Environmental Science Processes & Impacts

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

rsc.li/process-impacts

Cluster Analysis of Passive Air Sampling Data Based on the Relative Composition of Persistent Organic Pollutants

XIANDE LIU[1,*], FRANK WANIA[2]

[1] *State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China*

[2] *Department of Physical and Environmental Sciences, University of Toronto Scarborough, 1265 Military Trail, Toronto, Ontario, Canada M1C 1A4*

* Corresponding author, E-mail: liuxdlxd@hotmail.com

Abstract

The development of passive air samplers has allowed the measurement of time-integrated concentrations of persistent organic pollutants (POPs) within spatial networks on a variety of scales. Cluster analysis of POP composition may enhance the interpretation of such spatial data. Several methodological aspects of the application of cluster analysis are discussed, including the influence of a dominant pollutant, the role of PAS duplication, and comparison of regional studies. Relying on data from six regional studies in North and South America, Africa, and Asia, we illustrate here how cluster analysis can be used to extract information and gain insights into POPs sources and atmospheric transport contributions. Cluster analysis allows classification of PAS samples into those with significant local source contributions and those that represent regional fingerprints. Local emissions, atmospheric transport, and seasonal cycles are identified as being among the major factors determining variation in POP composition at many sites. By complementing cluster analysis with meteorological data such as air mass back-trajectories, terrain, as well as geographical and social-economic aspects, a comprehensive picture of the atmospheric contamination of a region with POPs emerges.

Environmental Impact Statement

Cluster analyses of data on POP composition obtained by passive air sampling networks reveal that distance between sites alone is not a good predictor for similarity in POP composition. Other factors, most notably proximity to, and seasonally variable strength of, local sources, have an important influence on POP composition and therefore clustering. On the other hand, the observation of clusters of sites with very similar POP composition is an indication of the existence of compositional profiles that are not unduly influenced by local sources and may thus be interpreted as regional or even (trans-) continental POP fingerprints, in particular if the sites within such a cluster are geographically far apart.

Introduction

While passive air sampling techniques have been applied to the measurement of volatile organic compounds (VOCs) and inorganic gaseous pollutants for many years, their use for persistent organic pollutants (POPs) only became common in the last decade, largely driven by the demand of implementing the Stockholm Convention on POPs signed in 2001.[1-3] Passive air samplers (PASs) made it possible to routinely obtain POP air concentration data from remote and harsh environments where lack of electricity and other logistical limitations make measurements relying on pumps very difficult. Moreover, as a cost-effective device, PASs can be deployed at a large number of sites covering areas of different scales.[4-6] PAS data at local, regional or global scales are now increasingly used to investigate the concentration levels, temporal and spatial variations and trends, atmospheric transport processes, source contributions and fate of POPs.[4-11] PASs have been deployed along urban-rural,[12,13] altitudinal[14-16] and latitudinal transects,[17] as well as a vertical gradient.[18] Based on well designed sampling networks, local sources and long range atmospheric transport contributions have been assessed.[6,19,20]

PAS data typically consist of concentrations of a set of POPs at a number of sites that cover an area such as a river delta, a country, or even an entire continent. Such data sets have two dimensions, with POPs as variables and samples obtained at specific sites as objects, and are thus suited for the application of multivariate analytical methods such as principal component analysis (PCA) and cluster analysis.[21,22] There have been several studies using those multivariable methods, but the number of applications has been limited.[12,23,24,25]

This work will discuss several aspects of applying cluster analysis to POP data obtained by PAS. In the first part we will address methodological issues, including a comparison of the results obtained using either concentration data or compositional data as input to the cluster analysis, how to deal with dominant compounds, and the role of replicate samples. In the second part we discuss in detail what information can be gained from the application of cluster analysis to POP data obtained by PASs. This includes the identification of local and regional source contributions, information on atmospheric POP transport and seasonal variations in POPs sources and atmospheric mixing.

Methods

Cluster analysis is a multivariate procedure that allows detection of groups of similar objects, based on their closeness in an *n*-dimensional space of variables. It establishes the relationship among objects (here samples) in an objective and quantitative way. In most cases the similarity among objects is defined as the Euclidian distance between the objects. While many clustering strategies have been developed, here we employ agglomerative hierarchical clustering. Its results are presented as a dendrogram, which aids visual evaluation and data interpretation by showing the progressive grouping of samples. Each sample unambiguously falls into one of the groups.[26] Samples are merged step by step; Ward's error sum method merges two clusters for which the resulting increase in the sum of the square distances of each element to the centroid of the cluster is minimized. This least square technique has been recommended in many clustering applications.[26]

During cluster analysis, the user is allowed to decide how many groups or clusters are deemed most suitable. The dendrogram usually includes a horizontal dashed line indicating the dissimilarity level where no further grouping is advised and where, therefore, the number of clusters is decided. When this line is moved up and down slightly along the similarity axis, the number of groups should not change rapidly, indicating a stable situation. Ultimately, the number of clusters should be driven by the possibility to give a reasonable interpretation for the grouping.

Using the XLSTAT software (Addinsoft, France),[26] we applied cluster analysis to several regional PAS data sets that have been published before (Table 1). We refer to earlier papers for details on the analytical methods and the quality control and quality assurance measures. Due to space limitations only a dendrogram, a compositional plot and major findings can be presented in the main paper. Additional information, including a short description of each study, a list of chemicals quantified, a map with the sampling sites, a table with the average POP composition for each identified groups, and a discussion of the characteristics of each group is available in the Electronic Supporting Information (ESI).

Data from PAS networks are usually reported in units of sequestered amount of a specific compound per sampler. In order to obtain concentration levels that can be compared with concentration data obtained by other means, efforts are often made to convert them to volumetric concentration data, e.g. in the unit of $ng \cdot m^{-3}$, using sampling rates.[2] Although sampling rates are compound and site-specific,[7,27] the differences to the compositional data that would be introduced through site and compound-specific sampling rates are generally too small to change the outcome of cluster analysis. Because the purpose of clustering was to find out the similarity among samples, the sequestered amount of a compound in ng per sampler, normalized to a common sampling period, was used as a measure of absolute concentration levels. Most cluster analyses were performed on compositional data, which were obtained from the concentration data through normalization. After selecting the POPs of interest, we summed up their levels in a sample and divided the level of each POP by this sum to get a fraction. Missing values, i.e. non-detects or levels reported as below the detection limit, are replaced with half of the MDL (Method Detection Limit).

While we use data obtained by XAD-resin based PAS here, the approach should also work for data obtained by other PAS. Whether data from different PAS can be used in the same cluster analysis may depend on whether they display different uptake characteristics, e.g.

with respect to the uptake of particulate phase POPs.[27]

Results and Discussion

Methodological Aspects of Cluster Analysis on Passive Air Sampling Data

We first use the Chengdu-Wolong Nature Reserve (WNR) regional study[16] as an example to explore several methodological aspects. Duplicate PAS were deployed for six-month intervals during ‘summer’ and ‘winter’ periods at seven sampling sites in WNR and one in Chengdu. The sample name consists of three parts. [16] First, a capital letter S or W indicates the sampling period (‘summer’ or ‘winter’). Second, one or two capital letter(s) describe the site; i.e. CD stands for Chengdu, N for sites in WNR which are ordered in increasing altitudes, e.g. N1 for Gengda school, N2 for Panda Center, N3 for Sandaoqiao, N4 for Dengsheng, N5 for Beimuping, N6 for 95 kilometer milestone, and N7 for Pass. Third, a lower-case letter (a or b) stands for one of the duplicates at a site. For instance, W-CDa is the first duplicate sample from Chengdu deployed during winter. For more details we refer to Section 1 in the ESI.

Using absolute or relative concentration data for cluster analysis

Analytical data of POPs are generally presented as concentrations and their absolute magnitude and seasonal and spatial variation have primary importance. However, when several POPs are measured, their relative composition can also be informative. During atmospheric transport from source regions to receptor areas POP concentrations decrease due to dispersion; concentration levels can also be altered substantially by weather conditions such as strong winds or inversion layer formation. Yet, the POP composition changes only slightly in these cases. Due to their different chemical and physical properties, the composition of POPs may change slightly during atmospheric transport. However, because POPs are relatively persistent in the atmosphere, the compositional similarity during transport should be maintained to a certain extent. For example, in Tianjin, a big city in Northern China, POP air concentrations changed by more than one order of magnitude between industrial, urban and rural sites, but the composition was fairly similar.[13, 23] In brief, the POP composition is reflective of an emission source profile and can serve as a kind of fingerprint of a POP mixture in the air of either a site or an extended source area. In fact, source profiles are generally presented in the form of relative composition.[28, 29]

Figures 1 and 2 present the results of cluster analysis based on either concentration or compositional data from the Chengdu-WNR study.[16] When using absolute concentration data, the calculated Euclidian distances can adopt numerical values in the hundreds and the ‘dissimilarity’ can be as high as 30000 (Fig. 1). Fractional composition data yield Euclidian distances less than 1 and dissimilarity levels of about 0.3 or less (Fig. 2).

Being located in a source area, Chengdu had elevated POPs concentrations in both summer and winter (Fig. 1). POP composition varied seasonally, with HCB higher in winter and HCHs and DDTs higher in summer (see ‘Summer- urban’ and ‘Winter-urban’ groups in Fig. 2). Levels were much lower in WNR, with sites falling into two ‘mixture’ groups (Fig. 1). The group ‘mixture, mainly summer samples’ consists of 13 samples with the lowest POPs levels, whereas the group ‘mixture, mainly winter samples’ comprises 11 samples including the summer Gengda duplicates (S-N1a, S-N1b) and all four samples from the Panda Center (W-N2a, W-N2b, S-N2a, S-N2b), which had slightly higher POP levels.

It may come as a surprise that winter samples from Gengda constitute a small group of their own, being assigned the label ‘Winter-windy’. ‘Summer-urban’ and ‘Winter-urban’ groups are very similar in terms of total POP levels. However, ‘Summer-urban’ first merges with ‘Winter-windy’ and then with ‘Winter-urban’ at a higher dissimilarity level (Fig. 1). Irrespective of season, samples from Gengda (N1) had always higher POPs levels than other WNR sites. However, POP composition at Gengda (N1) is consistent with those of most other WNR sites (see Fig. 2). The Gengda site was situated at a very windy location and the relatively high POPs levels at this site were attributed to the effect of wind on the kinetics of uptake in the XAD-PAS.[16,30] This wind effect was also observed for particularly windy sites in Western Canada[14] and on Changdao Islands, China [24].

It is worthwhile to note that the groups ‘Winter-urban’ and ‘Summer-urban’ have widely different appearances in Figures 1 and 2. Due to their unique composition, the ‘Summer-urban’ samples appear as an independent group in Figure 2, whereas the ‘Winter-urban’ group no longer exists. The Chengdu samples (W-CD) become part of a big ‘winter’ group with a number of samples from WNR, indicative of effective regional atmospheric mixing in winter. Even though summer samples fell into three different groups, these groups do merge at higher dissimilarity levels in Figure 2. Atmospheric transport processes are occurring during the summer monsoon period; meanwhile, confounding factors such as primary emission, secondary re-evaporation, and degradation processes are enhanced in warm conditions. This results in a much weaker influence of regional atmospheric transport and mixing in summer.

Summer samples from Sandaoqiao (S-N3) and Panda Center (S-N2) were not recognized as independent groups in Figure 1. However, their composition was marked by unusually high HCHs enriched in β-HCH. This is apparent in Figure 1, but much more obvious in Figure 2, where Sandaoqiao (S-N3) is identified as an independent group. As the most persistent isomer, β-HCH is enriched during aging; as the most water-soluble isomer, it is easily scavenged from the air. High β-HCH prevalence is thus indicative of local pollution. In the case of Sandaoqiao (N3) and Panda Center (N2) it was due to local soil pollution.[31] It seems the source emission profile of a site is more clearly identified when clustering is based on compositional data.

We believe it is worthwhile to perform clustering on both compositional and concentration data. Since during primary data interpretation attention is typically focused on absolute POP concentrations, we highlight here the benefit of also paying attention to their composition. Compositional data may provide additional information, sometimes allowing for new insights.

Cluster analyses with or without a dominant compound

The identity and number of target analytes of a study depends upon its purpose and scope. Frequently, one or more compounds dominate the relative composition, overwhelming the others. In the Chengdu-WNR study, HCB was the dominant compound,[16] accounting for 71.1% on average. HCB's fraction ranged from 54% to 84%; such large differences mean that HCB has significant weight in Euclidean distance calculations. The question arises, whether clustering results are solely determined by HCB and whether results would be different if the analysis is done without HCB data. Therefore, the same compositional data set, but without HCB, was used for another cluster analysis (Fig. 3). Similar to Figure 2 there are five groups. However, this time winter samples are better separated from summer samples. The 'mixed' group no longer exists; two Panda Center summer samples (S-N2a, S-N2b) are now identified as an independent group and two winter samples (W-N7, W-N6b) appear as a small sub-group in the group of 'all winter samples'. All of the points discussed in the previous section are holding true. In fact Figure 3 fully supports and confirms the major features in Figure 2.

Dominant compounds are measured at high levels and their concentrations therefore are generally more accurate. While it is reasonable and acceptable that dominant compounds should have more weight in a cluster analysis, it is worthwhile to examine how results change if a dominant compound is excluded. Performing the cluster analysis with and without a dominant compound ought to lead to slightly different, yet consistent results. For example, a study on the Tibetan Plateau identified groups labeled 'long range atmospheric transport' and 'urban' when all data were used, when HCB was excluded, an additional 'agricultural' group was separated.[25]

The role of replicate samples in cluster analysis

In the above cluster analyses, most of the fourteen duplicates samples behaved as would be expected, i.e., appeared side by side in the dendrogram or at least fell into the same group (Figs. 1 to 3). So what is the point of using individual samples instead of the average of replicates? Are they redundant information making things unnecessarily more complicated? Replication is primarily a quality control measure, ensuring the quality of a sampling and quantification procedure. Differences between replicates quantify the uncertainty associated with this procedure. POPs in environmental matrices are measured at trace levels. The lower the levels are, the greater the uncertainties will be. When performing cluster analyses on PAS data, it is of crucial importance that the data set reflects first and foremost real differences between sites and not the random error of the analytical procedure, which is inevitable and always present in analytical data. It is one of the basic requirements that analytical uncertainties are smaller than the differences between sites. To confirm that this requirement is met, we bring duplicates into the cluster analysis whenever possible. By comparing the similarity levels of duplicates with that of the groups, duplicates in cluster analysis can function as an ‘internal reference’.

For acceptable clustering results, duplicates must behave properly. If one or two pairs of duplicates fail to be in the same group, it might be tolerable. However, it should prompt a scrutinizing of original data and the analytical procedure. Above, most duplicates fell into the same group (Fig. 1-3); the only exception was the 95 km milestone site in winter (W-N6) in Fig. 2. If more than two pairs of duplicates or more than 20 % of the duplicates in a study fail to be in the same group, it may indicate that analytical uncertainties impacted the clustering and the results are no longer acceptable.

Because dominant compounds are present at relatively high concentration levels, their data generally have fewer uncertainties than those at lower levels. Caution should be taken when performing a cluster analysis without one or more of the dominant compounds, as one may run the risk of losing analytical quality of the data set and get poor clustering results that are no longer acceptable.

Applying Cluster Analysis to Passive Air Sampling Data

Distinguishing sites with local source influence from those representing regional background conditions

One of the main contributions cluster analysis can make to the interpretation of PAS network data for POPs is to separate sites influenced by local sources from those that are not. If several sites display strong similarity and are thus grouped into a cluster, this suggest firstly a lack of local source influence, and secondly the existence of a regional POP fingerprint. The Botswana study (Section 2 in the ESI) illustrates how cluster analysis can aid in identifying the influence of local sources. HCB was evenly distributed in this study. The larger the percentage of HCB at a site, the cleaner it is. Six sites in the Okavango Delta are in reasonably close vicinity (Fig. S2) and have a similar POP fingerprint with the highest HCB fraction. The remaining nine sites are influenced, to various extents, by local emissions, with compositions dominated by one or more POPs (Fig. 4, Table S1).

The Chilean study,[15] where 16 of the 20 sites are judged to reflect regional background conditions with limited local source influence, is particularly intriguing (Section 3 in the ESI). Even though the three transects, each consisting of sites in close geographical proximity, are far apart, site proximity is not the major driver of clustering, i.e., we do not see clusters of South, Central, and North. Instead clustering reveals three distinct fingerprints: one reflecting the clean Southern Pacific air mass, one reflecting that Pacific air mass plus some regional input, and one that is typical of Chilean urban areas (Fig. 5, Table S2). Importantly, all of these groups include sites from more than one of the three transects, even though they were very far apart (Fig. 6). Sites fall into different groups not merely based on latitude, but also based on their location downwind from regional sources, which in Chile to some extent may be related to longitude and altitude, those being further East/inland/higher having potentially more regional source influence.

This idea of the existence of common POP background fingerprints over very large spatial scales, as is apparent in the Chilean data, is reinforced in the analysis of the North American network (Section 4 in the ESI). Sites that are very far apart geographically can be clustered together, because of their similar fingerprints. 31 of the North America sites fall into three broad regional groups that can be regarded as reflecting regional background conditions, while the remaining nine sites, falling in groups with average concentrations of the sum of quantified POPs exceeding 100 ng/PAS, have significant local contributions (Fig. 7, Table S3). The sites in regional groups are not necessarily in close proximity. An example is the ‘waterside group’ (Fig. 7, Table S3), which comprises sites from the east and west coast of North America as well as locations near the Great Lakes, yet they have highly uniform compositions. It implies that representative compositional fingerprints exist on a continental scale. This is even true on the global scale: Within the GAPS network, a cluster representing North American/European sites without local source influence is apparent (Section 5, Figs. S4 and S5, Tables S5 and S6).

One could argue that a regional PAS network would be an intelligent way of making an informed decision on the sitting of one or more long-term POP air monitoring stations. Such stations could for example serve in the evaluation of the effectiveness of the Stockholm Convention in a region by recording the existence and rate of a decline in air concentrations over time. If the purpose of that monitoring is therefore to determine the regional situation rather than a local one, the cluster analysis could not only tell how many background stations would be required (one for each regional group), but also which sites would be good candidates because of the lack of local interference.

Revealing seasonal differences in POPs sources and transport

Both POP concentrations in air and their compositions can vary seasonally, as is apparent in Chengdu-WNR (Figs. 1-3). For example, even though the total amount of POPs at the Chengdu site was quite similar in winter and summer, the concentration of each of the POPs varied substantially. Seasonality may be caused by variations both in source intensity and in meteorological variables such as temperature, wind and precipitation.

Cluster analysis can aid in revealing seasonality of POPs levels, which is again illustrated using the Chengdu-WNR study. The 'winter' group contains thirteen winter samples from both remote WNR and urban Chengdu which are about 100 km away from each other (Fig. 2, Fig. S1). These samples had similar POP composition suggesting that they are in the same regional airshed, presumably as a consequence of effective atmospheric mixing. Since the prevailing wind situates the Chengdu plain upwind of the WNR, a regional fingerprint is consistent with the meteorological conditions.[16] Low temperatures suppress local sources in Chengdu during winter and the urban site shows the same regional fingerprint as the remote sites. The winter time cluster in WNR-Chengdu is interesting because neither site proximity nor site type (urban/remote), but season is a major driver of clustering.

In the warm summer period evaporation and degradation processes become enhanced. Most summer samples fall into three groups (Fig. 2), showing compositional differences and a much weaker manifestation of the regional airshed. When temperatures go up in summer, primary and secondary emissions become stronger and the Chengdu-WNR regional airshed breaks apart; during the cooler winter, emission strengths decline and the regional airshed merges again. In this annual cycle both seasonal variations and spatial differences vary in parallel. In summer, the Chengdu samples form a 'summer urban' group by themselves, with six POPs having maximum mass fractions (Table 2). Also the absolute concentrations were the highest in Chengdu during summer, about four times those in WNR (Fig. 1). The 'summer-local contamination' group consists of duplicate samples taken at Sandaoqiao (S-N3). Large mass fractions of α- and β-HCH at this site reflect local soil pollution with HCHs. This secondary local evaporation source is also more efficient in summer.[16,31,32] The 'summer' group contains nine samples from WNR, implying that during summer the east-facing slope remained a well mixed airshed with homogenous composition and that the soil pollution at Sandaoqiao had only limited local influence. Compared with the thirteen samples in the 'winter' group, the compositional similarity of the thirteen summer samples is lower by a factor of 6; as the three summer groups merge at a much higher dissimilarity level (Fig. 2).

Recognising influence of regional and long range atmospheric transport

PAS provide time-integrated concentrations, typically with a time resolution between three and twelve months. While it is no feasible to follow dynamic atmospheric transport processes with such data, it may be possible to investigate the consequences of atmospheric transport phenomena occurring during such long time periods. Cluster analysis of data from Western Canada identified three groups (Fig. S6, Table S7 in Section 6 in the ESI). Interestingly, it is not the sites belonging to a transect that are grouped together, but the high altitude sites from all three transects are assigned to a single group (Fig. 8, 9). In other words, instead of site proximity, elevation is the key characteristics for clustering.

The clustering results make intuitive sense, because altitude is correlated with distance from local sources that tend to be at lower elevations. Observation Peak is in a very remote area with virtually no permanent residents. The only humans present in the area are tourists driving on the Icefield Parkway during summer. The three lowest sites in the Yoho transect form a separate group; they are close to the Trans-Canada Highway, where there is year-round traffic (including trucks and trains) and some smaller settlements. The lower part of the Revelstoke transect, again grouped separately, is clearly influenced by the town at its base and again by the train and highway corridor. It is also reasonable that the higher elevations of the three transects are similar and truly reflect “clean” air masses not influenced by local sources.

A recent study in the Southern Alps on the South Island of New Zealand differentiated regional from long-range atmospheric transport, whereby the former did not extend beyond the mountain divide at about 700 m a.s.l.[28] Based on an analysis of relative compositional data it was concluded that ‘neither easterly nor westerly local-scale upslope winds resulted in the transport of pesticides to the other side of the mountain divide’ and also that ‘pesticide profiles at the highest mountain sites were mainly influenced by atmospheric transport via northwesterly synoptic-scale air flow’.[28] Adopting this line of reasoning, the formation of the R- and Y-groups is caused by local conditions and that of the ‘O plus’-group is due to samplers being exposed more to synoptic-scale air flow at high altitudes.

It is interesting to note that HCHs and HCB in soil samples were much lower along the Observation Peak transect than in those from the Revelstoke and Yoho transects (see Table S7 in ref.[15]). How then could HCHs and HCB in air be higher in the O-plus group than in the R- and Y-groups? It may be explained by more exposure to atmospheric air masses transported over long distances, possibly from the Pacific. This is also consistent with a relatively high α/γ-HCH ratio of 4.7.[33] The α/γ-HCH ratio showed a significant increasing trend from low to high latitudes on a cruise across the North Pacific Ocean.[34]

In this study the importance of site elevations was clearly revealed and two kinds of atmospheric transport, regional (valley) and synoptic flow (long-range), could be distinguished. There are other similar examples. In the Chilean study two regional groups were identified which are influenced by regional and long-range atmospheric transport contributions, respectively.[15] In a Tibetan Plateau study, four high altitude sites were designated as ‘long range atmospheric transport’ group as those sites had compositions very similar to those from the Indian sub-continent, namely were enriched in HCHs and endosulfans.[9,25]

Confirming source-receptor relationships

It would be desirable if a source-receptor relationship could be established directly based on a PAS network that includes sites in both source areas and receptor regions. The regional studies in both Chengdu-WNR and Tianjin-Changdao fulfill this premise. In the latter study 50 PAS were deployed for five consecutive three-month intervals during the 15 months between March 2007 and May 2008. There were ten monitoring sites, six in the city of Tianjin and four on the Changdao Islands. Several previous studies have established Tianjin as a major source area for organochlorine pesticides such as HCB, HCHs, and DDTs in Northern China.[13,35,36,37] The prevailing airflow situates Changdao ca. 400 kilometers downwind of Tianjin. Here we use the consensus on the existence of a source-receptor relationship between Tianjin and Changdao to find out what POP compositions are expected to look like at PAS monitoring sites in two regions linked as source and receptor areas.

Cluster analysis on the PAS compositional data identified four groups largely based on local source contributions and season (Fig. 10). Whereas two source groups, containing a total of 23 samples, reflect emission patterns in the area, the other two are characteristic of regional POP compositions (Table 3). The ‘winter’ group had the lowest POP levels, the ‘spring & autumn’ group had levels a factor of 2 higher, while the ‘DDXs-source’ and ‘HCHs-source’ groups had levels a factor of 4 and 9 higher, respectively.

The ‘HCHs-sources’ group contains ten samples with very high HCHs fractions, exclusively from Tianjin, particularly from Tanggu (T1). Summer samples from five Tianjin sites (all except Hangu, T2) are in this group, reflective of a seasonal and also regional signature. Of the 13 samples with high DDXs fractions in the ‘DDXs-sources’ group three are from Hangu (T2) in Tianjin, where there had been large scale DDT production in the past. More surprisingly, 10 are from sites on Changdao, particularly those at the meteorological (C2) and communication stations (C3). The reason is that fishing boats had been regularly treated with DDT-containing paint in spring and summer at the local harbors.[24, 38] Summer samples did not form a seasonal group, but instead fell into one of the two source groups. In other words, summer samples appear to be better indicators of source emission profiles.

The ’winter’ group, with a composition high in HCB, PCB28, and PCB52 and low in other POPs, contained only winter samples, four from Changdao and five from Tianjin. Winter samples from all sites except Tanggu (T1) fall into this group. The ‘spring/autumn’ group consisted exclusively of spring and autumn samples either from rural Tianjin (T6: Yuqiao and T5: Baodi) or from the sites on Changdao with the least local POPs (C1: county monitoring station, C4: county museum). Even though Tianjin and Changdao are 400 km apart, many sites had a highly similar POP composition in seasons other than summer, which can be seen as evidence of regional atmospheric mixing driven by the monsoon from continental Asia.[24]

The α-HCH/ γ-HCH ratio was 2.32 at the industrial Tanggu site (T1) in Tianjin, and 2.57 as mean value of four sites on Changdao, a slight increase indicating the effect of degradation during the atmospheric transport towards Changdao. Interpreting the *p,p'*-DDE/*p,p'*-DDT ratio is a bit complicated, as DDTs were also emitted on Changdao Island. The *p,p′*-DDE/*p,p′*-DDT ratio was 0.84 at the industrial Hangu site (T2) in Tianjin, and 1.14 at the monitoring station site (C1) at the northwestern tip of Changdao, upwind of the other three sites in Changdao (Fig. S7). The increasing relative abundance of DDE is consistent with a more degraded DDT signature. At the other sites in Changdao the *p,p′*-DDE/*p,p′*-DDT ratio was low (0.32, 0.43, and 0.68 for C2-meteorological station, C3-communication station, and C4-county museum, respectively) and well correlated with distance from where DDT-containing paints had been used, confirming the recent use of DDT.

To establish a link between source area and receptor sites, three conditions have to be met: a POPs emission source exists presently or in the past; the regional atmospheric circulation and prevailing wind connect the two regions; whereas POPs concentration levels are lower in the receptor area, the composition is similar in the two regions. Related to the latter, diagnostic ratios, such as α-HCH/γ-HCH and *p,p′*-DDE/*p,p′*-DDT, should also make sense, namely should either be uniform or see a slight increase in the relative abundance of the substance with the longer atmospheric residence time. Tianjin and Changdao fulfill these three conditions.

Comparison of several regional studies

Comparison between two or more dendrograms from different studies is not possible. Very often cluster analyses may be based on different sets of compounds, because the number and identity of compounds differs between studies. Therefore the calculated Euclidean distances are not directly comparable. We can, however, compare PAS samples from different studies by first making a synthetic data set that includes all PAS samples and then performing cluster analysis on the set of compounds that was measured in all studies. An example of this kind of cluster analysis is presented in Section 8 in ESI. The synthetic data set includes data for 10 compounds quantified in three regional studies (Western Canadian Mountains, Chile, Botswana). Most of the sites from Western Canada fall into a single group, while sites in Botswana fall into two groups, one clean and one polluted (Fig. 11). The 20 Chilean sites fall into five groups, which is not surprising considering that they form three altitudinal transects in the northern, central, and southern part of the country that are very far apart. Nevertheless, polluted sites are well separated from seven clean sites on the southern and central transects. Notably, in the Western Canadian Mountain study all 17 sites could be regarded as regional sites, i.e. the entire network experienced very limited local source influence.

Conclusion

Although we performed cluster analysis mostly on compositional POP data here, absolute concentration levels are equally important and carrying out clustering on both concentrations and compositions can provide additional insights into POPs sources and atmospheric transport. Similarly, it is worthwhile to perform cluster analyses with and without a dominant compound, thereby confirming or enhancing the reliability of the clustering results. PAS replication functions as a quality control measure in cluster analyses by providing an ‘internal reference’, against which the significance of differences between clusters can be judged.

Cluster analysis generally separates PAS samples into those with significant local source contributions and those that are regionally representative. This simple picture provides the basic framework for further interpretation. For data with time resolution of six months or less, cluster analysis reveals interactions between seasonal variations and site differences, which are closely related to factors such as emissions, secondary re-evaporation and atmospheric transport. Cluster analysis can structure the data sets from monitoring networks with a large number of sites in a way that facilitates further detailed in-depth investigations. In a comprehensive interpretation of clustering results the chemical information, i.e. the POP concentrations and compositions, needs to be complemented by other data, including meteorology (e.g. air mass back-trajectories), terrain, geographical and social-economical aspects.

Acknowledgements This study was supported by the National Basic Research Program of China (no. 2003CB415003) and by the Chinese Research Academy of Environmental Sciences (no. 2009KYYW13); FW acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

References

1. F. A. Esteve-Turrillas, A. Pastor, M. de la Guar, Passive sampling of atmospheric organic contaminants, Chapter 1.11 in book, Comprehensive Sampling and Sample Preparation. 2012 Elsevier Inc. (doi:10.1016/B978-0-12-381373-2.10011-0)
2. F. Wania, L. Shen, Y. D. Lei, C. Teixeira and D. C. G. Muir, *Environ. Sci. Technol.* 2003, 37, 1352-1359.
3. M. Shoeib and T. Harner, *Environ. Sci. Technol.* 2002, 36, 4142-4151.
4. L. Shen, F. Wania, Y. D. Lei, C. Teixeira, D. C. G. Muir and T. F. Bidleman, *Environ. Sci. Technol.* 2004, 38, 965-975.
5. C. Shunthirasingham, C. E. Oyiliagu, X. S. Cao, T. Gouin, F. Wania, S. C. Lee, K. Pozo, T. Harner and D. C. G. Muir, *J. Environ. Monitoring* 2010, 12, 1650-1657.
6. A. K. Halse, M. Schlabach, A. Sweetman, K. C. Jones and K. Breivik, *J. Environ. Monitoring* 2012, 14, 2580-2590.
7. C. Shunthirasingham, B. T. Mmereki, W. Masamba, C. E. Oyiliagu, Y. D. Lei and F. Wania. *Environ. Sci. Technol.* 2010, 44, 8082-8088.
8. J. N. Hogarh, N. Seike, Y. Kobara and S. Masunaga, *Environ. Sci. Technol.*, 2012, 46, 2600-2606

9. X. P. Wang, P. Gong, T. D. Yao and K. C. Jones, *Environ. Sci. Technol.,* 2010, 44(8): 2988–2993.
10. G. L. Daly, Y. D. Lei, C. Teixeira, D. C. G. Muir, L. E. Castillo, L. M. M. Jantunen and F. Wania. *Environ. Sci. Technol.* 2007, 41, 1124-1130.
11. G. L. Daly, Y. D. Lei, C. Teixeira, D. C. G. Muir, L. E. Castillo and F. Wania. *Environ. Sci. Technol.* 2007, 41, 1118-1123.
12. A. Motelay-Massei, T. Harner, M. Shoeib, M. Diamond, G. Stern and B. Rosenberg, *Environ. Sci. Technol.*, 2005, 39 (15), pp 5763–5773
13. X. Y. Zheng, D. Z. Chen, X. D. Liu, Q. F. Zhou, Y. Liu, W. Yang and G. B. Jiang, *Chemosphere* 2010, 78, 92-98
14. G. L. Daly, Y. D. Lei, C. Teixeira, D. C. G. Muir and F. Wania. *Environ. Sci. Technol.* 2007, 41, 6020-6025.
15. C. Shunthirasingham, R. Barra, G. Mendoza, M. Montory, C.E. Oyiliagu, Y.D. Lei and F. Wania, *Atmos. Environ.* 2011, 45, 303-309.
16. W. J. Liu, D. Z. Chen, X. D. Liu, X. Y. Zheng, W. Yang, J. N. Westgate and F. Wania, *Environ. Sci. Technol.,* 2010, 44, 1559-1565
17. L. Shen, F. Wania, Y. D. Lei, C. Teixeira, D. C. G. Muir and T. F. Bidleman, *Environ. Sci. Technol.* 2005, 39, 409-420.
18. Y. M. Li, Q. H. Zhang, D. S. Ji, T. Wang, Y. W. Wang, P. Wang, L. Ding and G. B. Jiang, *Environ. Sci. Technol.*, 2009, 43, 1030-1035
19. Y. M. Li, D. W. Geng, F. B. Liu, T. Wang, P. Wang, Q. H. Zhang and G. B. Jiang, *Atmos. Environ.*, 2012, 51, 140-145
20. P. Barthel, S. Thuens, C. Shunthirasingham, J. N. Westgate, F. Wania and M. Radke, *Environ. Pollut.* 2012, 166, 218-215
21. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman, Chemometrics: A text book. Volume 2 in Data Handling in Science and Technology. Elsevier, 1988.
22. D. L. Massart and L. Kaufman, The interpretation of analytical chemical data by the use of cluster analysis. Wiley, New York. 1983.
23. X. D. Liu, D. Z. Chen, X. Y. Zheng, W. Yang, Y. Liu and G. B. Jiang, *J. Chinese Mass Spec. Society*, 2011a, 32 (2): 65-70 (in Chinese with English abstract)
24. X. D. Liu, X. Y. Zheng, D. Z. Chen, Z. M. Hu, W. Yang and G. B. Jiang, *Res. Environ. Sci.*， 2011b, 24(9)， 967-974 (in Chinese with English abstract)
25. P. Gong, X. P. Wang, J. J. Sheng and B. Q. Xu, *Res. Environ. Sci.,* 2013, 26, 350-356. (in Chinese with English abstract)
26. XLSTAT software, Addinsoft, France. http://www.xlstat.com
27. X. M. Zhang and F. Wania, *Environ. Sci. Technol.* 2012, 46, 9563-9570.
28. K. S. Lavin and K. J. Hageman, *Environ. Sci. Technol.* 2013, 47: 1390-1398

29. Receptor Model Source Composition Library, EPA-450/4-85-002
30. X. M. Zhang, T. N. Brown, A. Ansari, B. Yeun, K. Kitaoka, A. Kondo, Y. D. Lei and F. Wania, *Environ. Sci. Technol.* 2013, 47, 7868-7875
31. X. Y. Zheng, X. D. Liu, W. J. Liu, G. B. Jiang and R. Q. Yang, *Chinese Sci. Bulletin* 2009, 54(5):743-751
32. D. Z. Chen, W. J. Liu, X. D. Liu, J. N. Westgate and F. Wania, *Environ. Sci. Technol.* 2008, 42, 9086-9091.
33. T. Harner, M. Shoeib, M. Kozma, F. A. P. C. Gobas and S. M. Li, *Environ. Sci. Technol.* 2005, 39: 724-731
34. X. Ding, X. M. Wang, Z. Q. Xie, C. H. Xiang, B. X. Mai, L. G. Sun, M. Zheng, G. Y. Sheng and J. M. Fu, *Environ. Sci. Technol.,* 2007, 41(15): 5204-5209
35. S. Tao, W. X. Liu, Y. Li, Y. Yang, Q. Zuo, B. A. Li and J. Cao, *Environ. Sci. Technol.*, 2008, 42, 8395–8400
36. F. M. Jaward, G. Zhang, J. J. Nam, A. J. Sweetman, J. P. Obbard, Y. Kobara and K. C. Jones, *Environ. Sci. Technol.* 2005, 39, 8638-8645.
37. X. Liu, G. Zhang, J. Li, L. L. Yu, Y. Xu, X. D. Li, Y. Kobara and K. C. Jones, *Environ. Sci. Technol.,* 2009, 43, 1316-1321
38. J. Li, G. Zhang, L. L. Guo, W. H. Xu, X. D. Li, C. S. L. Lee, A. J. Ding and T. Wang, *Atmos. Envir.* 2007, 41, 3889-3903

Table 1 Selected PAS network data sets reported in previous publications that were the subject of cluster analyses in this work

Data set	Type of study	Number of chemicals	Length of deployment	Year of study	No. of PAS sites/ No. of samples	Reference
Western Canadian mountains	elevation gradients	20	12 months	2003- 2004	17/17	14
Chile	elevation gradients	15	12 months	2006-2007	20/20	15
North America	continental	16	12 months	2000-2001	40/40	4, 17
Botswana	national	13	12 months	2006-2007	15/15	7
Chengdu-WNR, China	regional	11	6 months	2007-2008	8/30	16
Tianjin-Changdao, China	regional	11	3 months	2007-2008	10/50	13, 23, 24
Combined data set of 3 studies	synthetic data set	10	12 months		52/52	7, 14, 15
GAPS (Global atmospheric passive sampling)	global	11	12 months	2007-2008	32/32 for 2007 33/33 for 2008	5

Table 2 Average composition of the groups obtained in the cluster analysis of normalized compositional data from Chengdu-WNR. Maximum values are shown in bold font.

POPs	summer	summer-local contamination	mixed	summer-urban	winter
α-HCH	0.222	**0.249**	0.113	0.111	0.091
β-HCH	0.012	**0.087**	0.054	0.019	0.021
γ-HCH	0.059	0.080	0.043	**0.095**	0.039
HCB	0.649	0.550	0.719	0.539	**0.802**
p,p′-DDE	0.016	0.010	0.013	**0.074**	0.010
o,p′-DDT	0.032	0.013	0.012	**0.096**	0.006
p,p′-DDT	0.005	0.001	0.003	**0.028**	0.004
PCB28	0.004	0.000	0.023	**0.024**	0.014
PCB52	0.001	0.000	0.007	**0.007**	0.005

Table 3 Average composition of the groups obtained in the cluster analysis of normalized compositional data from Tianjin-Changdao. Maximum values are shown in bold font.

	HCHs-sources	Spring & autumn	DDXs-sources	Winter
Number of samples	10	18	13	9
α-HCH	**0.390**	0.215	0.153	0.095
β-HCH	**0.043**	0.020	0.034	0.003
γ-HCH	**0.146**	0.089	0.061	0.025
δ-HCH	**0.036**	0.009	0.009	0.000
HCB	0.301	0.525	0.300	**0.727**
p,p′-DDE	0.032	0.048	**0.098**	0.025
p,p′-DDD	0.001	0.002	**0.013**	0.005
o,p′-DDT	0.021	0.030	**0.131**	0.019
p,p′-DDT	0.015	0.032	**0.189**	0.029
PCB28	0.012	0.023	0.009	**0.060**
PCB52	0.003	0.007	0.003	**0.013**

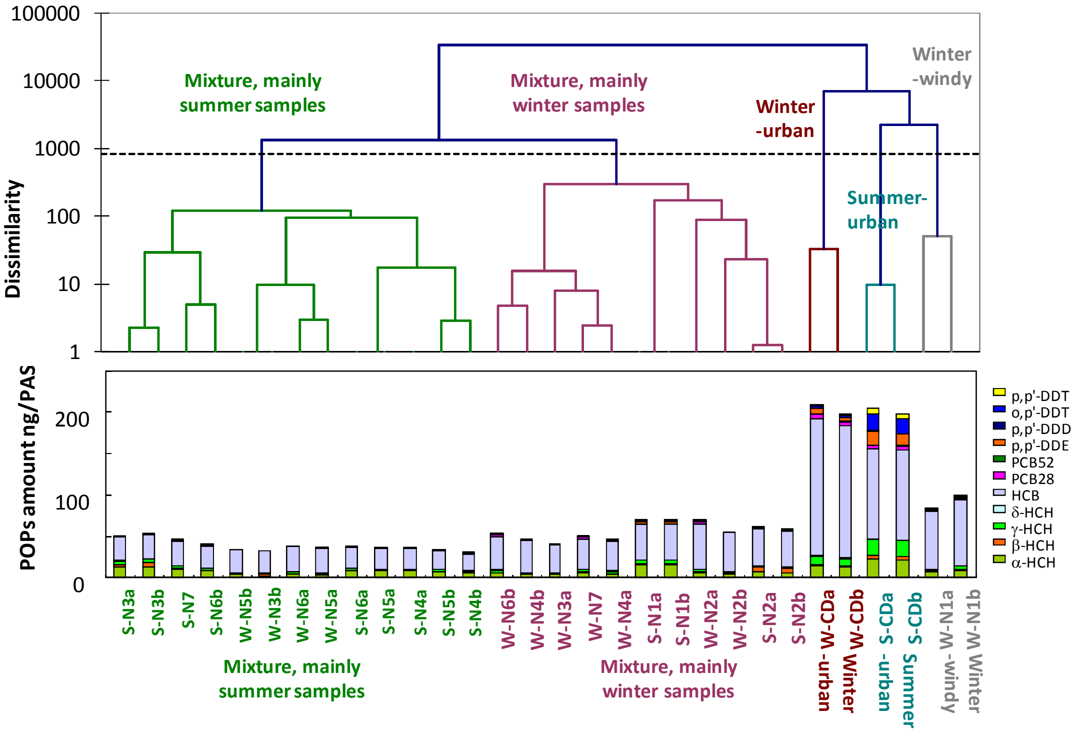


Figure 1 *Dendrogram (top, log scaled) and composition plot (bottom) for the cluster analysis based on absolute concentration data from the Chengdu-WNR study. The cluster analysis identified five groups: two big groups with low POPs levels (one group contains 13 samples, most of them summer samples; another contains 11 samples, most of them winter samples) and three small groups, each of them comprising duplicates from a single site ('Winter urban' and 'Summer-urban' groups at Chengdu; 'Winter-windy' group of Gengda winter samples).*

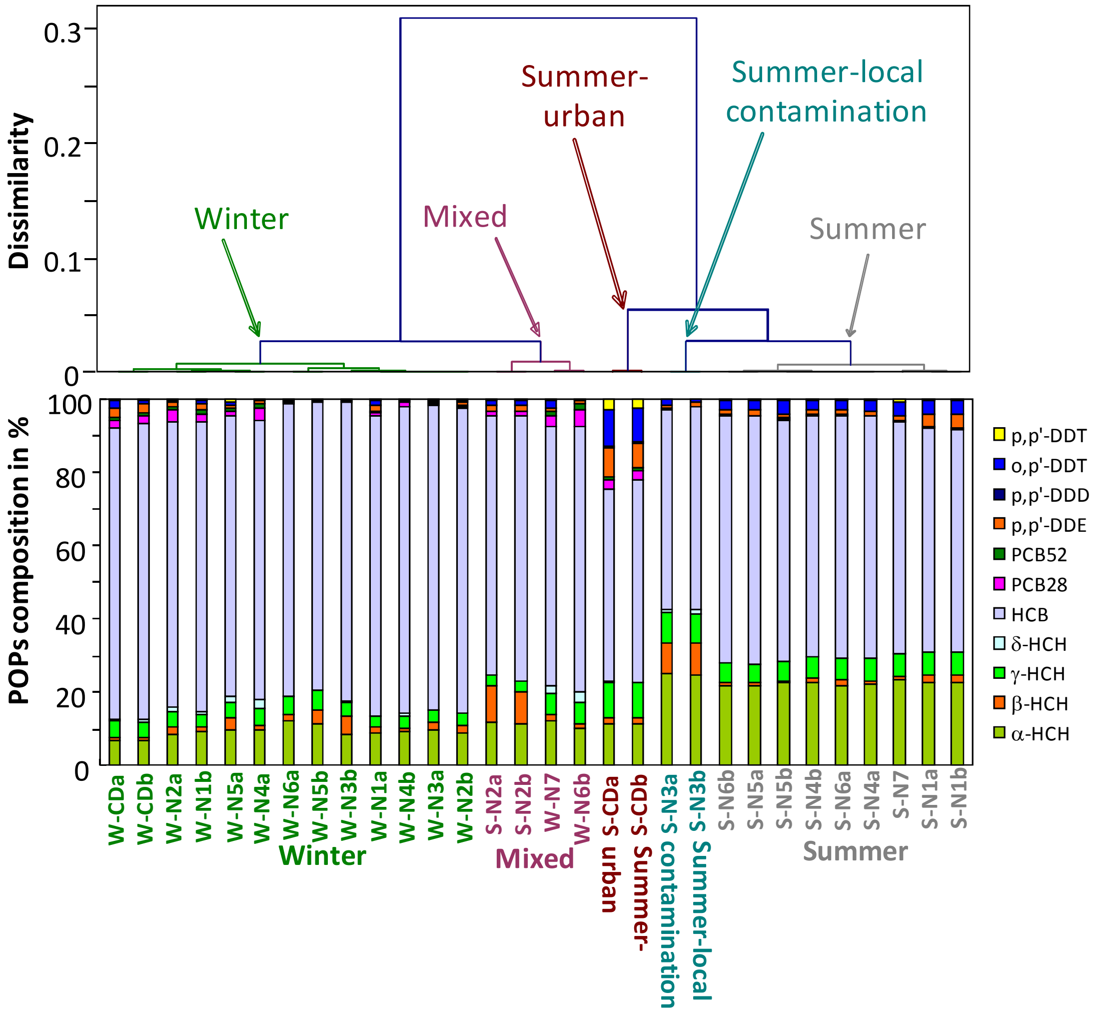


Figure 2 *Dendrogram (top) and composition plot (bottom) for the cluster analysis based on compositional data from the Chengdu-WNR study. The cluster analysis identified five groups. The winter group contains 13 winter samples. Three summer groups have 2, 2, 9 summer samples, respectively. In the middle, a 'mixed' group consists of two winter samples and two summer samples.*

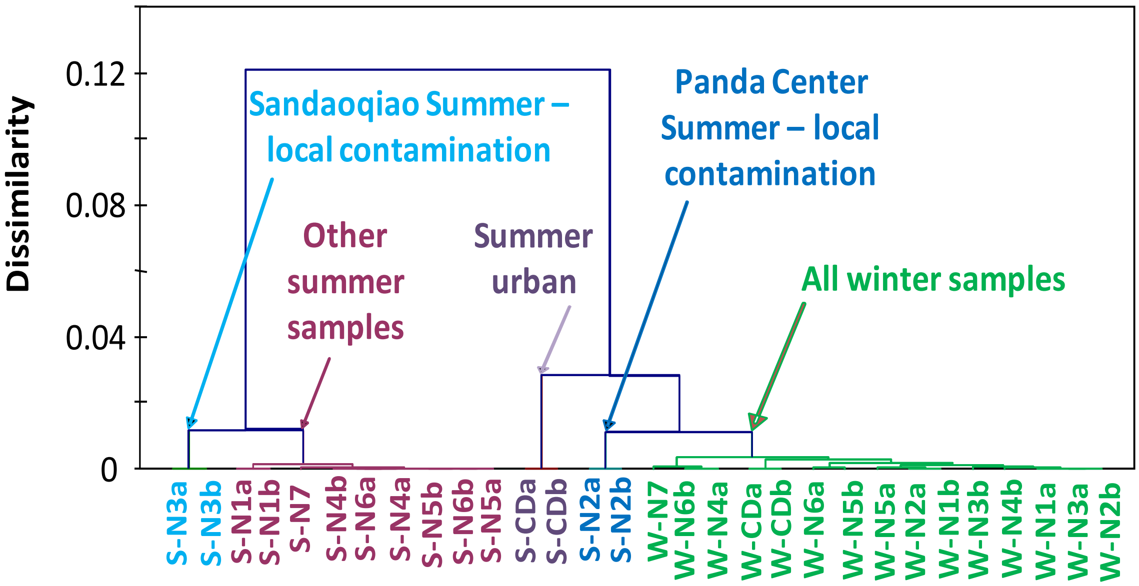
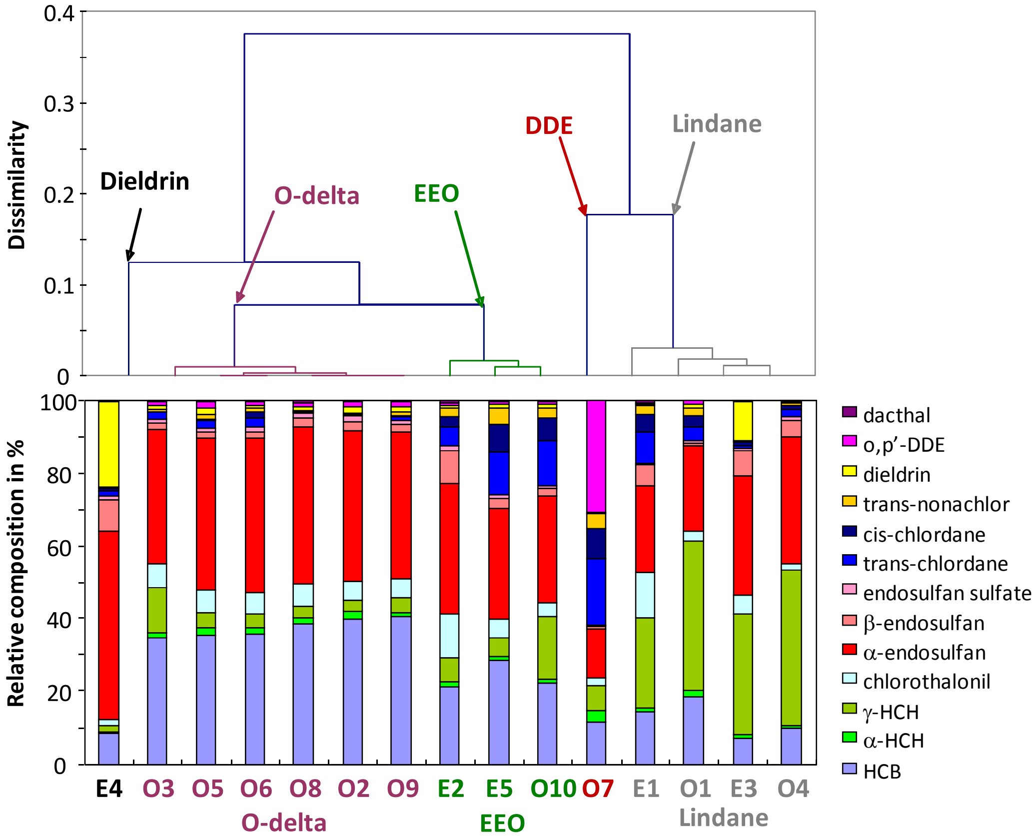


Figure 3 Dendrogram for the cluster analysis based on compositional data from the Chengdu-WNR study, with HCB data excluded. The analysis identified five groups. The big winter group contains all 15 winter samples. The four remaining groups consist of 2, 9, 2, and 2 summer samples, respectively.



559

560 *Figure 4* *Dendrogram and composition plot for Botswana study. The analysis identified*
561 *five groups. The six sites of the O-delta group were relatively clean while the*
562 *other nine sites were influenced by local sources.*

563

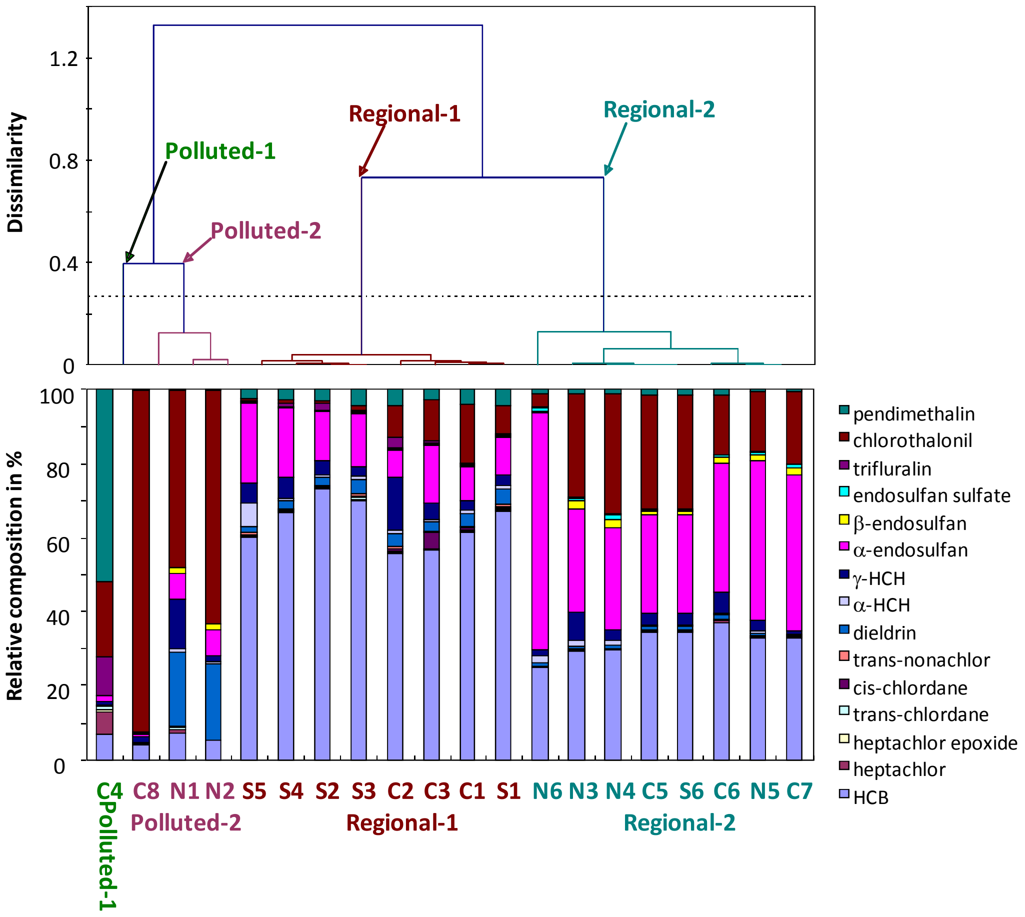


Figure 5 *Dendrogram and composition plot for the Chilean study. The analysis identified four groups. Sixteen sites in two regional groups were relatively clean, whereas the remaining four sites were influenced by local sources.*

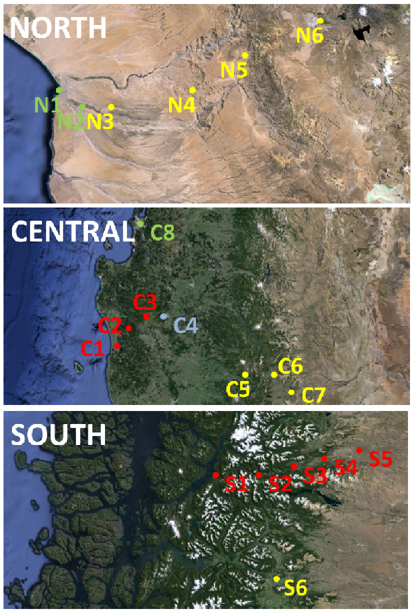


Figure 6 Map showing the passive air sampling sites in Northern, Central and Southern Chile [15]. Cluster analysis reveals that one cluster (red, including sites from two transects) reflect clean Southern Pacific air, whereas another (yellow, including site from all three transects) have higher proportions of endosulfan and chlorothalonil. The sites in green are in urban areas (Arica, Concepcion) dominated by chlorothalonil. The site in blue is in an agricultural area and has an unusual mix of pesticides.

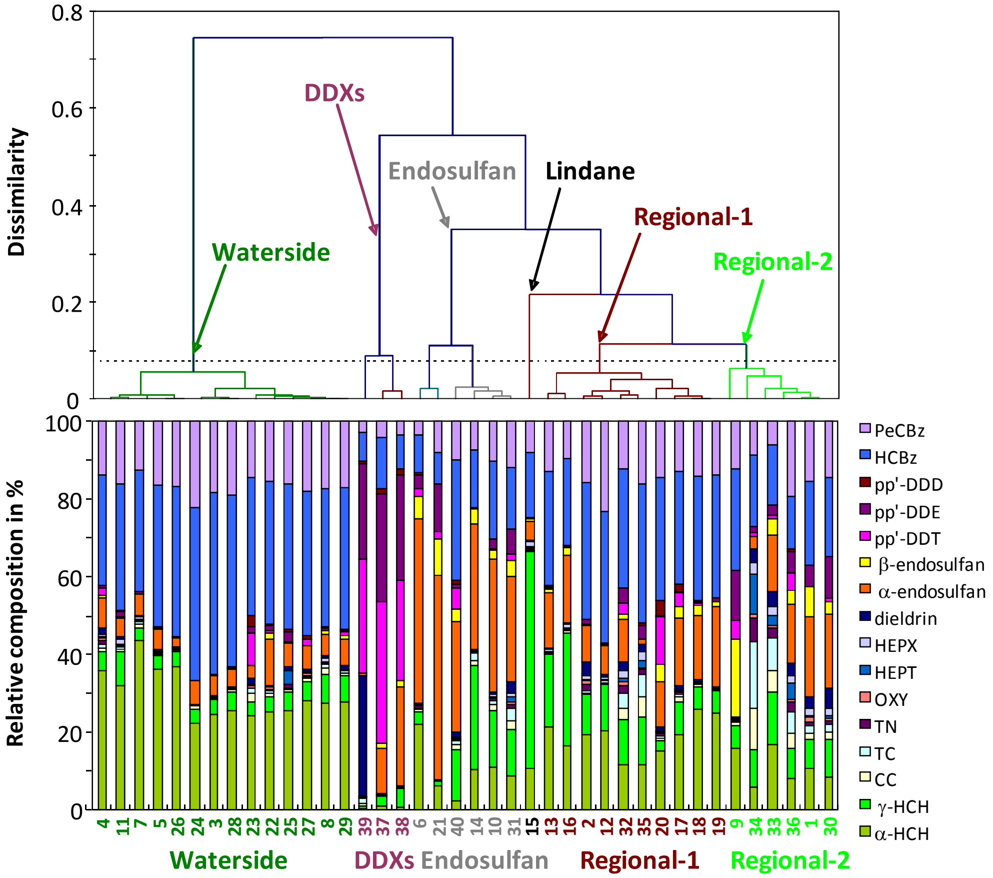


Figure 7 *Dendrogram and composition plot for the North American study. The analysis identified eight groups. Thirty one sites in three regional groups were relatively clean, whereas other ten sites were polluted.*

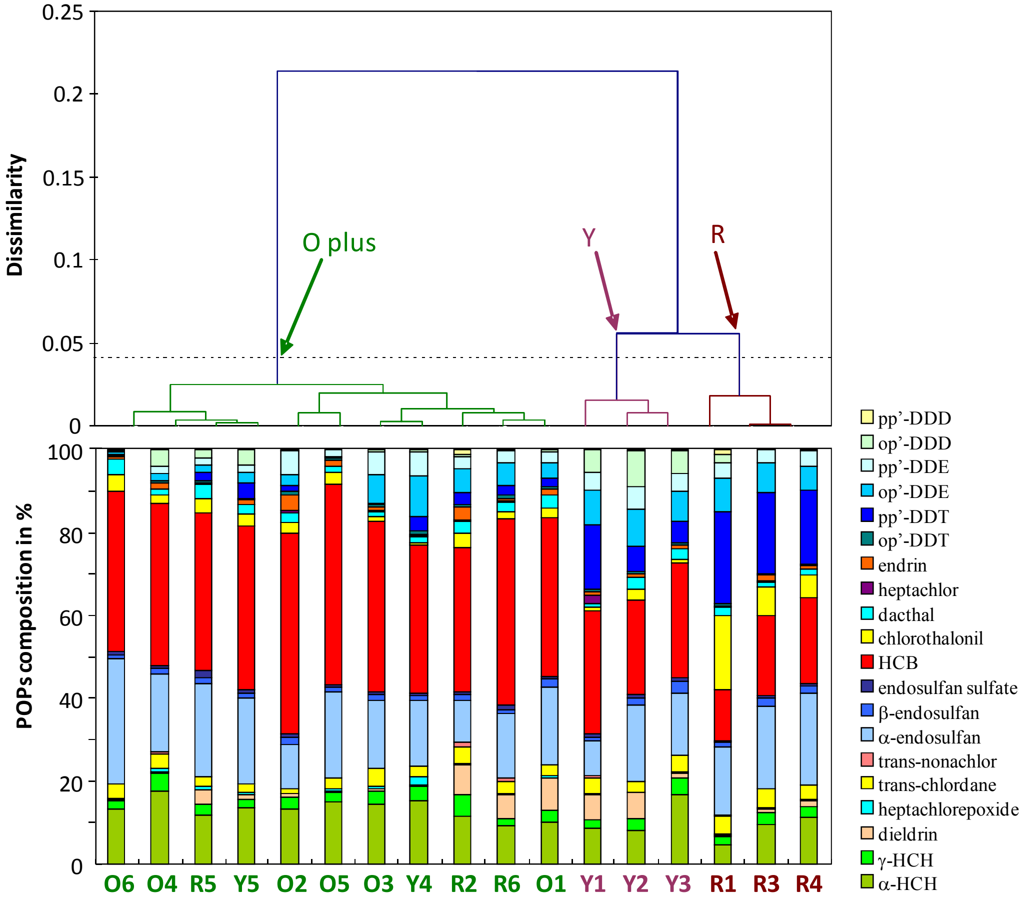


Figure 8 *Dendrogram and composition plot for the Western Canadian Mountains study. The analysis identified three groups for seventeen sites.*

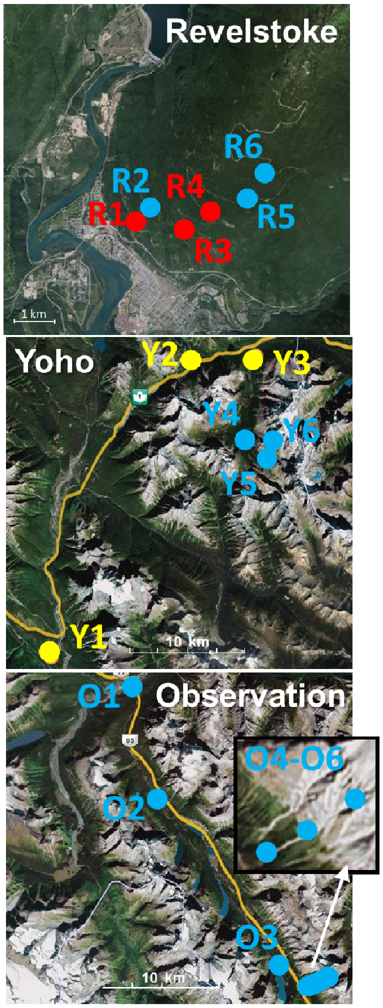


Figure 9 *Map showing passive air sampling sites in the mountains of Western Canada.*[14] *The blue cluster groups all sites from the remote Observation Peak transect with high altitude sites from the other two transects and reflects clean background air. The sites in yellow are close to a major highway, those in red close to the town of Revelstoke.*

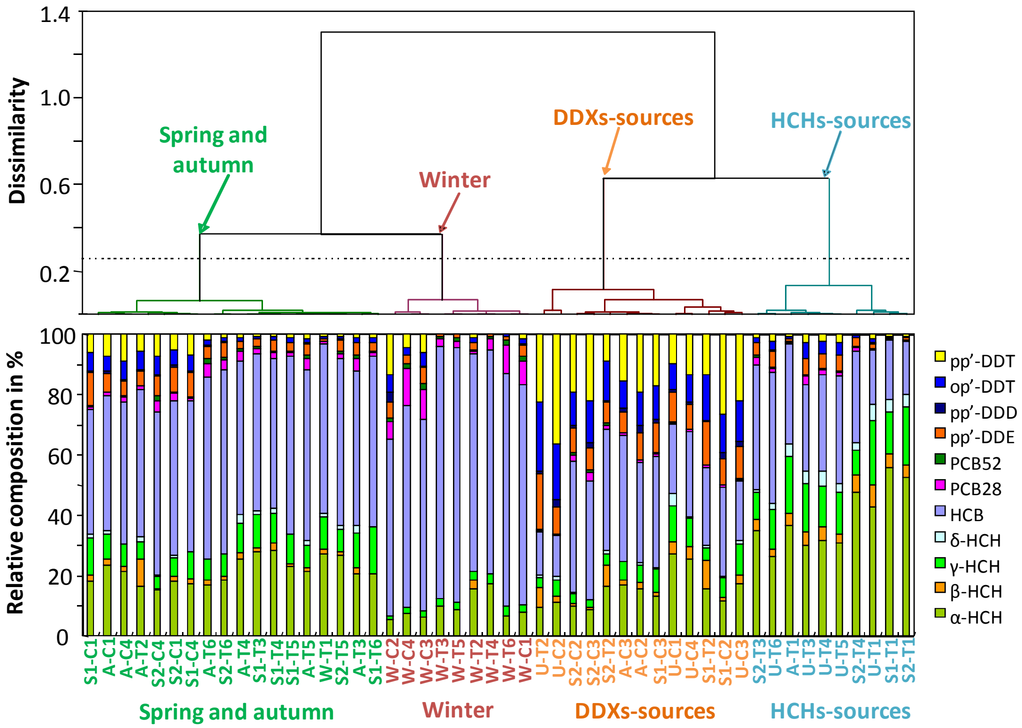


Figure 10 Dendrogram (top) and composition plot (bottom) for the cluster analysis based on compositional data from the Tianjin-Changdao Island study. The cluster analysis identified four groups. Two groups, containing 10 and 13 sites, are dominated by HCHs or DDXs source contributions, respectively. Two seasonal groups have 9 winter samples and 18 spring/autumn samples, respectively.

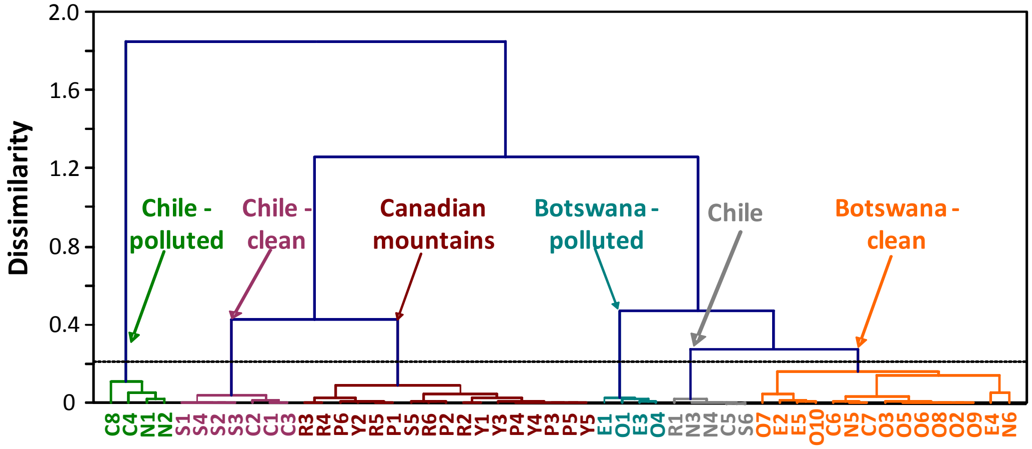


Figure 11 *Dendrogram for the combined data set of three regions. The analysis identified six groups.*

677x508mm (96 x 96 DPI)

Cluster analyses of POP data distinguish sampling sites influenced by local sources from those with regional or continental POP fingerprints.”