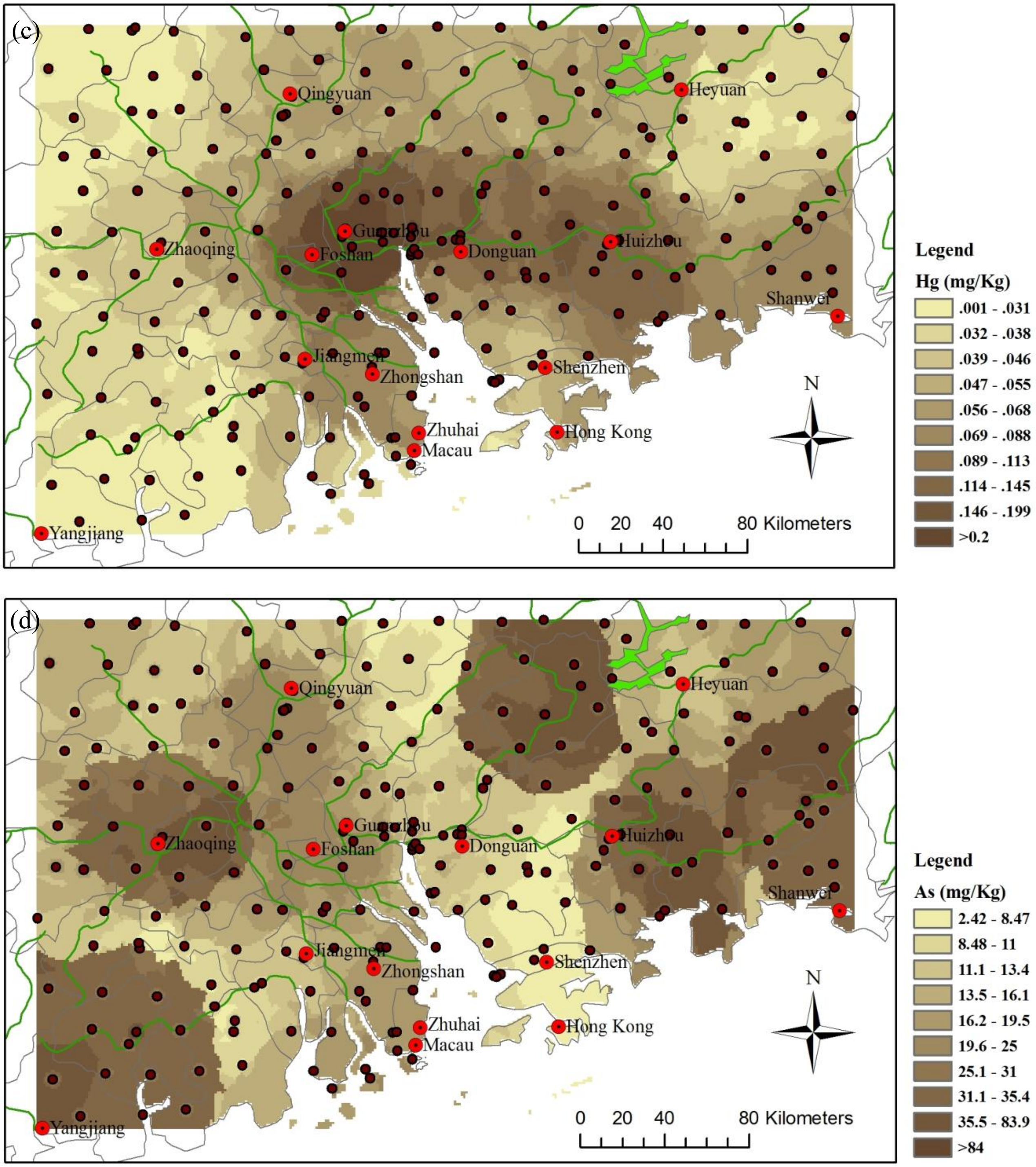


Figure S2. The spatial distribution maps of heavy metals (a) Cd, (b) Cr, (c) Hg, and (d) As in the PRD.

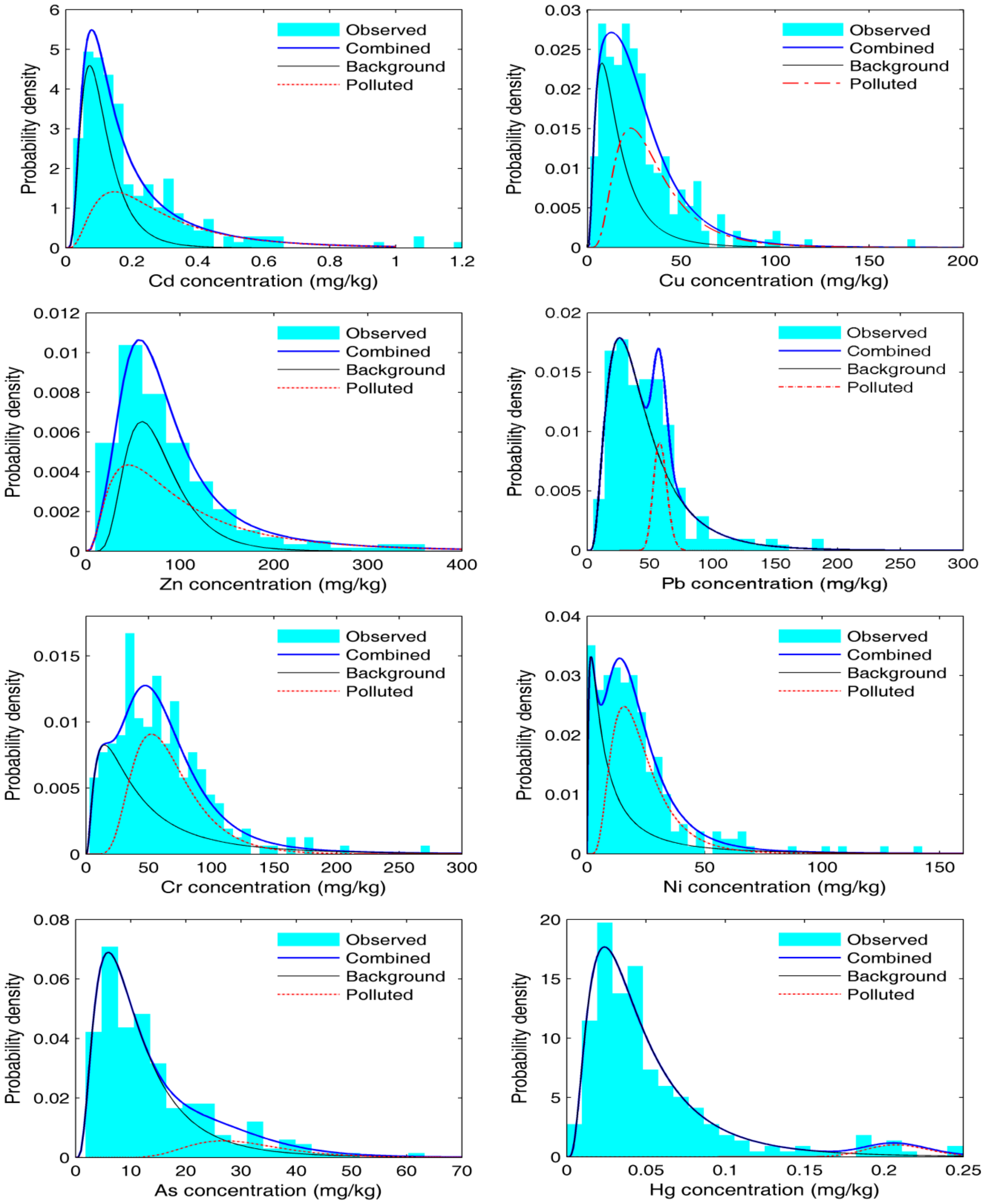


Figure S3. FMDM model fits for the concentrations of heavy metals (Cd, Cu, Zn, Pb, Cr, Ni, As, and Hg) in the surface soils of the PRD.

Literature Cited

1. Smith, W. H. Lead contamination of the roadside ecosystem. *J. Air Pollut. Control Assoc.* **1976,** *26* (8), 753-766.