Delimiting affinity zones as a basis for air pollution mapping in Europe

Danielle Vienneau[1,2,3]* and David J. Briggs[1]

1. Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, Imperial College London, London W2 1PG

2. Swiss Tropical and Public Health Institute, Basel, Switzerland

3. University of Basel, Basel, Switzerland

* Corresponding author: Danielle Vienneau, PhD

Updated contact information:

Tel: +41 (0)61 284 8398

Fax: +41 (0)61 284 8105

E-mail address: danielle.vienneau@unibas.ch

Mailing address: Department of Epidemiology and Public Health
Swiss Tropical and Public Health Institute
Socinstrasse 57,
CH-4051, Basel, Switzerland

Previous contact information:

Tel: +44(0)20 7594 3348

Fax: +44(0)20 7594 3193

E-mail address: danielle.vienneau@imperial.ac.uk

Mailing address: Department of Epidemiology and Biostatistics
MRC-HPA Centre for Environment and Health
Imperial College London, St Mary's Campus
Norfolk Place, London W2 1PG, UK

Abstract

Affinity zones are defined as areas within which air quality displays consistent behaviour over space and time. Constructed using multivariate statistical techniques and physiographic and landscape variables reflecting underlying sources and spatial patterns of air pollution, affinity zones provide a spatial structure suited to exploring the representivity of monitoring networks and as a basis for air pollution mapping and exposure assessment. The affinity zone method is demonstrated using European air pollution monitoring sites, and environmental data compiled within a 1km GIS. Organised into three main stages, this method involves: (i) indicator selection, using principal components analysis, (ii) zonation by cluster analysis to classify areas into distinct types, and (iii) site allocation, to confirm similarity within affinity zones in terms of monitored air pollution concentrations. Ten interpretable and coherent air pollution affinity zones were constructed for Europe, including two rural zones and eight related to different types of densely populated and built up environments. Concentrations between affinity zones differed significantly for NO_2 background and traffic sites and for PM_{10} traffic sites only. Not all zones, however, were found to be sufficiently represented by monitoring sites, illustrating the importance of affinity zones in identifying deficiencies in monitoring networks. Spatial modelling within affinity zones is also demonstrated, showing that simple kriging of background NO_2 concentrations within zones (compared to kriging ignoring zones) produced a ca. 22% reduction in errors and increased R^2 by 0.25 at reserved validation monitoring sites. The affinity zone method developed here is a robust, statistical approach that can be used for evaluating the representivity of routine monitoring networks often used in continental level environmental and health risk assessments.

Key Words

Affinity zones; environmental stratification; air pollution; exposure modeling; risk assessment; multivariate statistical analysis

1. Introduction

Ground-based measurements of air pollution provide crucial data for many different applications, including policy evaluation, environmental and health impact assessment and epidemiological research. These data nevertheless give only a limited perspective on patterns and levels of air pollution or resulting exposures. Pollutant concentrations vary greatly, often over short distances (Gilbert et al. 2003; Zhou and Levy 2007), and individual monitoring sites may therefore be representative only of a small surrounding area. Monitoring networks are also sparse, for example, with more than half of the countries having <0.01 rural and <2 urban NO_2 sites per 100km^2 (Table 1). Furthermore, these networks have often been established primarily for the purpose of monitoring compliance with air quality objectives. Sites therefore tend to be targeted in areas considered to be air pollution hotspots (e.g. close to industrial emission sources or on roadsides) rather than in accordance with population distribution. For all these reasons, individual monitoring sites can be considered representative only of a small local area, and of similar types of location, and considerable care is needed in extrapolating data from monitoring networks to the wider population.

<<Table 1 hereabouts>>

The problem of using site-based measurements as a basis for more generalised description of the environment is not unique to air pollution. The same issues arise with almost all environmental surveys, including soils (Webster and Burrough 1974) and ecological mapping (Bernert et al. 2003; Briggs and France 1983; Carter 1997; Fairbanks and Benn 2000; Harding and Winterbourn 1997; Metzger et al. 2005). One way of dealing with these problems is by defining what may be called affinity zones. These comprise areas within which environmental conditions can be considered relatively homogeneous. They thus form both a system of partitioning for sampling, and a framework for extrapolation – i.e. within which measured data can be generalised to the zone as a whole. This approach underlies, for example, the Institute of Terrestrial Ecology (ITE) classification (Bunce et al. 1996a, 1996b, 1996c) used in the UK Countryside Survey (Barr et al. 1993; Haines-Young et al. 2000) and more recent Environmental Stratification for Europe (Jongman et al. 2006; Metzger et al. 2005). To date, however, studies of air pollution seem to have made relatively little use of these techniques, though McGregor (1996) did show how affinity zones could be applied to generalise sulphur dioxide time-series data from

monitoring sites in Birmingham, UK, and similar methods have been used to help interpret data on ultra-fine particles (Harris et al. 2009).

Here, we demonstrate the methodology and potential of the approach at a wider, continental level. Using data on nitrogen dioxide (NO_2) and particulates (PM_{10}) from the Airbase data set, we first develop affinity zones for Western Europe, and then use the results as a framework within which to assess the representivity of the European monitoring network.

2. Methods

In the context of air pollution, and following McGregor (1996), an affinity zone can be considered as an area, typically defined in terms of its source characteristics and dispersion environment, within which air pollution behaviour is consistent and predictable. As such, affinity zones can be identified and applied in two main ways: i) by first grouping monitoring sites into different classes, on the basis of their air pollution profiles, and then delineating the boundaries to the zones that these represent; ii) by first defining zones in terms of their environmental or other characteristics, and then assigning monitoring sites to these zones depending on their location. We term the first of these the 'site-based' approach, and the latter the 'area-based' approach.

The choice of approach needs to take account both of the quality of the data and the purpose of the analysis. The site-based approach is useful where attention is focused on the full time-signal of monitored concentrations (e.g. the hourly or daily signature across the year). Where sufficient monitoring sites exist to enable the definition of clear and exclusive groups, the site-based approach ensures that the zones reflect real differences in air pollution behaviour. Where attention is on longer-term (e.g. annual average) concentrations, or where the number of monitoring sites or their spatial distribution is limited, the area-based approach is likely to be more reliable. In these situations, the area-based approach ensures more realistic zonation in that it allows for the recognition of zones for which no monitoring sites exist.

Here, we are concerned with trying to assess and map long-term (annual average) exposures to air pollution, on the basis of a monitoring network that is inherently uneven in its distribution and density. For this reason, the area-based approach is used. The

general approach is summarised in Figure 1. As this indicates, a three-stage analysis is used: i) indicator selection, ii) zonation and iii) site allocation.

<<Figure 1 hereabouts>>

The analysis was undertaken as part of the APMoSPHERE (Air Pollution Modelling for Support on Health and Environmental Risks in Europe) study (Beelen 2009). Analysis focused on NO_2 and particulates (PM_{10}), two of the main traffic-related pollutants. Air pollution data, as annual average concentrations ($\mu g/m^3$) for the year 2001, were derived from the European EUROAIRNET network (Larssen 2999) which is maintained in the Airbase database. This was supplemented with sites from the EMEP network. Pollution data for the two main site types, background and traffic, were used separately for this analysis. Only those sites having equal to, or more than 75% data capture for the study year were maintained. These were stratified by location (urban/rural) and country into a training and validation subset, respectively containing 75% and 25% of the sites. In addition to air pollution data, a 1x1km GIS database was constructed for the study area. The GIS database of ca. 2.8 million grid cells comprised data on land cover (100m grid: Corine land cover version 12/2000), transportation (1:300,000 vector by road type: AND Data Ireland Ltd), topography (1km grid of mean altitude (m): TOPO30) and climate smoothed to the 1km grid using kriging (50km grids: for meteorological parameters within Operational Data Sets from the atmospheric model archived at ECMWF). Variables were computed at three levels of measurement (i.e. 1km, 5km and 21km neighbourhoods) to represent local and regional variations in air pollution. This GIS is fully described in Beelen (2009), and the population data used in subsequent analysis relate to spatially resolved census data described in Briggs (2007). Analysis was conducted in ArcGIS 9, with statistical analysis undertaken in SPSS 12.

2.1 Indicator selection

Patterns of air pollution are typically influenced by several sets of characteristics, relating to emission sources, dispersion processes and the detailed local exposure environment. Different groups of variables, representing these different conditions or 'domains' (e.g. transport, land use, topography, climate) were therefore used to characterise affinity zones (Table 2). As both local and regional trends influence air pollution, and because these operate at varying spatial scales (Briggs 2005), the variables used to represent the domains likewise needed to be measured at (or aggregated to) different levels. Focal

functions were thus applied to some variables to create these additional regionalised variables (Beelen 2009). Here, three levels of measurement were used: 1km (the finest feasible resolution with the available data) to represent local conditions, 5km for intermediate, and 21km for the regional effect.

<<Table 2 hereabouts>>

The effectiveness of the area-based affinity zone method depends crucially on how well the variables selected for clustering represent the pollution environments under study. At the same time, it is important not to include large numbers of highly inter-correlated variables in the analysis, both because these may bias the results, and also greatly increase computational demands. Initial analysis was therefore undertaken to identify the minimum yet best set of indicators with which to define the zones.

The first stage in this analysis was to generate a set of meta-variables, or factors, for each domain, using principal components analysis (Hair et al. 1998) with a varimax rotation (Yu and Chang 2001). The aim of these was to provide an inclusive, interpretable and uncorrelated set of indicators that could be used in the subsequent cluster analysis. Factors were therefore accepted only where:

1. the domain comprised a sufficient number of variables (n > 3) to make FA worthwhile;
2. the constituent variables were highly correlated;
3. the constituent variables exhibited approximately normal distributions (departures from normality can diminish observed correlations in FA);
4. the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was >= 0.7 (Norusis 1988). The KMO measure of sampling adequacy is an index that compares the magnitude of the observed correlation coefficients with the magnitude of the partial correlation coefficients. Factor analysis is not recommended when the KMO value is small, as the correlations between pairs of variables cannot be explained by other variables.

The percentage of variance criterion, established *a priori* at 90%, and scree plot were used to determine the optimal number of principal components to extract. Where acceptable factors could not be constructed, the individual variables (Table 2) were retained. Data for these were first standardised by converting to z scores.

The indicators (i.e. factors and retained variables) were then screened by examining the pair-wise Pearson correlations within each domain. To reduce inter-correlation within the set of indicators, one variable from each pair of highly correlated variables/factors ($r >= 0.8$) was pruned prior to the cluster analysis in stage 2.

2.2 Zonation using Cluster Analysis

The set of indicators thus obtained were entered into a K-means cluster analysis (CA) in order to group the 1km grid cells on the basis of their environmental characteristics. K-means CA is an algorithm designed to handle a large number of cases by dividing the objects into a predefined number of clusters, and was selected here because computational demands of analysing over 2.8 million cells proscribed use of hierarchical classification. A range of cluster solutions from six to twenty-six, by steps of two, were analysed.

The success of clustering, and of the prior selection of indicators, was evaluated by analysing and mapping the resulting clusters. *A priori* criteria were defined to evaluate effective delineation of the affinity zones such that mapped clusters had to: a) be interpretable and easily named for descriptive purposes; b) represent logical, real-world geographic features that were realistic in size and structure (i.e. no single-cell/fragmented or unreasonably large clusters); and c) be relatively free of artefacts from the input data (i.e. no over-representation of a single input feature dominating the resulting clusters). Three indices were therefore devised to identify the optimal number of clusters to extract:

1. Statistical index: total (TSS) and mean (MSS) sum of squared distance to cluster centre portrayed as a scree plot.
2. Map granularity index: percentage of adjacent grid cells, within a 3x3km and 5x5km rectangle window, in the same cluster as the target cell. A low value indicated many small or fragmented clusters, while a high value indicated that the clusters were very large.
3. Map cluster index: resulting clusters were mapped and a visual inspection of the cluster patterns performed. Where possible, clusters were interpreted and named.

Overall, solutions were sought that minimised the statistical index while providing moderate granularity and map cluster indices.

2.3 Site allocation and pollution analysis

2.3.1 Delineation of affinity zones

The monitoring sites in each cluster provided a basis for specifying pollution characteristics of the zone. If the delineation of affinity zones was successful in this respect, zones might be expected to show small within-zone variations in monitored concentrations, and large between-zone variations. The extent to which this was true was evaluated by analysing within versus between variations in monitored concentrations across the different affinity zones, using ANOVA. Because of the relatively coarse resolution (1km) of the environmental data used for clustering, and thus of the resulting maps, the zones could not be considered capable of distinguishing between very local pollution environments. Roadside environments, for example, are often considered to be restricted to within 100-250 metres of a road, beyond which concentrations tend to fall to background levels (Gilbert et al. 2003; Zhou and Levy 2007). For this reason, monitored data were divided according to the site type (background and traffic) prior to testing of the zones. Training and validation sites were also combined for this purpose. Analysis was limited to affinity zones with at least twenty monitoring sites, in order to ensure reliable representation of the pollution conditions across zones.

2.3.2 Modelling within affinity zones

While air pollutant concentrations in the affinity zones derived from cluster analysis can be expected to be relatively homogeneous, some degree of variability may nevertheless be expected. This is likely to occur not only because of the influence of local emission sources (e.g. roads), but also because of more regional trends in air pollution, not reflected in the indicators. An exploratory analysis, for background NO_2 only, was therefore conducted to determine whether the affinity zones could provide a useful framework for further spatial modelling of the pollution surface. For the background NO_2 monitoring sites designated for model building (i.e. training sites), the affinity zone in which each site was located was identified. Two models were then applied to each zone:

a) A baseline model in which the mean concentration from the relevant training sites was applied to all 1km grid cells in the zone.
b) A kriging model where simple kriging with a Gaussian model was applied to the natural log of the concentrations at training sites. Kriging was applied only to zones with ≥ 40 training sites. The baseline model was assumed for zones not meeting this criterion.

These were compared against a global kriging model, ignoring the affinity zones, in which simple kriging with a Gaussian model was applied to all 714 NO_2 training sites.

The reserved 25% validation sites were then intersected with the modelled surfaces and goodness-of-fit evaluated using R^2, root mean square error (RMSE), fractional bias (FB) and factor-of-two (F2) (Vienneau 2009).

3. Results

3.1 Indicator selection

The four domains (Table 2) were examined in turn to evaluate the suitability of the constituent variables for factor analysis. As per the *a priori* criteria, factor analysis was not carried out for land cover (non-normality of the variables), or topography (insufficient number of variables). Factor analysis for the 1km transport domain (KMO = 0.54) was rejected, leaving only the climate domain (KMO = 0.7), for which three principal components were extracted explaining 90.6% of the variation. Component 1 related to temperature, radiation and vapour pressure, component 2 related to wind speed, and component 3 was a rainfall factor (Table 3).

<<Table 3 hereabouts>>

The final set of indicators thus identified is listed in Table 4. The 21km altitude variable, which was highly correlated with altitude at the 1km level, was excluded from this set.

<<Table 4 hereabouts>>

3.2 Zonation using Cluster Analysis

A repeated process of cluster analysis was carried out to obtain a range of solutions, with different numbers of clusters. These were then evaluated against statistical and map granularity indices. Interrogation of the scree plots (not shown) suggested that the optimal number of clusters could be 10 (where the plot initially flattened) or 16 (where it flattened and tailed off). The granularity index at both stages was also moderate, indicating that these solutions were not dominated by small and fragmented or very large clusters. A wave-like structure of the granularity index and step-like structure of the statistical index suggested that there are some clusters which, when initially broken up, formed looser clusters that must be further subdivided several times before settling into a stable partition of the data. Both the 10 and 16 cluster solutions were interpreted and mapped to facilitate examination of the distribution of the clusters across the study area. Table 5 describes the resulting clusters.

<<Table 5 hereabouts>>

Comparison of the results showed that the 16 cluster solution provided a better division of the large mixed rural class: cluster 10-1, for example, splits approximately into 16-14 and 16-15; minor roads were incorporated into the mixed rural (10-1) and forested upland (10-5) clusters for the 10 cluster solution, but were represented by their own class (16-7) in the 16 cluster solution. When mapped, however, the 16 cluster solution resulted in a relatively poor map. In Germany (data not shown), for example, the map was dominated by minor roads, while a unique cluster arose representing built-up areas in the Netherlands only (16-10). The 10 cluster solution, shown in Figure 2, was therefore selected as the preferred set of affinity zones upon which the third stage, site allocation and pollution analysis, was based.

<<Figure 2 hereabouts>>

3.3 Site allocation and pollution analysis

3.3.1 Characteristics of affinity zones

Figure 3 shows the differences in concentrations between zones, by pollutant and site type, for training and validation sites combined. The ANOVA confirmed that the overall differences in concentrations were significant (p-value < 0.05) for all analyses except background PM_{10}. None of the rural or traffic-related zones had a sufficient number of monitoring sites to be included in the analysis of traffic sites.

The highest concentrations of background NO_2 tended to occur in zone 4 (metropolitan centres) while moderately high concentrations were seen in zones 2 (high density residential), 3 (low density residential), 9 (suburban) and 10 (industrial). The lowest concentrations occurred in zones 5 (forested uplands), 1 (mixed rural) and 7 (major road corridors). It should be noted that zone 7 was not, in general, heavily influenced by urban areas. A Bonferroni *post hoc* test (p-value < 0.05) further confirmed that background NO_2 concentrations within many of the affinity zones were significantly different from other zones. Significant differences in concentrations were found between zone 4 and all other affinity zones for NO_2 and between zone 4 and zone 3 and 10 for PM_{10}.

<<Figure 3 hereabouts>>

Descriptive statistics were computed to evaluate the site density in each affinity zone by area (km^2/site) and population (people/site), as presented in Table 6. Marked differences were found in site density both by area and population. In terms of area/site, there was an under-representation of rural areas, especially the forested uplands. Monitoring generally appeared more evenly and densely distributed for NO_2 compared to PM_{10}. Examination of the people/site values for the densely populated zones showed that suburban areas, for both pollutants, appeared to be under-monitored compared to the high and low density residential and metropolitan centre zones.

<<Table 6 hereabouts>>

3.3.2 Pollution mapping by affinity zone

Table 7 shows the mean and standard deviation of NO_2 concentrations by affinity zone for the NO_2 training sites. For the baseline model, the mean for each affinity zone was applied to all cells within that particular zone. This approach, however, falsely assumes that the pollution surface within each zone is flat and completely ignores the within-zone variability. For the kriging representation, a Gaussian model was fitted to the variogram for each affinity zone with sufficient number of monitoring sites (parameters shown in Table 7). An example of these resulting surfaces is provided in Figure 4. These two models are then compared against a global kriging model in which the zone structure is ignored. Evaluation of the goodness-of-fit for each model, based on comparisons at the reserved validation sites, is presented in Table 8. For all models, fractional bias was small. Kriging within zones, however, gave the highest R^2 (0.51) and F2 (92%) and lowest RMSE (7.42 $\mu g/m^3$). Both models which recognise the affinity zones, however, outperformed the global kriging model. This supports the notion that representivity of monitoring is likely to differ depending on the zone.

<<Table 7 hereabouts>>

<<Table 8 hereabouts>>

<<Figure 4 hereabouts>>

4 Discussion

This paper describes the development and testing of a methodology, using multivariate statistical techniques, for defining affinity zones as a way of characterising the air pollution environment in Europe. Affinity zones are delineated as a basis to explore representivity of EUROAIRNET - the regulatory air pollution monitoring network in Europe, comprising sites from national networks and reported in the Airbase database. Ten interpretable and coherent affinity zones are identified, four related to densely populated areas (metropolitan centres, high density residential, low density residential and suburban), four representing industry and transport infrastructure, and two types of rural areas.

A monitoring network with adequate spatial coverage is essential for air quality management. The extent to which existing networks achieve adequate coverage is largely unknown, in part because of the difficulty of testing how representative the networks are of pollution levels across the wider landscape. This problem is especially important if the data from monitoring sites is to be used as a basis either for pollution mapping or exposure assessment. The relevance of data from routine networks for exposure assessment is often limited given that the areas covered by monitoring do not always reflect the target areas of interest – e.g. where people live (WHO 1999). A further major consideration in this context is that ground-based routine monitoring networks, such as the European network used here, are normally designed for assessing whether air quality standards have been exceeded. The area-based affinity zone method, therefore, provides a helpful framework within which to evaluate the representivity of monitoring networks. Affinity zones can identify areas where monitoring is potentially inadequate, and can also help determine the suitability of extrapolating measurements to unmonitored locations.

Results of this analysis highlight some of these issues related to potential deficiencies in the monitoring network. The ANOVA and box plots (Figure 3), for example, show that while NO_2 background sites generally provide good coverage of the affinity zones (with 96% of the total population residing in the eight represented zones), at least half of the zones for other pollutants/site types did not. Although most of the highly populated residential zones (representing 85% of the population) are covered by background PM_{10} sites, substantially less of the population (61%) is represented by the four zones included in the traffic analyses. Notably, none of the traffic-related affinity zones have sufficient traffic sites for ANOVA. This fact alone indicates that some of the affinity zones are not adequately monitored, which presents a problem for air pollution modelling more

generally. This is further evidenced by the descriptive analysis of site density by affinity zone (Table 6). Here we find that rural areas, where approximately 26% of the population reside, are generally under-represented. Suburban areas are also identified as being potentially under-represented by monitoring sites, however only a small proportion of the population (4%) reside within this zone. In general, monitoring is seen to be more evenly and densely distributed for NO_2, but shows greater deficiencies for PM_{10}. The former is also true of sulphur dioxide and ozone, while carbon monoxide monitoring in 2001, like PM_{10}, is sparse in many of the zones (data not shown).

Overall zones with distinct and homogeneous background pollutant concentrations are defined for NO_2: the two rural zones, for example, have significantly different concentrations and are both significantly different from the four urban zones. The box plots show that the situation for PM_{10}, however, is far less coherent. This seems to be partly a function of the smaller number of monitoring sites for PM_{10} (making characterisation of the zones more difficult), but also because PM_{10} seems to show less geographic variation than NO_2, perhaps due to the greater influence of long-range particulate transport. In assessing representativeness, therefore, variability within the affinity zones must also be taken into account. A small number of sites might, for example, be considered representative if the affinity zone shows little variation in concentrations and/or the population within the zone is small. On the other hand, more sites might be needed if the variability in concentrations is great and/or the affinity zone has a very large population.

In principle, the affinity zones define areas representative of the monitoring sites that they contain. Where a sufficient number of sites exist, therefore, it should be possible to extrapolate the monitored concentrations from these sites to all unmonitored locations in the same zone. The way in which such an extrapolation is done, however, must reflect the spatial variations of pollutant concentrations within the zones. Comparisons at reserved validation monitoring sites illustrate, as expected, that the simplistic baseline model provides only a moderate characterisation of the pollution surface. Kriging within affinity zones (which is not possible in all zones), compared to kriging without zones, reduces the errors (RMSE) in predictions of background NO_2 concentrations by approximately 22% and gives marked improvement in the R^2. This demonstrates the clear potential for using affinity zones as a framework for further modelling. A more refined method such as co-kriging or land use regression (Briggs et al. 2000; Beelen et al. 2009; Hoek et al. 2008; Jerrett et al. 2005), which includes covariates as proxies of emissions

and the dispersion environment, is expected to be even better at characterising the local variations in air pollution. Likewise, the spatial framework offered by the affinity zones could be used as a basis for more sophisticated spatial-temporal modelling of air pollution. These methods, however, will also be limited by the number of available monitoring sites, and should also consider that variation in pollution concentration occurs at different scales depending on the environmental characteristics.

The delineation of affinity zones is a multi-stage process and the main drawback relates to the practical challenges of the procedure. Running repeated clustering algorithms on 2.8 million cases (i.e. 1km grid cells), for example, is time consuming. It also takes time and effort to correctly interpret the clusters. Extensive knowledge of the study area or maps of the input variables, however, can aid in successful interpretation of the clusters.

The first stage of the area-based approach involves an exploratory analysis to select the optimum set of indicators for use in cluster analysis, while the second stage is concerned with the selection of the most appropriate clustering algorithm. To arrive at the best partition of the landscape and physiographic features to define air pollution affinity zones, the approach, in the words of Carter (1997), had to be exploratory and iterative. The approach is also based on the assumption that the most appropriate division of the landscape and environmental features is achieved. The nature of the methodology, however, might result in a different set of affinity zones depending on the decisions taken at various stages. Although validation of the landscape clusters was not possible, sensitivity analyses were conducted to demonstrate that the resulting clusters were robust to changes in the clustering method and minor changes in the input variables.

In general, the main challenge is in choosing the correct number of clusters to extract - a process which is somewhat subjective and requires trial and error. While K-means CA can provide the best partition of a dataset when the number of clusters is known, it is often difficult to determine the 'optimal number' *a priori*. A range of solutions was thus sought and assessed against several predefined criteria, based on statistical and geographical indices, collectively to identify the appropriate number of clusters that might represent air pollution affinity zones. Hierarchical CA is advantageous when there is no *a priori* information on the number of partitions. It cannot, however, be run on large data sets. As a sensitivity analysis, therefore, hierarchical CA on repeated samples of the full data set was investigated. This resulted in a set of 12 clusters which were compared to the 10 defined using K-means CA. The overall agreement between cluster solutions was

significant (kappa value of 0.936), with percent agreement ranging from 52% for two clusters to >93% for more than half of the clusters. K-means CA was ultimately selected as it was the more straightforward process. Despite the approximately 3-fold longer computer processing time, interpretation of the K-means CA results was also easier and quicker than for the hierarchical CA method.

As previous studies of ecological mapping has shown (Briggs and France 1983; Metzger 2005; Bunce et al. 1996a), the success of these classification methods are, to a great extent, dependent on the initial choice of input variables to reflect the conceptual model. For this reason emphasis is placed on indicator selection prior to cluster analysis in stage 2. The transport and land cover variables associated with built-up, highly-trafficked areas, were included as important indicators of local source activity. Topographic and climatic variables, which represent regional phenomena, were included to characterise the regional dispersion environment (e.g. areas of high altitude, topographic exposure or wind speed are likely to be associated with lower levels NO_2 and PM_{10}). Seasonal (summer and winter) annual temperatures were also included to allow for differentiation between the varying regional climatic zones in the study area (e.g. Mediterranean, continental and arctic). A sensitivity analysis was conducted to assess whether CA should be customised by pollutant using variables specifically thought to influence NO_2 or PM_{10}. Local agricultural activities involving the exposure of bare soil, such as tilling fields, may contribute to PM_{10}; thus, an additional variable to reflect such activity was included for PM_{10} only. For ease of comparison, a 14 cluster solution for CA with and without the tillage variable (PM_{10} and NO_2 set of variables, respectively) was selected and the concordance between the two results examined. Thirteen similar clusters were obtained, and the overall agreement between cluster solutions was significant (kappa value of 0.984). Eleven of the 14 clusters were highly similar with >99% agreement. As a result, the use of a common set of input variables to define clusters for both NO_2 and PM_{10} was considered justified.

Overall, the concordance analyses showed that slight changes in the input variables or clustering method did not significantly alter the outcome of the analysis. Even so, there will usually be no single, ideal set of clusters that can characterise a monitoring network. However, the process of constructing and comparing different solutions can help to reveal some of the factors that influence pollution patterns in an area, and some of the limitations of the monitoring network, as illustrated here. In terms of updateability, stage 1 and 2 of the analysis would only have to be repeated if affinity zones were to be

generated for pollutants with very different sources (e.g. ozone), or as updates for the input data (e.g. land cover and transport) were made available. Determination of affinity zones from the clusters (stage 3) can be easily updated as new pollution measurements become available or as new monitoring sites are added to the network.

5 Conclusions

Air quality monitoring provides only a partial picture of the air pollution situation across the wider environment or population. This limits the reliability of the monitored data as a basis for air quality management, exposure assessment and many other purposes. An excessively dense network of monitoring stations is required to resolve these limitations completely. Nevertheless, network design could be improved, in many cases, by strategic siting of the monitoring sites, such that they can provide a better basis for interpolation. Affinity zones offer a potentially useful tool in this respect, both for evaluating existing networks and as a framework for designing new ones. The affinity zone map for Western Europe is available for non-commercial use from the authors.

Acknowledgements

This work was largely undertaken in the framework of the EU-funded APMoSPHERE project (EVK2-2002-00577). It was also partly supported through the EU 6th Framework Programme INTARESE project (018385-2). The authors gratefully acknowledge the financial support given by the funders, and the scientific input and advice of colleagues working on these projects.

References:

Barr CJ, Bunce RGH, Clarke RT, Fuller RM, Furse MT, Gillespie MK, et al. Countryside Survey 1990 Main Report. Department of the Environment; 1993.

Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K, Briggs DJ. 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Sci Total Environ 2009; 407(6):1852-1867.

Bernert JA, Eilers JM, Sullivan TJ, Freemark KE, Ribic C. A quantitative method for delineating regions: An example for the western corn belt plains ecoregion of the USA. Environ Manage 2003;21(3):405-420.

Briggs DJ. The role of GIS: Coping with space (and time) in air pollution exposure assessment. J Toxicol Environ Health Part A 2005;68:1243-1261.

Briggs DJ, de Hoogh C, Gulliver J, Wills J, Elliott P, Kingham S, et al. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci Total Environ 2000;253(1-3):151-167.

Briggs DJ, Gulliver J, Fecht D, Vienneau DM. Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote Sens Environ 2007;108:451-466.

Briggs DJ, France J. Classifying landscapes and habitats for regional environmental planning. J Environ Manage 1983;17:249-261.

Bunce RGH, Barr CJ, Clarke RT, Howard DC, Lane AMJ. Land classification for strategic ecological survey. J Environ Manage 1996a;47(1):37-60.

Bunce RGH, Barr CJ, Clarke RT, Howard DC, Lane AMJ. Special Paper: ITE Merlewood land classification of Great Britain. J Biogeogr 1996b;23(5):625-634.

Bunce RGH, Barr CJ, Gillespie MK, Howard DC. The ITE land classification: Providing an environmental stratification of Great Britain. Environ Monit Assess 1996c;39(1-3):39-46.

Carter SE. Spatial stratification of Western Kenya as a basis for research on soil fertility management. Agric Syst 1997;55(1):45-70.

Fairbanks DHK, Benn GA. Identifying regional landscapes for conservation planning: a case study from KwaZulu-Natal, South Africa. Landscape Urban Plan 2000;50(4):237-257.

Gilbert NL, Woodhouse S, Stieb DM, Brook JR. Ambient nitrogen dioxide and distance from a major highway. Sci Total Environ 2003;312(1-3): 43-46.

Haines-Young R H, Barr CJ, Black HIJ, Briggs DJ, Bunce RGH, Clarke RT, et al. Accounting for nature: assessing habitats in the UK countryside. Department of the Environment, Transport and the Regions, London; 2000.

Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate Data Analysis, 5th edn, Prentice-Hall International, Inc.; 1998.

Harding JS, Winterbourn MJ. An ecoregion classification of the South Island, New Zealand. J Environ Manage 1997;51(3): 275-287.

Harris P, Lindley S, Gallagher M, Agius R. Identification and verification of ultrafine particle affinity zones in urban neighbourhoods: sample design and data pre-processing. Environ Health 2009;8(Suppl 1):S5 doi:10.1186/1476-069X-8-S1-S5.

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 2008;42(33):7561-7578.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 2005;15(2):185-204.

Jongman RHG, Bunce RGH, Metzger MJ, Mücher CA, Howard DC, Mateus VL. Objectives and Applications of a Statistical Environmental Stratification of Europe. Landscape Ecol 2006;21(3):409-419.

Larssen S, Sluyter R, Helmis C. Criteria for EUROAIRNET. The EEA Air Quality Monitoring and Information Network. Technical Report No. 12. European Environment Agency; 1999.

McGregor GR. Identification of air quality affinity areas in Birmingham, UK. Appl Geogr 1996;16(2):109-122.

Metzher MJ, Bunce RGH, Jongman RHG, Mücher CA, Watkins JW. A climatic stratification of the environment of Europe. Global Ecol Biogeogr 2005;14:549-563.

Norusis MJ. SPSS/PC+ Advanced Statistics V2.0 SPSS Inc., Chicago; 1988.

Vienneau D, de Hoogh K, Briggs D. A GIS-based method for modelling air pollution exposures across Europe. Sci Total Environ 2009;408(2):255-266.

Webster R, Burrough PA. Multiple discriminant analysis in soil survey. J Soil Sci 1974;25(1):120-134.

WHO. Monitoring Ambient Air Quality for Health Impact Assessment, Regional Office for Europe, Copenhagen; 1999.

Yu TY, Chang LF. Delineation of air-quality basins utilizing multivariate statistical methods in Taiwan. Atmos Environ 2001;35(18):3155-3166.

Zhou Y, Levy J. Factors Influencing the Spatial Extent of Mobile Source Air Pollution Impacts: A Meta-Analysis. BMC Public Health 2007;7:89 doi:10.1186/1471-2458-7-89.

Table 1 **Number of sites per area and people, by country**

Country	Sites/100km^2				Sites/100,000 people	
	NO_2		PM_{10}			
	Rural (n=360)	Urban (n=1082)	Rural (n=128)	Urban (n=553)	NO_2	PM_{10}
Austria	0.068	7.89	0.012	2.92	1.61	0.46
Belgium	0.026	0.55	0.004	0.40	0.28	0.17
Denmark	0.005	0.33	0.002	0.05	0.15	0.04
Finland	0.002	1.06	0.001	1.39	0.37	0.38
France	0.007	2.66	0.003	0.96	0.59	0.21
Germany	0.029	1.86	0.015	1.19	0.52	0.32
Great Britain	0.004	0.57	0.002	0.37	0.14	0.09
Greece	0.003	2.10	0.002	0.99	0.19	0.09
Ireland	0.004	0.55	-	0.92	0.15	0.13
Italy	0.010	1.36	0.001	0.46	0.21	0.06
Netherlands	0.077	0.66	0.031	0.31	0.28	0.12
Portugal	0.005	2.73	0.001	1.13	0.20	0.08
Spain	0.015	3.82	0.005	2.12	0.46	0.22
Total	**0.013**	**1.66**	**0.005**	**0.85**	**0.39**	**0.19**

Note: based on background and traffic sites in 2001 with >= 75% data capture

Table 2 Indicator variables, by measurement level and domain

Variable	Measurement level		
	Regional 21km	**Intermediate 5km**	**Local 1km**
Transport Domain: length of road (km/neighbourhood)			
Motorways		✓ (major roads)	✓
A-roads			✓
B-roads			✓
Minor roads			✓
Land Cover Domain: percent land cover area (%)			
HD residential	✓ (total built-up)	✓ (non-residential built up)	✓
LD residential			✓
Industry			✓
Transport			✓
Airports			
Seaports			
Construction			
Urban greenery			✓
Forestry			✓
Topography Domain#			
Altitude (mean elevation (m) above sea level)	✓		✓
Topex (height difference between 1km centroid and centroid of 24 surrounding 1km cells)			✓
Distance to sea (distance in km to nearest coastline)			✓
Climate Domain*			
Summer temperature (ºC)	✓		
Winter temperature (ºC)	✓		
Annual radiation (Kj/m^2)	✓		
Annual wind speed (m/s)	✓		
Annual calm days (%)	✓		
Annual vapour pressure (hPa)	✓		
Annual rainfall (mm)	✓		

Notes:

Merged cells indicate grouped variables with group name provided in brackets

#Variables in the topography domain were rescaled using the following transformations:

Altitude calculated as √(nalt/max(nalt)), where nalt = altitude – min(altitude)

Topex calculated as √(ntopex/max(ntopex)), where ntopex = topex – min(topex)

Distance to sea calculated as √(minimum distance/max(minimum distance))

* Climate domain variables were available at 1km, representative of regional variations (derived from 50km ECMWF data)

Table 3 Rotated principal component loadings for climate domain

Variable	Component 1	Component 2	Component 3
Summer temperature	**0.895**	-0.136	-0.351
Winter temperature	**0.939**	0.256	0.045
Annual radiation	**0.882**	-0.181	-0.276
Annual wind speed	0.032	**0.922**	0.099
Annual calm days	-0.071	**-0.896**	-0.029
Annual vapour pressure	**0.922**	0.238	-0.002
Annual rainfall	-0.176	0.085	**0.974**

Note: High factor loadings shown in bold

Table 4 Indicators used in cluster analysis by measurement level

Domain	Variables		Factors
	Z Scores	Measurement Level	1km Level
Transport	Motorways	1km	
	A-roads	1km	
	B-roads	1km	
	Minor roads	1km	
	Major roads	5km	
Land Cover	HD residential	1km, 5km	
	LD residential	1km, 5km	
	Industry	1km	
	Other built up	1km	
	Urban greenery	1km	
	Forest	1km	
	Non-residential built up	5km	
	Total built up	21km	
Topography	Altitude	1km	
	Topex	1km	
	Distance to sea	1km	
Climate			High temperature
			Windy
			Rainy

Table 5 Description of clusters for 10 and 16 cluster solutions

10 Cluster Solution			16 Cluster Solution		
No.	**% Area**	**Description**	**No.**	**% Area**	**Description**
10-1	61.3	Mixed rural	16-1	0.1	Urban green space
10-2	0.4	High density residential	16-2	2.3	Low density residential
10-3	3.1	Low density residential	16-3	0.5	Mixed industrial
10-4	0.1	Metropolitan centres	16-4	0.2	Motorway intersections
10-5	28.3	Forested uplands	16-5	0.3	Other built-up
10-6	1.2	Motorways	16-6	0.3	Suburban
10-7	4.5	Major road corridors	16-7	13.9	Minor roads
10-8	0.2	Motorway intersections	16-8	1.2	Motorways
10-9	0.3	Suburban	16-9	0.2	Industrial
10-10	0.5	Industrial	16-10	0.4	Built-up areas – NL*
			16-11	20.1	Northern rural
			16-12	0.3	High density residential
			16-13	4.3	Major road corridors
			16-14	26.8	Southern rural
			16-15	29.0	Central rural
			16-16	0.1	Metropolitan centres

Note: * Unique class in the Netherlands only

Table 6 Monitoring site density by affinity zone

	Affinity Zone	Background		Traffic		Total	
		Area/ Site (km^2)	People /Site ('000)	Area/ Site (km^2)	People /Site ('000)	Area/ Site (km^2)	People /Site ('000)
NO_2							
1	Mixed rural	9,854	479	110,363	5,363	9,046	483
2	High density residential	146	475	90	294	56	199
3	Low density residential	228	337	576	855	163	249
4	Metropolitan centres	55	475	28	241	18	164
5	Forested uplands	12,576	170	226,361	3,064	11,914	167
6	Motorways	2,128	345	4,255	690	1,419	235
7	Major road corridors	3,023	478	7,467	1,182	2,152	354
8	Motorway intersections	840	851	327	331	235	243
9	Suburban	303	484	699	1,117	211	349
10	Industrial	150	302	487	981	115	240
PM_{10}							
1	Mixed rural	26,701	1,297	827,719	40,219	25,866	1,380
2	High density residential	237	771	169	551	99	351
3	Low density residential	469	695	1,153	1,709	333	508
4	Metropolitan centres	99	861	59	514	37	329
5	Forested uplands	45,272	613	339,542	4,596	39,946	560
6	Motorways	11,348	1,841	8,511	1,381	4,863	805
7	Major road corridors	7,934	1,255	14,105	2,232	5,078	836
8	Motorway intersections	5,878	5,957	653	662	588	608
9	Suburban	534	854	1,514	2,421	395	652
10	Industrial	303	610	682	1,373	210	440

Table 7 Background NO_2 training sites: monitored concentrations and kriging parameters

	Affinity Zone	Monitored Concentrations (Baseline Model)			Kriging Model Parameters		
		N	Mean ($\mu g/m^3$)	SD ($\mu g/m^3$)	Range (km)	Partial sill (ln $\mu g/m^3$)	Nugget (ln $\mu g/m^3$)
1	Mixed rural	125	14.16	8.11	1,637	0.38	0.16
2	High density residential	50	30.30	9.91	630	0.12	0.07
3	Low density residential	297	26.31	9.08	217	0.12	0.07
4	Metropolitan centres	54	38.06	11.83	1,208	0.18	0.02
5	Forested uplands	40	8.73	4.80	-	-	-
6	Motorways	13	25.62	10.97	-	-	-
7	Major road corridors	36	15.64	7.21	-	-	-
8	Motorway intersections	5	36.00	6.52	-	-	-
9	Suburban	25	26.72	9.23	-	-	-
10	Industrial	69	26.70	9.29	568	0.14	0.07
	All zones (global kriging)	714	-	-	1,970	0.38	0.22

Table 8 Evaluation of NO_2 background models across all affinity zones

Model	N	R^2	RMSE	F2	FB
Baseline	217	0.42	8.13	88.9%	0.05
Kriging within zones	217	0.51	7.42	91.7%	0.02
Global kriging	217	0.27	9.04	84.8%	0.01

Figure 1 **Area-based affinity zone method**

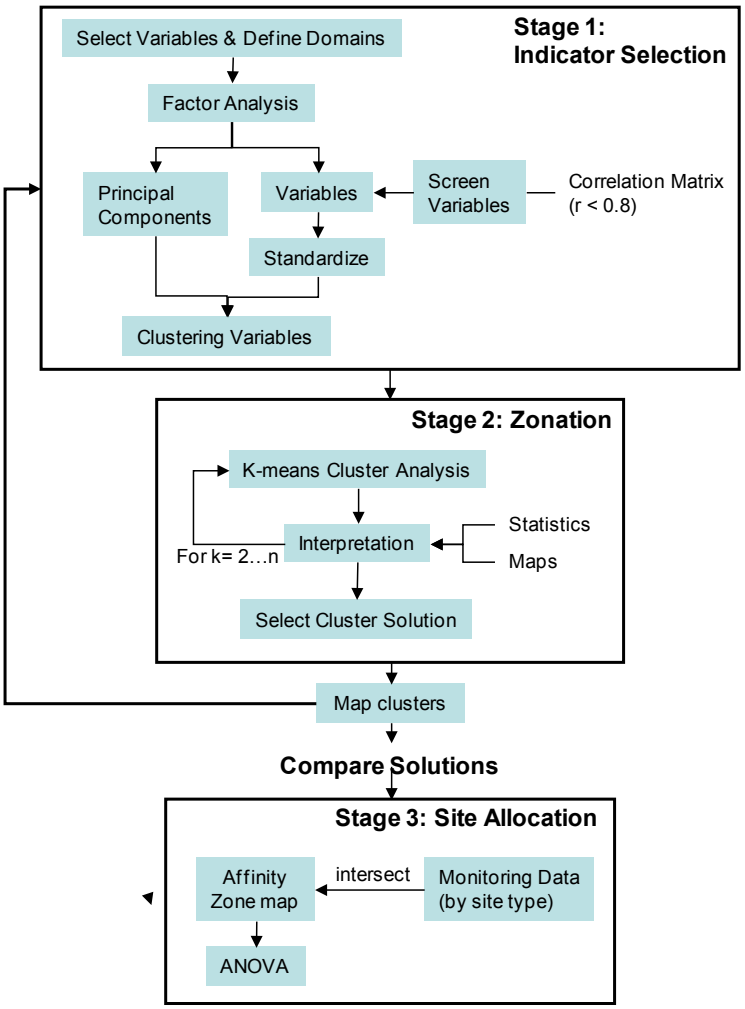

Figure 2 Affinity zone map

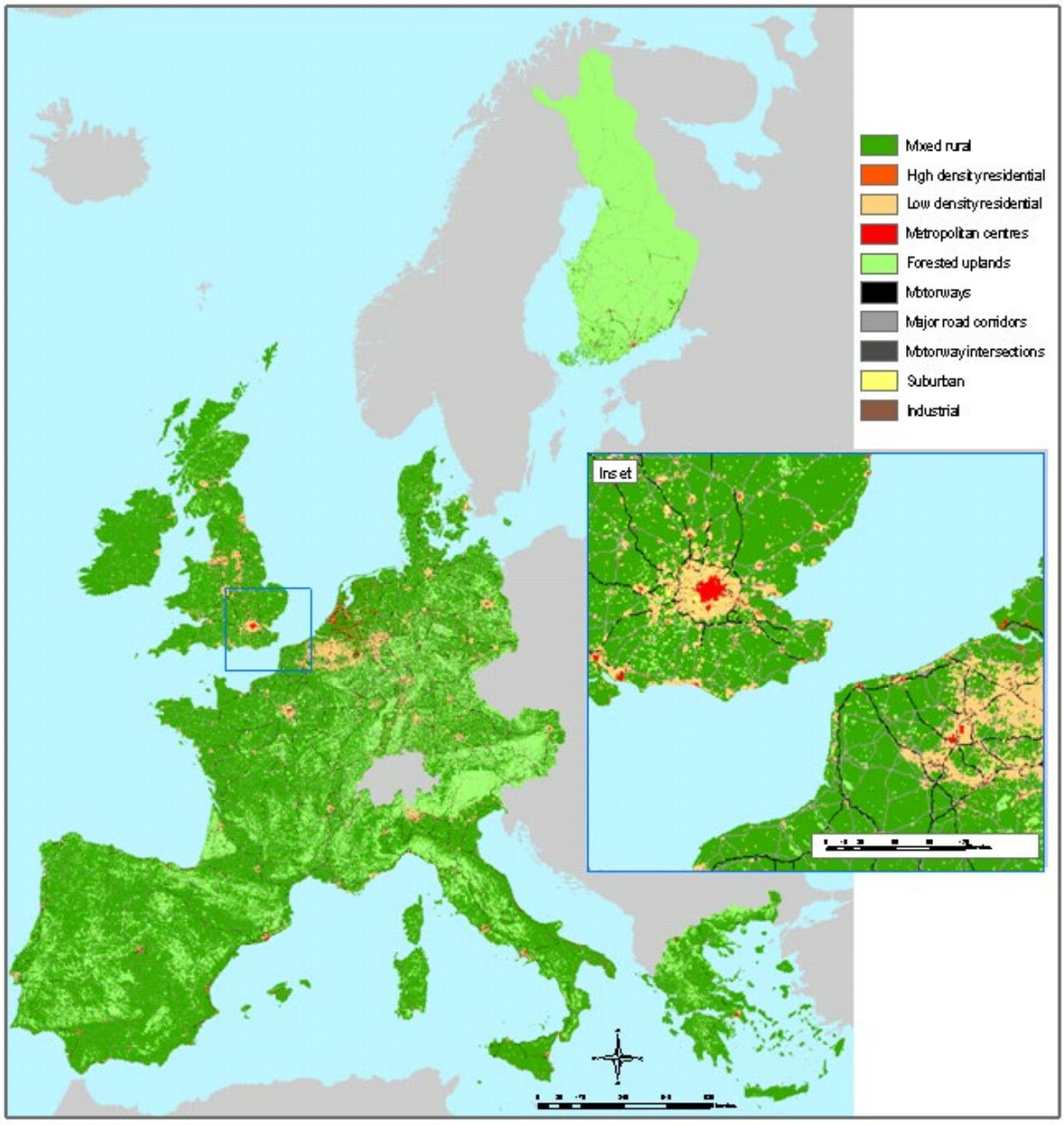

Figure 3 **Air pollutant concentrations by affinity zone**

(Training and validation sites combined, zones with <20 monitoring sites excluded)

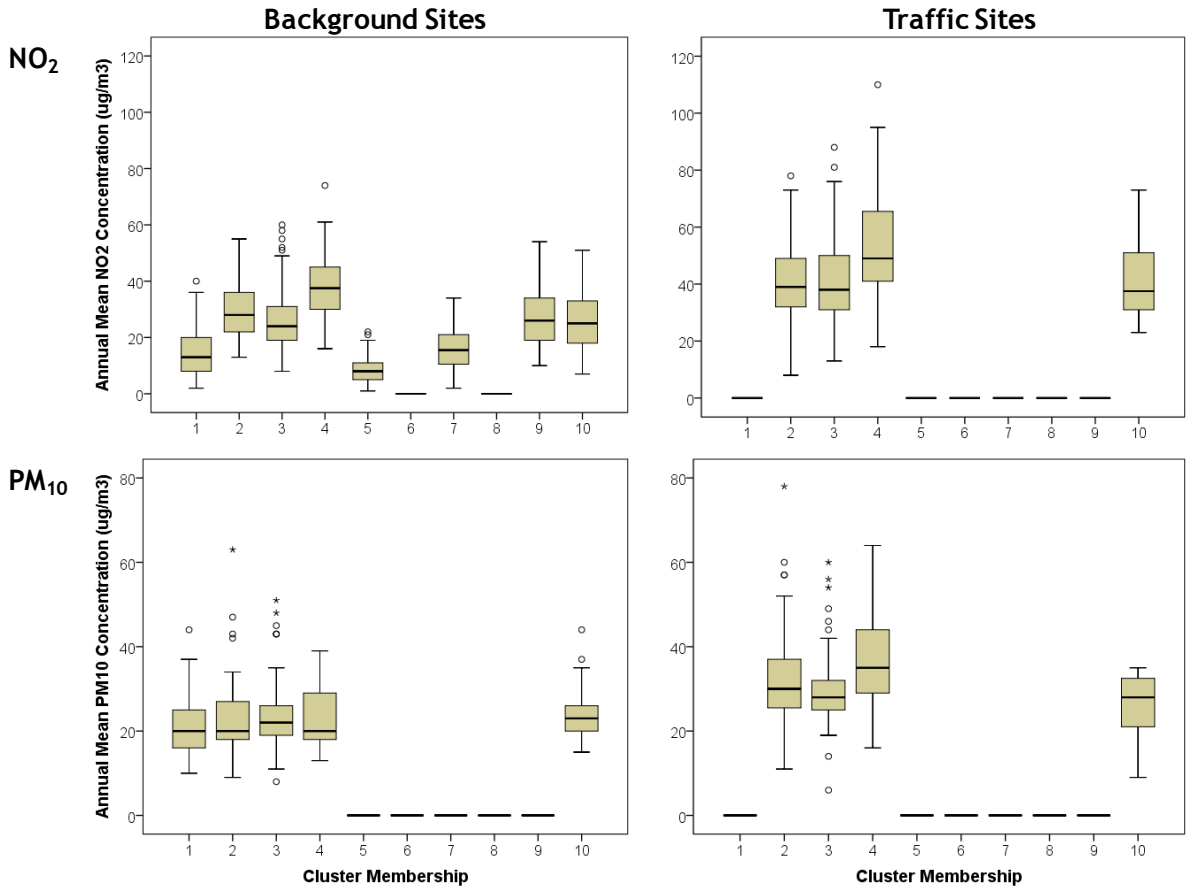

Figure 4 **Comparison of baseline vs. kriging model**

Background NO_2 concentrations ($\mu g/m^3$) in area surrounding Madrid, Spain