

SPATIAL MODELLING OF AIR POLLUTION FOR EXPOSURE ASSESSMENT

Danielle M. Vienneau

**Department of Epidemiology and Public Health
Imperial College London**

**Submitted for the degree of Doctor of Philosophy
University of London**

2006



Abstract

This thesis addresses the need for high resolution (ca. 1km²) data for assessing risks associated with ambient air pollution across the European Union. GIS techniques are used to model the population distribution and air pollution concentrations as a basis for exposure assessment. A high resolution GIS is developed by linking and enhancing data from ground-based, satellite and census sources including data on land cover, emissions, topography, meteorology, population, and routine air pollution monitoring.

NUTS 5 population totals are disaggregated to the 1km level on the basis of land cover and light intensity using dasymetric mapping and regression techniques. The resulting population density map is validated against postcode and headcount data in Great Britain ($r > 0.93$). Affinity zone stratification is then used to assess the spatial coverage of the European air pollution monitoring network, and used as a basis for air pollution mapping ($R^2 = 0.51$ for NO₂). Pollution surfaces for NO₂ and PM₁₀ are also created using a moving window approach in which emissions are dispersed using distance decay functions. The moving window approach is shown to perform ($R^2 = 0.61$ for NO₂) as well as a more conventional approach involving regression mapping and kriging.

Integrated into a spatial database, the population density and pollution maps are used to assess exposure to annual mean NO₂ and PM₁₀ for 2001. The GIS-based exposure assessment allows for quick interpretation of the areas in which guideline levels are exceeded, and provides a good approximation of the number of people residing in polluted areas. Across the EU, 4 and 6% of the population were found to be potentially exposed to above guideline levels of PM₁₀ and NO₂, respectively, with the largest proportion of exposed people resident in the southern European countries and in the densely populated affinity zones.

Declaration of Own Work

I declare that the work presented in this thesis is my own, and that where material relating to the work of others is presented, it is fully referenced and/or appropriate acknowledgement is given.

This research was funded by, and conducted in parallel with, the APMoSPHERE project. APMoSPHERE, Air Pollution Modelling for Support to Policy on Health and Environmental Risk in Europe, was a thematic project under the Global Monitoring for Environment and Security initiative as part of the European Unions Fifth Framework Programme. The project included partners at Imperial College, Utrecht University, the University of Bath, the University of Oslo, the National Observatory of Athens and AEA Technology, National Environmental Technology Centre (Netcen) in the UK.

My contribution to the APMoSPHERE project, with advice from my supervisors and principle investigators, included: building and cleaning the majority of the GIS and distributing the data to partners; population modelling; affinity zone stratification; developing and applying focalsum methods for pollution modelling; mapping the project results; intersects and summary tables for the exposure assessment. Output from the APMoSPHERE project used in this thesis that was generated by other researchers included: the emission models and maps; and pollution modelling results from the regression mapping and geostatistical methods.

Signature of author *Danielle Vennart* Date *4 Dec 06*

Acknowledgements

I would like to express my appreciation to those who gave me the opportunity to complete this thesis. In particular, I am deeply grateful to my supervisor Prof. David Briggs, whose suggestions, support and patience have been welcomed throughout my research and writing up. I am also very appreciative of the funding provided by the APMoSPHERE project and the Universities UK Overseas Research Students Awards Scheme.

At some stage along the way, every member of the SAHSU GIS group at Imperial College has helped me with this endeavour. Most importantly is Dr. Kees de Hoogh, my second supervisor. Thank you, Kees, for your continuous guidance and words of encouragement along the way. I am grateful to Daniela Fecht and her predecessor John Mitchell for helping out with the boring bits, Dr. John Gulliver for sharing his knowledge of GIS and air pollution modelling (and for suffering through early thesis drafts), Dr. Linda Beale for her programming guidance and words of wisdom, and Chloe Morris who was always willing to provide remote assistance when I was working from home. I would also like to thank members of the SAHSU Database group, in particular Peter Hambly, Dr. Norman Cobley and Dr. Nariman (Mojgan) Amin, for resolving my tricky data format issues.

Many thanks as well to the APMoSPHERE project team - Dr. Chris Dore, Stephen Pye, Dr. Gerard Hoek, Dr. Edzer Pebesma, Dr. Kor de Jong, Rob Beelen, Dr. Michael Petrakis, Asbjørn Aaheim and especially Dr. Gavin Shaddick for his continued patience in explaining statistics to me.

Finally, I would like to thank my parents for inspiring and encouraging me always and my husband Seb for being so supportive and patient.

Table of Contents

1	INTRODUCTION	16
1.1	Rationale	16
1.2	Aims	19
1.3	The APMoSPHERE project	19
1.4	Thesis overview.....	20
2	LITERATURE REVIEW	21
2.1	Air pollution overview	21
2.1.1	The policy context.....	21
2.1.2	Pollutants of interest.....	23
2.1.3	Temporal and spatial variations	27
2.2	Exposure assessment.....	30
2.2.1	Nearest pollution monitor	31
2.2.2	Proxies for exposure.....	33
2.2.3	Exposure modelling.....	34
2.3	Air pollution modelling methods	40
2.3.1	Spatial interpolation.....	40
2.3.2	Stochastic methods	48
2.3.3	Dispersion modelling	50
2.4	Conclusion	52
3	GIS DEVELOPMENT.....	54
3.1	Geographical context.....	54
3.1.1	Study period and area.....	54
3.1.2	Base geography.....	56
3.2	Derived 1 x 1km data	56
3.2.1	Regional variables	60
3.2.2	Land cover	61
3.2.3	Population.....	64
3.2.4	Transport networks	66
3.2.5	Topography	68
3.2.6	Light intensity	71
3.2.7	Meteorological data	75
3.2.8	Distance to sea	78
3.2.9	Derived data - summary statistics	79
3.3	Air pollution data.....	81
3.3.1	Information sources.....	81
3.3.2	Data limitations	83
3.3.3	Classification of site location	85
3.3.4	Spatial and temporal coverage	90
3.3.5	Training and validation data	94
3.4	Conclusion	95

4	POPULATION MODELLING.....	96
4.1	Introduction	96
4.1.1	Need for high resolution population data	96
4.1.2	Current population models	99
4.2	Methods.....	104
4.2.1	Data requirements	104
4.2.2	Modelling approach	105
4.3	Results	110
4.3.1	No-light versus light intensity models.....	111
4.3.2	Light intensity models.....	111
4.3.3	Model selection and mapping	116
4.4	Validation.....	120
4.4.1	ED1991 and OA2001.....	121
4.4.2	KM1991 and KM2001	121
4.4.3	Model comparison.....	121
4.5	Discussion.....	123
4.6	Conclusions.....	126
5	AFFINITY ZONE STRATIFICATION	127
5.1	Introduction	127
5.2	Methods.....	131
5.3	Analysis and results	134
5.3.1	Variable screening	134
5.3.1.1	Variable selection approach	135
5.3.1.2	Factor analysis approach	137
5.3.1.3	Combined approach	140
5.3.2	Cluster analysis.....	146
5.3.2.1	K-means cluster analysis	149
5.3.2.2	Hierarchical cluster analysis	155
5.3.3	Concordance analysis.....	158
5.3.4	Pollution characteristics of affinity zones.....	160
5.4	Pollution mapping by affinity zone	163
5.4.1	Procedure	164
5.4.2	Model validation.....	164
5.4.3	Results.....	165
5.5	Discussion.....	169
5.5.1	Creating air pollution affinity zones	169
5.5.2	Characterising pollution within affinity zones	171
5.5.3	Mapping concentrations with affinity zones	172
5.5.4	Assessing representativeness.....	173
5.6	Conclusions.....	174

6	MODELLING AIR POLLUTION USING MOVING WINDOW METHODS.....	175
6.1	Introduction	175
6.2	Methods.....	178
6.2.1	Data requirements	178
6.2.2	Focalsum modelling.....	185
6.2.3	Mapping concentrations.....	191
6.2.4	Model validation.....	191
6.2.5	Comparison with reference method.....	191
6.3	Results	195
6.3.1	NO ₂ models.....	195
6.3.2	Residual concentrations and traffic analysis.....	205
6.3.3	Other pollutant models	209
6.4	Discussion.....	217
6.4.1	Possible sources of error	217
6.4.2	Potential enhancements to the focalsum approach	221
6.5	Conclusion	223
7	GIS-BASED EXPOSURE ASSESSMENT - A DISCUSSION.....	224
7.1	Evaluation of the exposure modelling components	225
7.1.1	Modelling population density.....	226
7.1.2	Affinity zone stratification	229
7.1.3	Pollution modelling	239
7.2	Exposure assessment.....	243
7.2.1	Methods	243
7.2.2	European exposure to NO ₂	243
7.2.3	European exposure to PM ₁₀	246
7.2.4	Uncertainty in the exposure estimates	249
7.3	Exposure modelling using GIS.....	255
7.3.1	Place in epidemiology	256
7.3.2	The policy need	258
8	CONCLUSION	260
8.1	Future work	261
8.1.1	Population modelling	261
8.1.2	Affinity zone stratification	262
8.1.3	Focalsum modelling.....	262
8.1.4	The way forward	263
	REFERENCES.....	264

List of Figures

Figure 2.1	Trend in European NO ₂ and PM ₁₀ , 1997-2001	28
Figure 2.2	The variogram.....	45
Figure 3.1	Study area extent	55
Figure 3.2	Moving window example: Focalsum.....	60
Figure 3.3	Circular window weights used to calculate the regional variables.....	61
Figure 3.4	Total built-up land (% area).....	65
Figure 3.5	A-road density (km/km ²)	67
Figure 3.6	Reprojected, area-weighted altitude grid	69
Figure 3.7	Altitude (m)	70
Figure 3.8	Topex calculation.....	71
Figure 3.9	Slope-aspect calculations	71
Figure 3.10	1 x 1km light intensity	74
Figure 3.11	2001 annual mean temperature (°C).....	76
Figure 3.12	Measuring angles to calculate wind direction in ArcInfo.....	78
Figure 3.13	Distance to sea (m).....	80
Figure 3.14	AIRBASE sites with incorrect coordinates	85
Figure 3.15	Discrimination of site location by land cover.....	89
Figure 3.16	2001 annual mean NO ₂ concentration (µg/m ³)	92
Figure 3.17	2001 annual mean PM ₁₀ concentration (µg/m ³)	93
Figure 4.1	EU population density based on NUTS 5 regions	98
Figure 4.2	Area-weighting example.....	100
Figure 4.3	Population modelling procedure	106
Figure 4.4	Population density: national models (strategy 2).....	117
Figure 4.5	Scatter plot EU strategy 1 vs. strategy 2.....	118
Figure 4.6	Population density: EU model (strategy 1)	119
Figure 4.7	Residual population in Great Britain (EU strategy 1 model).....	124
Figure 5.1	Affinity zone stratification approach	133
Figure 5.2	Example input for the variable selection method	136
Figure 5.3	Example of blocky clusters.....	137
Figure 5.4	Principal components derived from full data set	140
Figure 5.5	Correlation net for potential variables used in AZS	143
Figure 5.6	Scree plot for climate domain.....	144
Figure 5.7	Correlation net for variables and factors used in AZS.....	146
Figure 5.8	Statistical index.....	150
Figure 5.9	Map granularity index.....	151
Figure 5.10	10 cluster solution (K-means CA).....	153
Figure 5.11	16 cluster solution (K-means CA).....	154
Figure 5.12	Example scree plots for hierarchical CA	156
Figure 5.13	12 cluster solution (HCA).....	157
Figure 5.14	Box plots of air pollutant concentrations by cluster	162
Figure 5.15	Comparison of baseline vs. kriging model	169

Figure 6.1	Emissions modelling protocol	181
Figure 6.2	Eight segments for wind direction	184
Figure 6.3	Basic focalsum procedure	186
Figure 6.4	Annulus and wedge windows for 13x13 neighbourhood	187
Figure 6.5	Validation stratified by urban and rural sites	201
Figure 6.6	Annual mean background NO ₂ (µg/m ³) - Reference method	202
Figure 6.7	Annual mean background NO ₂ (µg/m ³) - Focalsum	203
Figure 6.8	Residual NO ₂ (µg/m ³) at background sites	205
Figure 6.9	Residual NO ₂ (µg/m ³) by country and affinity zone.....	206
Figure 6.10	Modelled background vs. measured NO ₂ at traffic sites	207
Figure 6.11	Residual NO ₂ (µg/m ³) by country and affinity zone - traffic	208
Figure 6.12	Annual mean background PM ₁₀ (µg/m ³) - Reference method.....	213
Figure 7.1	NO ₂ sites by affinity zone population density (50km blocks).....	237
Figure 7.2	Variability of monitoring sites by affinity zone	238
Figure 7.3	Exposure to NO ₂ concentrations (µg/m ³) - modelled surface	244
Figure 7.4	Exposure to NO ₂ concentrations (µg/m ³) - nearest monitor.....	244
Figure 7.5	Cumulative population by NO ₂ exposure class - country.....	245
Figure 7.6	Cumulative population by NO ₂ exposure class - affinity zone.....	246
Figure 7.7	Exposure to PM ₁₀ concentrations (µg/m ³) - modelled surface	247
Figure 7.8	Exposure to PM ₁₀ concentrations (µg/m ³) - nearest monitor	247
Figure 7.9	Cumulative population by PM ₁₀ exposure class - country	248
Figure 7.10	Cumulative population by PM ₁₀ exposure class - affinity zone.....	249
Figure 7.11	Population exposure to NO ₂ concentrations (µg/m ³) - traffic.....	251
Figure 7.12	Residuals at background NO ₂ sites - Focalsum	253
Figure 7.13	Cumulative population distribution by pollution errors.....	254

List of Tables

Table 2.1	Air quality limits for the protection of human health.....	22
Table 2.2	Source sector emissions totals in the EU-15 for 2001	24
Table 2.3	Air pollution modelling methods	41
Table 3.1	Input data sets	58
Table 3.2	Derived variables	59
Table 3.3	Land cover classification	63
Table 3.4	EUROAIRNET site selection for assessment of population exposure.....	82
Table 3.5	AIRBASE definitions: site type and location	86
Table 3.6	Discrimination of monitoring sites and 1km cells into urban and rural...88	
Table 3.7	Urban and rural area based on discriminant analysis	90
Table 3.8	Number of NO ₂ and PM ₁₀ monitoring sites by year and site type	90
Table 3.9	Number of sites in 2001 with >= 75% data capture by country	91
Table 3.10	Monitoring sites by type and location	95
Table 4.1	Summary statistics for NUTS 5 census data (1991)	97
Table 4.2	Regression modelling strategy	108
Table 4.3	Regression statistics for land cover area (no-light) models	113
Table 4.4	Regression statistics for light intensity models	114
Table 4.5	Strategy 3 models (with light intensity).....	115
Table 4.6	Validation results for Great Britain	122
Table 5.1	Variables (by scale) used in AZS.....	142
Table 5.2	Total variance explained for climate domain	144
Table 5.3	Rotated principal component loadings	145
Table 5.4	Description of clusters (10 and 16 cluster solutions - K-mean CA)	152
Table 5.5	Description of clusters (12 cluster solution - HCA)	158
Table 5.6	Percent agreement between NO ₂ and PM ₁₀ results	159
Table 5.7	Percent agreement between HCA and K-means CA results.....	160
Table 5.8	One-way ANOVA results	162
Table 5.9	Pollutant concentration by cluster - training sites	166
Table 5.10	Simple kriging (Gaussian model) parameters - background NO ₂	167
Table 5.11	Model validation results across all affinity zones	168
Table 5.12	Model validation results by affinity zone	168
Table 6.1	Emissions modelling predictor variables by SNAP category	182
Table 6.2	Description of APMoSPHERE emissions grids	182
Table 6.3	Emissions grids, by pollutant, used in focalsum modelling.....	183
Table 6.4	Variables used in APMoSPHERE pollution modelling	193
Table 6.5	Reference method - NO ₂ background model.....	194
Table 6.6	Reference method - PM ₁₀ background model.....	194
Table 6.7	Reference method - SO ₂ background model	195
Table 6.8	NO ₂ model building.....	197
Table 6.9	Background NO ₂ modelling results	198
Table 6.10	Regression results - traffic component.....	208
Table 6.11	PM ₁₀ model building	211
Table 6.12	Background PM ₁₀ modelling results.....	212
Table 6.13	SO ₂ model building	215
Table 6.14	Background SO ₂ modelling results.....	216

Table 7.1 Site representativeness by affinity zone 235

Table 7.2 Percent 50km blocks containing NO₂ monitoring sites 236

Table 7.3 Kriging errors by affinity zone - background NO₂ 252

Table 7.4 Effect of uncertainty on background NO₂ exposure estimates..... 255

List of Appendices

- A1. Base coverage..... 277
- A2. ECMWF meteorological data 278
- A3. Sea distance..... 281
- A4. Summary statistics for derived data..... 284
- B1. Population modelling AMLs..... 289
- B2. Population modelling results..... 305
- C1. Cluster analysis - map granularity index..... 312
- D1. Weighted wind direction 314
- D2. Focalsum AMLs..... 318
- E1. Evaluation of the monitoring network - 50km analysis 322
- E2. Exposure assessment results 326

Abbreviations

ACS	American Cancer Society
AIRBASE	European Air Quality Information System
AirView	Air quality Visualisation instrument for Europe on the web
AML	Arc Macro Language
ANOVA	Analysis of variance
APHEA	Air Pollution and Health: a European Approach
APIS	Air Pollution Information System
APMoSPHERE	Air Pollution Modelling for Support to Policy on Health and Environmental Risk in Europe
AHSMOG	Adventist Health and Smog study
AZS	Affinity zone stratification
CA	Cluster analysis
CAFE	Clean Air for Europe
CI	Confidence interval
CIESIN	Center for International Earth Science Information Network
CLC2000	CORINE Land Cover 2000
CLC90	CORINE Land Cover 1990
CO	Carbon monoxide
CORINE	Coordination of Information on the Environment
DA	Discriminant analysis
DALY	Disability-adjusted life years
DM	Dispersion modelling
DMSP	Defence Meteorological Satellite Program
ECMWF	European Centre for Medium-Range Weather Forecast
ECRHS	European Community Respiratory Health Survey
ED	Enumeration district
EEA	European Environment Agency
EMEP	European Monitoring Evaluation Programme (Co-operative Programme for Monitoring and Evaluation of Long-Range Transmission of Air Pollutants in Europe)
EO	Earth observed
EOI	Exchange of Information Decision (Decision 97/101/EC amended by Decision 2001/752/EC)
EPER	European Pollutant Emission Register
EPIC	European Prospective Investigation into Cancer and Nutrition
EU	European Union - refers to the 15 Member States
AT	Austria
BE	Belgium
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GR	Greece
IT	Italy
IE	Ireland
LU	Luxembourg
NL	the Netherlands
PT	Portugal
SE	Sweden
UK	United Kingdom

EUROAIRNET	European Air Quality Monitoring Network
EXPOLIS	Air Pollution Exposure Distributions within Adult Urban Populations in Europe
F2	Factor-of-two
FA	Factor analysis
FB	Fractional bias
FMO	Figure-of-merit
GB	Great Britain
GIRAFE	Guide d'Information sur les Réseaux de qualité de l'Air Fonctionnant en Europe/Information System for Operational European Air Quality Monitoring Networks
GIS	Geographic information systems
GLC2000	Global Land Cover 2000
GMES	Global Monitoring for Environment and Security
GPW-3	Gridded Population of the World (version 3)
IDW	Inverse distance weighting
ITE	Institute of Terrestrial Ecology
JRC	Joint Research Centre
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
LAU	Local Administrative Unit
LOO	Leave-one-out
MANTLE	Mapping Night-Time Light Emissions
MARS	Monitoring of Agriculture through Remote Sensing
MEET	Methodologies for Estimating Air Pollutant Emissions from Transport
MSS	Mean sum of squared distance
MST	Minimum spanning tree
NAEI	National Atmospheric Emissions Inventory
Netcen	National Environmental Technology Centre
NI	Northern Ireland
NO	Nitric oxide
NO ₂	Nitrogen dioxide
NO _x	Oxides of nitrogen (NO + NO ₂)
NUTS	Nationales Unites Territoriales Statistiques
O ₃	Ozone
OA	Output area
ONS	Office for National Statistics (United Kingdom)
PCA	Principal components analysis
PM	Particulate matter (e.g. PM _{2.5} and PM ₁₀)
RMSE	Root mean square error
SABE	Seamless Administrative Boundaries of Europe
SAHSU	Small Area Health Statistics Unit - Imperial College London
SAVIAH	Small Area Variations in Air Quality and Health
SCA	Spatial correlation analysis
SD	Standard deviation
SEE	Standard error of the estimate
SNAP	Selected Nomenclature for sources of Air Pollution
SO ₂	Sulphur dioxide
SOI	Sphere of influence
STEMS	Space-Time Exposure Modelling System
TRAPCA	Traffic-Related Air Pollution on Childhood Asthma
TSP	Total suspended particulates
TSS	Total sum of squared distance

UDM-FMI
WHO

Urban Dispersion Modelling System - Finnish Meteorological
Institute
World Health Organization

1 Introduction

1.1 *Rationale*

Problems of air pollution have long been recognised, but it was only in the 20th century that serious concern arose about its possible effects on health. Several severe pollution episodes associated with fossil fuel combustion in the middle of the twentieth century (i.e. Meuse Valley, Belgium in 1930; Donora, Pennsylvania in 1948; and London in 1952 and 1962) provided clear evidence that high levels of air pollution can have adverse effects on human health (Colls 2002). Air pollution has, thus, been one of Europe's main policy concerns since the 1970s and remains a public health priority today.

As a result of this concern, the current focus of European Union (EU) air pollution policy is to develop and put into practice appropriate measures to improve air quality. This is being addressed through the Sixth Environment Action Programme where environmental health (including air pollution) is identified as one of the four main areas of effort. The Sixth Environment Action Programme aims to achieve levels of air quality that do not present unacceptable risks to human health and the environment. Currently the focus is on implementing air quality standards and improving coherency of air quality legislation and related policy. Clean Air for Europe (CAFE) is a programme designed to lead to the adoption of a thematic strategy on air pollution. It aims to develop a long-term, strategic and integrated policy for the protection of human health and the environment from the adverse effects of air pollution (Europa 2002).

The World Health Organization's global burden of disease initiative estimated that, in the year 2000, ambient air pollution caused almost 800,000 premature deaths worldwide. While the burden of disease due to air pollution predominantly occurs in developing countries, developed countries are not free from burden. In Western Europe, ambient air pollution accounted for an estimated 12,000 and 11,000 premature deaths in men and women, respectively. Ambient air pollution also attributed to an estimated 73,000 and 44,000 disability-adjusted life years (DALY) for men and women, respectively, where one DALY equals the loss of one year of healthy life (Ezzati et al. 2002; WHO 2002). Clearly, the health effects of air pollution cannot be ignored.

There have been numerous epidemiological studies related to air pollution, many of which have focused on exposure to particulates or the pollution mixture. Recent evidence has identified associations between air pollution and health in both short-term studies looking at day-to-day variations in pollution (de Hartog et al. 2003; Desqueyroux et al. 2002; Sunyer et al. 2003) and in longer-term studies examining the exposures over time (Dockery et al. 1993; Hoek et al. 2002a). Health effects are often related to respiratory symptoms and disease (Desqueyroux et al. 2002; Wolf 2002), and associations with cardiovascular disease have also been identified (Dockery et al. 1993; Hoek et al. 2002a; Maheswaran & Elliott 2003; Peters et al. 2001). The epidemiological evidence, however, has not been entirely definitive in terms of the levels of risk or the causative agents. One reason is that these studies have been hampered by problems of exposure assessment.

Hoek et al. (2002b), for example, argue that long-term epidemiological studies on air pollution effects have been held back by complications in characterizing the spatial variations in air pollution. Air pollution represents a complex mixture, and the spatial variations of this mixture are influenced by several factors including source activity, meteorology, topography and land use characteristics. While purpose-designed monitoring can be undertaken to help characterise the pollution surface for a particular area, this level of effort may be too time consuming or expensive within an epidemiological study. Furthermore, to assess long-term exposure, measurements recorded over a longer time-period are required. Such measurements may be unrealistic to obtain via purpose-designed monitoring. As a result, researchers often have had to rely on permanent monitoring networks. In general, the spatial coverage of these networks is often far from ideal as a basis for characterising the variations in pollution at a resolution suitable for individual-level exposure assessment. The sparse nature of these networks has meant that epidemiologists have had to rely on data from few monitoring sites at the city level or on indirect measures of exposure. The quality of the available monitored data can, thus, lead to exposure misclassification which may weaken the power to detect associations with health.

In an attempt to address exposure misclassification, measurements from the monitoring network can be enhanced by spatial modelling using geographic information systems (GIS). Recent modelling efforts in exposure assessment have characterised the spatial variations in air pollution at the city (Briggs et al. 1997; Liu & Rossini 1996) and country level (de Kluizenaar et al. 2001; Stedman et al. 1997).

The lack of a consistent modelling method in epidemiological studies, however, leads to difficulty in making comparisons between such studies and has contributed to continuing uncertainties about dose-response relationships and attributable health risks from air pollution. Additionally, the lack of reliable exposure data inhibits policy development, and monitoring of policy performance. To monitor policy performance better, therefore, models built using consistent methods and derived from a monitoring network spanning the whole area (e.g. the European air quality monitoring network - EUROAIRNET) would be very beneficial. Such consistent, trans-national data on air pollution are not only needed to guide European policy on air pollution, but could also be used as a basis for broad-scale epidemiological studies (with sufficient power to detect health effects). Studies designed to model air pollution at a geographic scale as large as the EU, encompassing a broad range of topography and land uses, however, are uncommon. GIS offer a powerful environment in which large data sets can be managed and manipulated to facilitate the development of high resolution models over vast areas with practicable computer processing demands.

As well as providing a means by which to enhance the data from the monitoring sites, GIS also offer useful tools for estimating exposures by intersecting the air pollution surfaces with population data. Further, using this approach, the population exposed to elevated annual mean concentrations can be visualised through maps. Such maps can be used to identify pollution 'hot-spots' and 'at-risk' populations in order to highlight where pollution control efforts are most urgently needed. Maps of the current situation, expressed either in terms of ambient concentrations or population exposures, also provide a baseline against which to compare potential or actual policy effects, and thus provide a basis for health risk assessment of policies. Maps which are easily updated as new input data become available will also allow the EU to track the progress of the Member States as they work toward achieving the target limits.

This study aims to use GIS techniques to model both the population distribution and air pollution concentrations as a basis for exposure assessment. The methods and detailed maps derived from this study are directly relevant to agencies concerned with air quality management in Europe and can strengthen the science base upon which policy is built. In its own right, the map-based exposure assessment presented here can help identify potential European populations 'at-risk' of exposure to air

pollution above guideline levels. Additionally, this study addresses the need for a common, reliable methodology and data set upon which future exposure assessments conducted across the EU may be based, to facilitate comparability between epidemiological investigations. This thesis focuses on the broad-scale exposure of the European population to the traffic-related air pollutants nitrogen dioxide (NO₂) and particulates (specifically PM₁₀).

1.2 Aims

The overall objective of this study is to develop, test and apply methods for generating EU-wide pollution and population density models and maps as a basis for assessing European exposure to air pollutants within a GIS.

The specific aims are:

- To develop an accurate, high resolution (ca. 1km²) EU-wide model and map of population density.
- To assess the spatial coverage of the European air pollution monitoring network as a basis for air pollution mapping and exposure assessment.
- To develop and apply a novel GIS-based interpolation technique for modelling and mapping small-area variations in annual average PM₁₀ and NO₂ concentrations across the EU, and compare the performance with currently accepted methods using a reserved set of validation data.
- To assess European exposure to annual mean concentrations of PM₁₀ and NO₂, on the basis of the pollutant and population density maps, and identify potential populations exposed to above guideline pollution levels.

1.3 The APMoSPHERE project

This research was conducted in parallel with the APMoSPHERE project. APMoSPHERE, Air Pollution Modelling for Support to Policy on Health and Environmental Risk in Europe, was a thematic project under the Global Monitoring for Environment and Security (GMES) initiative as part of the European Unions Fifth Framework Programme. The APMoSPHERE project aimed to compare different spatial models, including interpolation, stratification, Bayesian hierarchical models, and direct monitoring from satellites for the prediction of long-term air pollution across the EU as a basis for scientific research and policy support.

1.4 Thesis overview

The aims of this particular research will be addressed in the remainder of this thesis, organised as follows:

- Chapter 2 presents a literature review related to air pollution policy and epidemiology, exposure assessment and the 'state of the art' in air pollution modelling.
- Chapter 3 details GIS development and data sets used in the study.
- Chapter 4 focuses on the development of a high resolution population density model and map for Europe.
- Chapter 5 explores affinity zone stratification, based on multivariate statistical techniques, as a basis for modelling air pollution.
- Chapter 6 introduces a novel GIS-based modelling method, using moving window analysis, as a basis for modelling air pollution. The resulting maps are compared to those created by currently accepted air pollution modelling methods.
- Chapter 7 provides results of the exposure assessment, and discusses the challenges and opportunities of the GIS-based exposure assessment across large geographic areas.
- Chapter 8 presents a summary of the findings and concludes the thesis.

2 Literature Review

2.1 *Air pollution overview*

2.1.1 The policy context

The European Environment Agency (EEA) defines an air pollutant as:

a “substance in air that could, at high enough concentrations, harm human beings, animals, vegetation or material. Air pollutants may thus include forms of matter of almost any natural or artificial composition capable of being airborne. They may consist of solid particles, liquid droplets or gases, or combinations of these forms” (EEA 2006).

In the wake of the major air pollution episodes of the 20th century, measures were taken throughout much of the western world to reduce air pollution from industrial and domestic sources. In many ways these have had considerable success, both by controlling emissions and by dispersing pollution more widely (e.g. through the ‘tall stacks’ policy). Largely as a result, concentrations of traditional pollutants, such as sulphur dioxide (SO₂) and black smoke, have declined in most cities over the past few decades (EEA 2005b). Concerns about the health effects of air pollution, however, have not been fully abated as far-travelled particulates continue to contribute to major pollution episodes such as that which affected much of southeast England in March 1996 (Stedman 1997). The growth of road traffic has also led to increased concentrations of a wide range of pollutants including nitrogen dioxide (NO₂), fine particulate matter (PM), carbon monoxide (CO) and ozone (O₃). In association with these changes has come growing concern about their implication in a wide range of respiratory and cardiovascular diseases.

Given that approximately 75% of the European population live in urban areas (PRB 2005), and that city inhabitants are often exposed to elevated pollution levels, air quality remains an important policy initiative not only for the European Union (EU) but in many countries around the world. The World Health Organization (WHO 2000) and international governments have thus identified pollutants considered harmful to human health and adopted air quality guideline limits accordingly.

Air quality policy in the European Union is implemented through the “Council Decision 96/62/EC on air quality assessment and management” that came into force in 1996 (Official Journal of the European Communities 1996). Referred to as the Framework Directive, it aims to protect the environment and human health. To this end, it provides the umbrella legislation which includes the revision of existing legislation, and incorporates new air quality standards for previously unregulated pollutants. It also establishes a framework and timeline for addressing air quality at the national level through a series of daughter directives targeted at specific air pollutants, including: sulphur dioxide, nitrogen dioxide, particulate matter, lead, ozone, benzene, carbon monoxide, poly-aromatic hydrocarbons, cadmium, arsenic, nickel and mercury.

The first daughter directive sets air quality limit values for the protection of human health (Table 2.1) and the environment for five pollutants including NO₂, SO₂ and particulate matter (Official Journal of the European Communities 1999). Given that the time between exposure and the onset of adverse health effects can differ by pollutant, there are typically two sets of limits: (1) a short-term average (i.e. daily or hourly, with the number of allowable exceedances); and (2) a long-term average (i.e. annual mean). For some pollutants, health effects occur very quickly (acute effects) while for others cumulative exposures are required to cause an effect (chronic effects). Some pollutants, for example NO₂ and PM₁₀, have effects after both short and long exposures: thus, both air quality limits are defined. The long-term averages are relevant to this research.

Table 2.1 Air quality limits for the protection of human health

Pollutant	Value	Average time	Number of allowed exceedances	To be met by
EU Guidelines				
NO ₂	200 µg/m ³	1h average	< 19 hours/year	2010
	40 µg/m ³	annual mean	none	2010
PM ₁₀ (stage 1)	50 µg/m ³	24h average	< 36 days/year	2005
	40 µg/m ³	annual mean	none	2005
PM ₁₀ (stage 2)	50 µg/m ³	24h average	< 7 days/year	2010
	20 µg/m ³	annual mean	none	2010
SO ₂	350 µg/m ³	1h average	< 25 hours/year	2005
	125 µg/m ³	24h average	< 4 days/year	2005

Notes: Council Directive 1999/30/EC (Official Journal of the European Communities 1999)

Monitoring activities are undertaken across the EU to assess if, and where, the Member States are in compliance with these guidelines. As such, harmonisation of

monitoring data is considered key to facilitating the implementation of the daughter directives. Harmonisation is addressed through the EU Council Decision 97/101/EC amended by Decision 2001/752/EC on the Exchange of Information (EOI) (Garber et al. 2002). The EOI aims to provide a reciprocal exchange relating to data collected from the air pollution monitoring networks and stations in the Member States. Specifically, a European air quality monitoring network (EUROAIRNET) and information system (AIRBASE) were developed to provide background information suitable for conducting air quality assessments at the European scale, in order to inform policy decisions at the EU and Member State levels. The intent of EUROAIRNET is to: (1) enable exposure assessments based on monitoring alone; (2) enable exposure assessments based on monitoring and modelling; and (3) support cost-effective abatement strategies (Larssen et al. 1999). In line with the EUROAIRNET objectives, the research presented in this thesis uses spatial modelling to enhance the monitoring network data in order to assess European exposures to air pollutants.

2.1.2 Pollutants of interest

As already mentioned, there has recently been a substantial decline in sulphur dioxide emissions due to the move away from sulphur-based fuels and the introduction of emission controls for industry in Europe. Mainly owing to the increased number of motor vehicles and effects of long-range transport, attention has thus shifted to nitrogen dioxide and particulates. These pollutants - which derive from a variety of sources including power generation, industrial processes, agriculture and transport - are explored in this study. The total 2001 EU emissions for NO_x and PM₁₀ by major source sector (i.e. SNAP or Selected Nomenclature for sources of Air Pollution category) are outlined in Table 2.2.

Table 2.2 Source sector emissions totals in the EU-15 for 2001

SNAP Code	SNAP Category Description	NO _x		PM ₁₀	
		Gg	%	Gg	%
1	Combustion in Energy Production and Transformation	1,662	17	63	9
2	Combustion in Commercial, Institutional, Residential and Agriculture	509	5	56	8
3	Combustion in Industry	1,049	11	60	9
4	Production Processes	164	2	205	30
5	Extraction/Distribution of Fossil Fuels	0	0	35	5
6	Solvent Use	0	0	0	0
7	Road Transport	4,568	48	80	12
8	Other Transport and Machinery	1,561	16	12	2
9	Waste Treatment and Disposal	9	0	6	1
10	Agriculture, Forestry and Land Use Change	4	0	158	23
11	Nature and Other	0	0	0	0

Notes: Data from the European Monitoring Evaluation Programme (EMEP). These emissions data are based on officially reported emissions to the extent possible, but some of the officially reported data have been corrected and gaps filled (EMEP 2005).

Air pollution represents a complex mixture. Individual air pollutants, however, are often classified as either primary or secondary in nature. Primary pollutants are those emitted directly from a source, whereas secondary pollutants are formed in the atmosphere through chemical reactions. Oxides of nitrogen and particulate matter may occur as either primary or secondary pollutants. The nature, sources, distribution and typical concentrations of the pollutants across Europe are described next.

Oxides of nitrogen

The two main oxides of nitrogen are nitric oxide (NO) and nitrogen dioxide (NO₂), together known as NO_x. Oxides of nitrogen are emitted from natural sources such as volcanoes, bacterial action in soil and by lightning. To a larger extent, however, NO_x are emitted from man-made sources mainly due to fossil fuel combustion from stationary (e.g. power plants) and mobile (i.e. road, air and water transport) sources (Working Group on Nitrogen Dioxide 1997). As shown in Table 2.2, stationary and mobile combustion sources respectively contributed approximately 35 and 64 percent to the total EU-15 NO_x emissions in 2001 (EMEP 2005).

During combustion processes, NO_x are formed though oxidation with atmospheric nitrogen, primarily producing nitric oxide and a small proportion of nitrogen dioxide.

Most combustion systems typically comprise more than 90 percent NO with NO₂ making up the remainder (Equation 2.1). Further oxidation of NO to NO₂ is limited by the availability of ozone (O₃) (Equation 2.2). In urban areas, the emissions close to roads are dominated by NO (e.g. high NO/NO_x ratio) which combines with ozone to form NO₂. Downwind from the emissions source, the NO₂/NO_x ratio is higher, shifting the balance of Equation 2.2 so that NO₂ dissociates again to form ozone. The highest concentrations of ozone thus tend to occur in rural areas (Colls 2002).



Although NO₂ is not the major NO_x species in all areas, it is one of the most important urban air pollutants and most significant when it comes to human health effects. Consistent associations between NO₂ and health effects, however, have not been found (Working Group on Nitrogen Dioxide 1997). Asthmatics seem to be the most sensitive subgroup to the effects of NO₂ exposure, although individuals differ in their sensitivity (WHO 2000). Some evidence suggests short-term effects of NO₂ on mortality and hospital admissions. For example, an increase of 50 µg/m³ 1-hour maximum NO₂ concentration was associated with a 1.3% (95% confidence interval [CI] 0.9-1.8%) increase in the number of daily deaths in the APHEA (Air Pollution and Health: a European Approach) study (Touloumi et al. 1997). To date there are few epidemiological studies to estimate long-term NO₂ exposures likely to be associated with health risks. The Californian Adventist Health and Smog (AHSMOG) study is one example. This study, however, showed no statistically significant associations between NO₂ and chronic disease outcomes or change in severity of respiratory symptoms (Abbey et al. 1993) (further described in Section 2.2.3). More consistent associations between air pollution and health effects have been found for particulates which are described next.

Particulate matter

Particulate matter is a complex mixture of organic and inorganic substances, present in the atmosphere as both liquids and solids. Particulate matter may originate from natural sources, such as sea spray or volcanic emissions, or from man-made sources -

most commonly those related to traffic, power generation plants and other combustion sources. EMEP (2005) reports the most significant sources of particulate emissions in the EU-15 as transport (14%), agriculture (23%), combustion processes (26%) and production processes (30%) in 2001 (Table 2.2). Particulate matter emissions from road transport, thought to be the major source of PM_{10} in urban areas, derive from tyre and brake wear, the re-suspension of road dust and as direct emissions from vehicles (Technical Working Group on Particles 1997).

The chemical composition and size distributions of particulate matter varies according to the emissions source. Three main size fractions are often defined: nucleation mode, accumulation mode and coarse particles. The nucleation mode, or ultrafine particles, are smaller than $0.1\mu m$ in diameter and originate from local combustion sources such as vehicular traffic. They account for the majority of particles by number but only a few percent of the total mass of airborne PM. The accumulation mode, or fine particulates ($PM_{2.5}$), are between 0.1 and $2.5\mu m$ in diameter and are typically secondary in nature. They usually account for most of the particulate surface area and a substantial part of the total mass of airborne PM. They mainly derive from gas-to-particle conversion of atmospheric gases (e.g. inorganic and organic acids) and agglomeration of ultrafine particles (Colls 2002). Secondary particles such as nitrates and sulphates, for example, form as a result of photochemical oxidisation of NO_2 and SO_2 (Koutrakis & Sioutas 1996). Coarse particles are those with an aerodynamic diameter greater than $2.5\mu m$. This fraction comprises a large part of the aerosol mass and is mainly composed of primary particles such as those formed by mechanical processes and re-suspended dust from road traffic, agriculture and industry. Vehicular traffic is one of the major sources of ultrafine particles and is also responsible for the re-emission of coarse, deposited particles (Bayer-Oglesby et al. 2004).

Depending on their size, particles have different potentials to travel either long or short distances. Coarse PM has a lifetime of minutes to hours and can travel over approximately one to tens of kilometres. Fine PM is more conducive to long-range transport with a lifetime of days to weeks and can travel distances in the order of hundreds to thousands of kilometres (Spengler & Wilson 1996). The atmospheric lifetime of ultrafine particles in high concentrations is very short, but elevated concentrations observed in urban air can last for several hours (Bayer-Oglesby et al. 2004).

The particular fraction of interest in this study is PM_{10} . These are particles with a 50% cut-off diameter of $10\mu m$ and represent the thoracic fraction of inhaled particles that is deposited within the lung airways and gas-exchange region (Technical Working Group on Particles 1997). Numerous time-series studies have observed associations between particulate air pollution and health outcomes ranging from mortality, hospitalisation for respiratory and heart disease, aggravation of asthma, incidence and duration of respiratory symptoms, effects on lung function, and restricted activity (Dockery & Pope 1996). The acute effects of particulate exposure are well established for total non-accidental, respiratory, cardiopulmonary and cardiac daily mortality (Katsouyanni 2003) with a relative risk for daily deaths in the order of 1.0074 (95% CI 1.0062-1.0086) with a $10\mu g/m^3$ increase in PM_{10} (WHO 2000). More recent studies in Europe and the United States have shown similar relative risks (Katsouyanni et al. 2001; Samet et al. 2000). Guideline limits for protection against the effects of chronic exposure to particulates derive from two American cohorts. Dockery et al. (1993) reports relative risks for total mortality per $10\mu g/m^3$ increase in PM_{10} and $PM_{2.5}$ as 1.10 (95% CI 1.03-1.18) and 1.14 (95% CI 1.04-1.24), respectively. The corresponding estimate from Pope III et al. (1995) is 1.07 (95% CI 1.04-1.11) for $PM_{2.5}$. These two studies are further described in Section 2.2.1.

2.1.3 Temporal and spatial variations

Air pollution varies in both time and space. These variations are driven by several forces including: the distribution and activity of emission sources, meteorological processes influencing dispersion, and chemical processes such as the interaction of primary pollutants with sunlight to form photochemical smog.

The temporal trends in air pollution cover a range of time scales from short-term patterns influenced by meteorology to long-term changes, such as the general decline in sulphur dioxide with a change in energy consumption (Mage et al. 1996) and policy. Figure 2.1 presents the European trend in annual mean NO_2 and PM_{10} concentrations from 1997-2001, computed on the basis of data from EUROAIRNET and the EMEP monitoring networks. The figure indicates that NO_2 concentrations appear relatively stable over the five-year period while PM_{10} exhibits a slight decline; in 2001, some monitoring stations for both pollutants continue to exceed the annual mean guideline limit (i.e. $40\mu g/m^3$ for both pollutants) though the majority recorded average concentrations below this threshold.

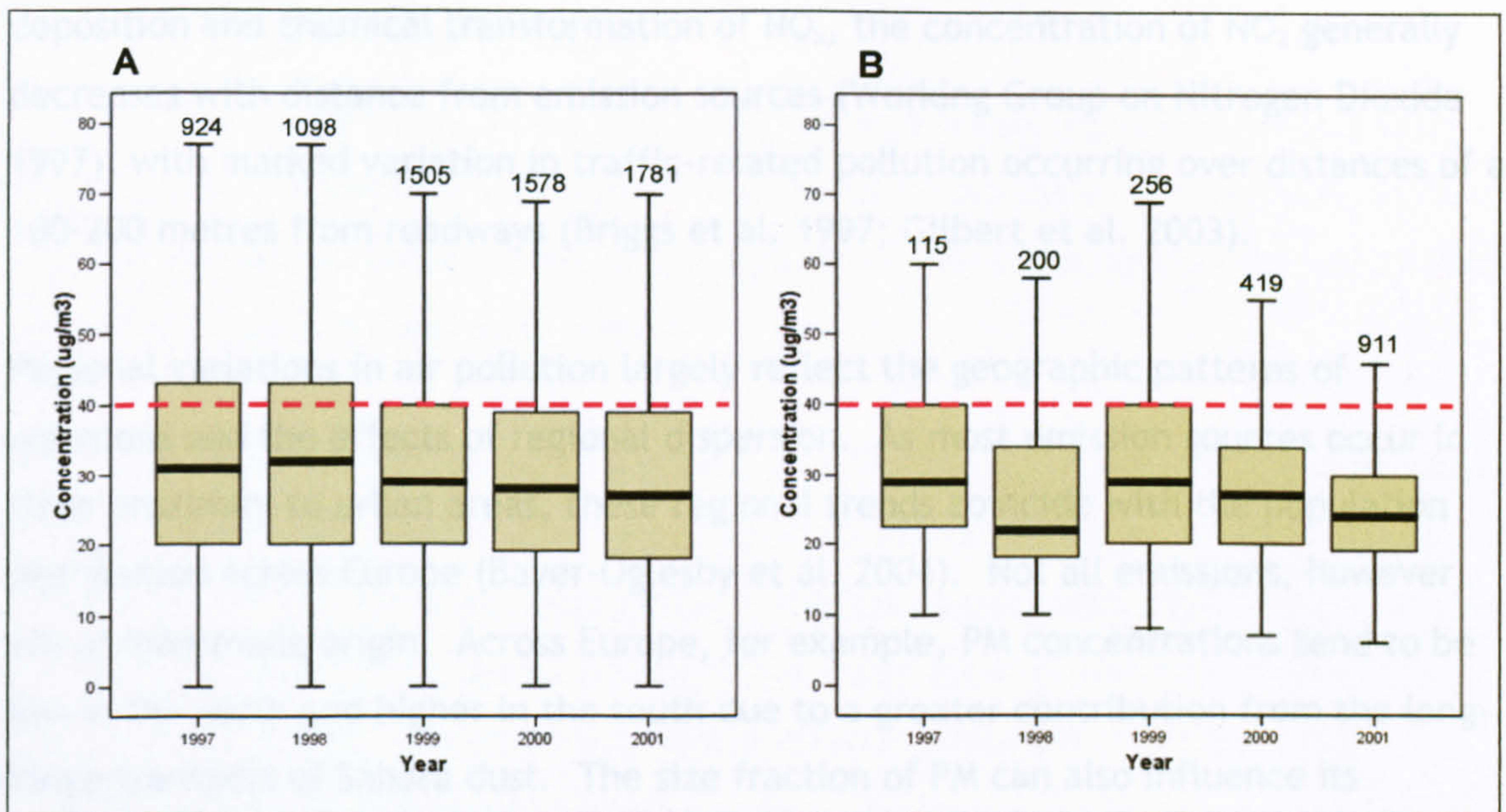


Figure 2.1 Trend in European NO₂ and PM₁₀, 1997-2001

Annual mean (A) NO₂ and (B) PM₁₀ concentrations (µg/m³) showing the median and interquartile range. Numbers represent the number of recording monitors, and the red line indicates the current air quality limit. Includes the EU-15 plus Norway, based on data downloaded from AIRBASE (ETC 2004) and EMEP (EMEP 2004).

There are also clear seasonal trends in air pollutants across Europe, related to changes in both source activity and meteorology. Emissions from power stations, for example, tend to vary according to fluctuations in energy demand throughout the year. While particle concentrations are generally higher in the south due to the favourable conditions for accumulation and photochemical reactions in the atmosphere, they often peak in the north during the winter heating season (Technical Working Group on Particles 1997).

At the finer temporal scale, daily and weekly patterns exist in cities influenced by regular cycles of traffic activity. The diurnal pattern for PM₁₀ and NO₂ closely follows the pattern of rush hours, and is characterised by peaks in the morning and early evening. On a weekly basis, there tends to be a pronounced decrease in traffic-related pollution during the weekend period when the number of commuters is reduced (Flemming et al. 2005; Ping Shi & Harrison 1997). There is also a clear spatial pattern of traffic-related pollutants in the urban outdoor environment, with the highest concentrations of NO_x and NO₂ generally along busy streets. In highly polluted areas, for example, NO can be oxidised by radicals through very quick reactions that can lead to high NO₂ concentrations at roadsides. Due to dispersion,

deposition and chemical transformation of NO_x, the concentration of NO₂ generally decreases with distance from emission sources (Working Group on Nitrogen Dioxide 1997), with marked variation in traffic-related pollution occurring over distances of a 100-200 metres from roadways (Briggs et al. 1997; Gilbert et al. 2003).

Regional variations in air pollution largely reflect the geographic patterns of emissions and the effects of regional dispersion. As most emission sources occur in close proximity to urban areas, these regional trends coincide with the population distribution across Europe (Bayer-Oglesby et al. 2004). Not all emissions, however, are of man-made origin. Across Europe, for example, PM concentrations tend to be low in the north and higher in the south due to a greater contribution from the long-range transport of Sahara dust. The size fraction of PM can also influence its distribution. Whereas fine particulates, for example, can be transported long distances leading to relatively uniform distributions over large areas, PM_{10-2.5} concentrations are generally more dependent on local sources, resulting in greater variability at the local scale (Houthuijs et al. 2001).

Pollutant concentrations are also influenced by the prevailing meteorological conditions. Broad-scale air mass movements and weather systems play an important role in determining long-range dispersion patterns of point source emissions. Many of the most severe episodes in northern Europe, for example, occur when upper tropospheric winds associated with anticyclonic systems transport pollution from industrial areas of central Europe and the Ruhr valley (Bayer-Oglesby et al. 2004).

The extent of dispersion is also influenced by the height of the emission source. Many industrial sources, for example, emit from tall stacks so that the pollutants are dispersed over long distances (i.e. tens or even hundreds of kilometres) before reaching the ground. Road transport, however, is a low-level source which means that the pollutants tend to be concentrated close to the traffic sources (Bayer-Oglesby et al. 2004).

This brief discussion highlights the importance of recognising the temporal and spatial variations in air pollutants when assessing exposures. As one might expect, exposures can change depending on when and where people spend their time.

2.2 Exposure assessment

Exposure is defined as “an event that occurs when a person comes in contact with the pollutant” (Monn 2001). Exposure to ambient air pollution is, therefore, determined by the concentration of the mixture of gases and particulate matter that people encounter in their daily lives.

The health effects of exposure to ambient air pollution have been the topic of many recent studies, only a small selection of which is mentioned here. Studies usually focus either on the temporal associations between air pollution and health or on the spatial associations. The distinction between these areas of research is that the former are generally concerned with acute effects, while the latter are more concerned with chronic health effects.

There have been many investigations on the short-term effects of air pollution, with emphasis on mortality and hospital admissions (Brunekreef & Holgate 2002). The most consistent associations are found for particulates, in particular the fine fraction (Laden et al. 2000; Neuberger et al. 2004; Samet et al. 2000; Schwartz 2000). All-cause, respiratory and cardiovascular mortality (Dockery 2001; Goldberg et al. 2001; Katsouyanni et al. 1997) have all been associated with ambient air pollution. Respiratory effects include exacerbation of asthma, increased hospital admissions, and decreased lung function (Boezen et al. 1999; Brunekreef et al. 1995; Dockery & Pope III 1994; Hoek et al. 1998). Specific cardiovascular effects that have been observed include myocardial infarction (Dockery & Pope III 1994; Peters et al. 2001), increased cardiac arrhythmias (Peters et al. 2000) and stroke mortality (Maheswaran & Elliott 2003; Maheswaran et al. 2005). Many studies focus on susceptible groups including the elderly, children or asthmatics. In particular, associations between increased pollution levels and increased respiratory symptoms in asthmatics have been well documented (Desqueyroux et al. 2002; Mortimer et al. 2002; Neuberger et al. 2004). Long-term studies have also identified associations with lung cancer and cardiopulmonary disease (Dockery et al. 1993; Nafstad et al. 2003; Pope III et al. 1995; Pope III et al. 2002).

For several reasons, assessing exposure to air pollution is not a straightforward task. The complexities derive from the temporal and spatial variability of air pollution as well as the nature of air pollution, which is a complex mixture. A further complication is the fact that humans are mobile and may move between various

microenvironments (e.g. home, office, outdoors, transit) as they go about their day. Each of the microenvironments has potentially different pollution characteristics - leading to potentially different exposure patterns.

The ideal way in which to assess exposure is, therefore, via direct personal monitoring for all the individuals in the population of interest. As one can imagine this can be very costly and is often unrealistic in epidemiological studies. As an alternative, the population may be split into smaller groups on the assumption that individuals within each group have similar exposures (e.g. all living near the same major source). Estimates of exposure can then be made for a representative sample of individuals within each group and the data extrapolated to unmonitored members. A large number of samples, however, might still be required to achieve a sample size large enough to provide reliable exposure estimates for the wider population (Nieuwenhuijsen 2003).

Given the limitations of personal monitoring for individuals or groups, studies related to air pollution generally rely on exposure modelling. The approaches may vary greatly in complexity. Exposures are often assessed simply by taking the measured concentrations from the nearest available monitoring sites (Section 2.2.1). In other cases, exposure proxies are used to reflect sources of air pollution - for example, traffic volume in the neighbourhood (Section 2.2.2). A third approach is to use mapping or modelling techniques to determine the exposure; for example, using air pollution models to calculate exposures for individuals or populations (Section 2.2.3). This study focuses on the use of GIS for exposure assessment in spatial studies, concerned with chronic effects. The following discussion related to measures of exposure thus highlights the studies that have been most influential in this context - those which are typically of cohort design.

2.2.1 Nearest pollution monitor

Early air pollution cohort studies assumed homogeneous exposure within cities and used the nearest monitor to assign exposures to individuals. A benchmark study is the Six Cities Study (Dockery et al. 1993) which examined mortality from a 14-16 year (between 1974-1991) follow-up of over 8,000 adults living in six American cities. Ambient concentrations of total suspended PM, SO₂, O₃ and suspended sulphates were measured in each community at a single central monitoring station and mean

concentrations were calculated for each city. Cox proportional hazards models, controlling for individual risk factors (i.e. smoking, sex and occupational exposure), were then used to examine associations of single pollutants with health outcome. The original study concluded that particulate air pollution was positively associated with lung cancer and cardiopulmonary disease mortality. In terms of all-cause mortality, a relative risk of 1.26 (95% CI 1.08-1.47) was found for the most versus least polluted cities.

In a similar but larger study Pope III et al. (1995) examined the health effects of particulate air pollution in a cohort of over 500,000 adults from the American Cancer Society (ACS) Cancer Prevention Study II. This study prospectively followed adults in 151 metropolitan areas between 1982 and 1989. The mean sulphate (for 1980, available for all 151 metropolitan areas) and the median $PM_{2.5}$ (for 1979-1983, available for 50 metropolitan areas) concentrations were calculated from the monitors in each metropolitan centre. Separate Cox proportional hazards models were then used to estimate lung cancer, cardiopulmonary disease and all-cause mortality while adjusting for confounding. Similar to the Six Cities Study, Pope III et al. (1995) concluded that particulate air pollution was associated with both cardiopulmonary and lung cancer mortality. The relative risk of all-cause mortality for the most versus least polluted areas was found to be 1.15 (95% CI 1.09-1.22) and 1.17 (95% CI 1.09-1.26) for sulphates and fine particulates, respectively.

Both the Six Cities and ASC II studies, which played an instrumental role in establishing the USA's National Ambient Air Quality Standard for $PM_{2.5}$, however, were subject to much debate. While some reviewers voiced concern that potential confounders had not been included, others said that the monitored data used did not adequately characterise the pollution over a long enough time period. This eventually led to an extensive reanalysis of the cohorts (Health Effects Institute 2000). The reanalysis showed that the mortality risk estimates for $PM_{2.5}$ and sulphates reported in the original studies were highly robust to alternative risk models with additional covariates. Models extended to incorporate a spatial random-effect to allow for spatial autocorrelation (i.e. that communities closer together may have more similar air pollution than those far apart) also supported the associations reported in the ACS II study.

2.2.2 Proxies for exposure

Rather than assuming homogeneous air pollution and assigning exposure on the basis of the nearest monitoring station, several cohort studies have used proxies for exposure. Proxy measures may be used to reflect distance from the pollution source. The assumption that exposures decline with increasing distance from source is often reasonable simply because many environmental pollutants tend to dilute as they are transported (Briggs 2005). Proxies for the level of source activity in the area of residence, for example traffic volume in the neighbourhood, may also be defined.

There are few examples of cohort studies using proxies for exposure. One example is provided by Hoek et al. (2002a) who used distance from road in combination with NO₂ data. In this study a random sample of 5,000 subjects from the Netherlands Cohort Study on Diet and Cancer were used to examine the association between mortality and individual-level traffic-related air pollutants from 1986 to 1994 (as part of the TRAPCA study - Traffic-Related Air Pollution on Childhood Asthma). The long-term exposure to the pollutants was estimated for the home address of the subjects. Regional and urban background concentrations of black smoke and NO₂ were interpolated from the monitoring stations using average concentrations for 1987 and 1990. The small-scale spatial variations in concentrations were also quantified by calculating proximity to major roads. Subjects living within 100m of a freeway or 50m of a major road were considered exposed. Persons exposed to freeway traffic were assigned exposures of 4.4 µg/m³ black smoke and 11 µg/m³ NO₂ while those exposed to major roads were assigned 13 µg/m³ black smoke and 8 µg/m³ NO₂. Using a Cox proportional hazards model, adjusting for potential confounders including smoking, Hoek et al. (2002a) found associations between living near a major road and all-cause mortality with a relative risk of 1.41 (95% CI 0.94-2.12). A stronger association with a relative risk 1.95 (95% CI 1.09-3.52) was found for cardiopulmonary deaths.

More often proxies for exposure have been used in other epidemiological designs (e.g. case-control, cross-sectional) to study, for example, the effects of traffic-related pollution on health. Both distance from busy roads (Livingstone et al. 1996; Maheswaran & Elliott 2003) and/or traffic density (English et al. 1999; Janssen et al. 2001; Wilkinson et al. 1999) have been used. These proxies, however, are often measured differently in the various studies. As one example, Maheswaran & Elliott (2003) used an ecological design to study the association between stroke mortality at

the census enumeration district (ED) level and living near major roads in England and Wales. The exposure proxy was calculated as the distance from the centroid of each ED ($n = 113,465$) to the nearest major road (i.e. motorways, primary roads and A-roads). Based on 1991 population denominators, stroke mortality from 1990 to 1992 for men and women ≥ 45 years of age was explored using log-linear Poisson regression adjusting for age, sex, socioeconomic deprivation, regional variation, urbanisation and metropolitan area. By using several distance thresholds (i.e. $<200\text{m}$, 200 to $<500\text{m}$, 500 to $<1,000\text{m}$ and $\geq 1,000\text{m}$), Maheswaran & Elliott (2003) found that the risks diminished with increasing distance from road. Living within 200m of a major road was associated with a 7% (rate ratio 1.07; 95% CI 1.04-1.09) and 4% (rate ratio 1.04; 95% CI 1.02-1.06) increase for men and women, respectively, compared to those residing more than 1km from a major road.

2.2.3 Exposure modelling

Both the nearest monitor approach and use of proxy measures have important limitations in terms of exposure assessment. Proxy measures, for example, can provide useful general indicators of relative exposures from specific sources or source types, and as such can be used to address public health questions such as 'Are people who live close to busy roads at greater risk of disease X?' By their nature, however, they are often extremely crude and inevitably non-specific in terms of the pollutants concerned. Problems also occur in standardising the measures because of the lack of any formal procedure for classifying the sources or defining the metrics to be used. Studies using distance from road as a proxy for exposures to traffic-related air pollution, for example, have used different distance thresholds or bands, and have classified roads in different ways (Hoek et al. 2002a; Livingstone et al. 1996; Maheswaran & Elliott 2003). As a result, studies may be difficult to compare.

Estimates based on local monitoring stations are in turn limited by the density (and thus representativeness) of the available monitoring network. This is seldom sufficient to reflect spatial variability of air pollution. Air pollution is routinely monitored by ground-based monitoring networks operated by municipal, regional or national governments. These networks are often designed to assess whether air quality standards have been exceeded rather than to provide a basis for evaluating health effects, and as such are not designed to describe spatial patterns of air pollution. In addition, monitoring stations provide information for specific point

locations at the time of sampling; over what surrounding area the measured values might be considered representative is rarely known. In many cases, however, air pollution gradients may be expected to be steep (Bayer-Oglesby et al. 2004; Janssen et al. 2001; Roorda-Knape et al. 1998), so that the representative area - what McGregor (1996) has called the 'affinity zone' - for any measurement station is likely to be small.

A very dense network of monitoring stations would therefore be required to describe the spatial and temporal variations in air pollution to a high degree of accuracy. Costs prohibit this level of monitoring on a routine basis, and even when purpose-designed monitoring is carried out as part of epidemiological studies, the number of monitoring sites is rarely adequate. The value of data from routine monitoring networks is also limited because the areas covered by monitoring and where people live are often different (WHO 1999). For an exposure assessment to be effective, therefore, the air quality measured at monitoring sites must somehow be translated to the resident population. In other words, we want to be able to predict exposures at unsampled locations, representing places or areas where people live or spend their time. As a result, there are obvious benefits in using some kind of modelling to improve the spatial and temporal resolution of concentration data obtained from the monitoring networks. The resulting pollution maps can then be intersected with population data (e.g. resident addresses or population density maps) to estimate exposures. Ideally, exposure assessments based on modelled pollution surfaces would be validated with personal monitoring, but where this is not possible the pollution surfaces themselves can be validated with independent concentration data. Section 2.3 focuses on the 'state of the art' in air pollution modelling, while this section focuses on exposure modelling - i.e. relating these pollution models to the population to assess exposures for large numbers of people at a local or individual scale.

Amongst others, the AHSMOG study (Abbey et al. 1993) provides a good example of exposure modelling. This study comprised a subgroup of 6,000 individuals in California studied over ten years to evaluate associations between ambient NO₂ concentrations and respiratory symptoms, myocardial infarction, cancer and all-cause mortality. Unlike the original Six Cities and ASC II studies, which assessed air pollution exposure at an aggregated city level, this investigation assessed exposure at the individual level. Cumulative NO₂ concentrations were estimated by month for

each participant by inverse distance weighting (IDW) the monthly average concentrations from the nearest three monitoring stations within 50km of the subjects home and work zip code centroids. The monthly values were then further cumulated to the time period before and after symptoms were assessed (e.g. 1973 through March 1977 and April 1977 through March 1987, respectively). Additional personal and ambient monitoring was undertaken for 650 subjects to develop regression equations by which to adjust the ambient measures obtained by interpolation. Analysis of the data, however, showed no statistically significant associations between NO₂ and the chronic disease outcomes, development of new respiratory cases or change in severity of respiratory symptoms. The authors identified the low levels of ambient NO₂ concentrations as a possible reason for the lack of positive associations.

Using the Firestone cohort, Finkelstein et al. (2003) examined the relationship between income, air pollution and mortality in 5,228 Canadians from the Hamilton-Burlington area of Ontario. The cohort comprised individuals referred to the Firestone Institute for Respiratory Health for pulmonary function testing between 1985 and 1999. Three-year mean SO₂ and particulate concentrations were interpolated by universal kriging from a network of 19 and 29 monitoring sites, respectively, and were then assigned to the subjects by their address postcode. The Cox proportional hazards model was used to model risk of death from non-accidental causes and cardiopulmonary disease. Although smoking was not included as a confounder, adjustments were made for age, sex, pulmonary function, body mass index and diagnosis of chronic disease. Compared to the most favourable group (i.e. high income level and low pollutant concentrations), subjects in other groups had higher risk of mortality. The relative risk for all-cause mortality for the low income and high pollutant level group was found to be 2.62 (95% CI 1.67-4.13) for particulates and 2.40 (95% CI 1.61-3.58) for SO₂. Relative risks for cardiopulmonary causes for the same group were 2.66 (95% CI 1.43-4.97) for particulates and 3.36 (95% CI 2.12-5.32) for SO₂.

A large European cohort conducted in Norway involved over 16,000 Oslo men that were followed over 27 years (1972/73 to 1998) (Nafstad et al. 2004; Nafstad et al. 2003). Individual-level measures of annual pollution were assigned to the home address of subjects. Yearly SO₂ exposures were determined by first using detailed models to calculate 1km SO₂ concentrations for 1979 and 1995. Emissions data and

observed concentrations were then incorporated to derive concentrations for the other years. Exposures to NO_x were modelled on the basis of the SO₂ concentration fields with NO_x emissions data, and an additional exposure variable based on annual average daily traffic estimates was assigned to addresses linked to the 50 busiest streets. Controlling for confounding, Cox proportional hazards models were used to estimate associations between four exposure categories as well as per 10 µg/m³ increase in SO₂ and NO_x exposure. Associations between mortality and increased exposure were only identified for NO_x. The relative risk for a 10 µg/m³ increase in NO_x between 1975 and 1978 was 1.08 (95% CI 1.06-1.11), while the corresponding adjusted risk ratios for respiratory disease, lung cancer and ischemic heart disease were 1.16 (95% CI 1.06-1.26), 1.11 (95% CI 1.03-1.19) and 1.08 (95% CI 1.03-1.12), respectively (Nafstad et al. 2004).

In addition to obtaining measures of exposure for cohort participants, exposure modelling is applicable for assessing exposures for larger populations over vast geographic areas. Such an exercise can be conducted on the basis of population density maps. Estimates of the total number of people at risk are often needed to guide or target policy decisions. This method can also be used to assess the health effects of concern which, while often quite small at the individual level, may be important at the population level due to the large numbers of people exposed. To facilitate exposure assessments of this scope, census information are often used in studies of ecological design.

Stedman et al. (1997) presents an example of a map-based approach to estimate the proportion of the UK population potentially exposed to annual mean concentrations of NO_x and NO₂ above the threshold levels in 1994 (only NO₂ is described here). First a map of rural annual mean NO₂ was calculated by bilinear interpolation of NO₂ diffusion tube measurements from 32 sites. This map was then calibrated with automatic monitoring sites, using linear regression, to incorporate the effect of local emissions on ambient concentrations. For each automatic monitoring site, the difference between the underlying NO₂ rural map and measured concentration was calculated and used as the dependent variable. The independent variables, NO_x emissions from major roads and percentage of urban plus suburban land cover, were calculated for five circular buffers (i.e. 1, 4, 25, 100 and 400km²). The resulting map was validated with several hundred diffusion tube measurements not used to derive the model ($r = 0.60$). The exposure assessment was conducted by overlaying census

data and the modelled pollution map. The analysis showed that an estimated 8% of the UK population in 1994 was potentially exposed to mean NO₂ concentrations exceeding the then current EU annual mean directive value of 26 ppb.

In ecologic studies, reliable information on the location and distribution of the population is essential for assessing exposures. While census data provide regularly updated information on population counts, the way in which census tracts are defined (or in some cases the large areas they cover) can limit the value of these data for assessing exposure. Area-weighting techniques can be used to redistribute population to a smaller area. This method, however, falsely assumes that the population is uniform across the census tract. More sophisticated models, therefore, are required to more appropriately estimate small-area population from the census information. In particular, a GIS-based approach to generate small-area estimates for EU population is explored in Chapter 4 of this thesis. A review of the current population modelling methods is also provided in that chapter.

While the complex temporal and spatial variations of air pollution are often considered in exposure modelling, one important concern with the exposure modelling approaches described thus far is that they typically assume that the population is static. This is obviously a false assumption. Although this topic is not explored in this thesis, it is certainly an important issue to consider in assessing human exposures to air pollution. The time-space interaction between the population and the pollution surface is briefly discussed next.

Time-activity models

We know that people generally move between different places throughout the day, and that these microenvironments are likely to differ in pollution levels. If an individual resides in the countryside and commutes to the city for work, for example, we would expect them to have higher exposures to traffic-related pollutants during their commute and while at the office compared to home. More sophisticated models that take account of movements through a dynamic pollution field can provide a better realisation of how individuals are exposed to variable concentrations (Briggs 2003).

In Europe, one of the major initiatives in this context has been the EXPOLIS (Air Pollution Exposure Distributions within Adult Urban Populations in Europe) project.

This is a multicentre study in which the time-activity approach is being explored to assess exposures. In Helsinki, for example, location information was recorded over 48 hours, at 15 minute intervals, by a sample of working age citizens to develop a database of time-activity for four microenvironments (i.e. home, workplace, traffic and other). Pollutant levels were simultaneously recorded by microenvironmental and personal monitors. Using the EXPOLIS Helsinki database, Hanninen et al. (2003) calculated time-weighted average exposure to PM_{2.5} for a large number of simulated individuals through random selection of input variables for time spent and concentration in the different microenvironments from specified probability distributions. Comparison between the simulations and the exposure distributions measured in the EXPOLIS study showed that the microenvironment model can provide reliable estimate exposure distributions given the required input data. Additional research based on the EXPOLIS database by Kousa et al. (2002) used an approach that involved dispersion models to calculate NO₂ concentration fields. In combination with information on the spatial location of the population and time-activity, the data were used to estimate the average exposure of the population with reasonable accuracy.

The GIS-based Space-Time Exposure Modelling System (STEMS) uses Hagerstrand's (1970 in Briggs 2005) concept of lifelines - paths through space and time that we take during our life, within which we encounter our life experiences - to model exposures to traffic-related air pollution. The system comprises a number of models including traffic and dispersion models that are calibrated against monitored data to derive high resolution, hourly pollution maps. The time-activity model then incorporates individual-level data derived from activity diaries (such as that from EXPOLIS) or simulated data based on national surveys to model the movement of individuals across the changing pollution surface to calculate exposure profiles (Briggs 2005; Gulliver & Briggs 2005).

While time-activity models such as EXPOLIS and STEMS are outside the scope of this research, they are a key advance in devising a more realistic representation of population exposures to air pollution. GIS-based methods for modelling the air pollution surface are discussed next.

2.3 Air pollution modelling methods

Various methods can be used to model and map air pollution. Table 2.3 presents an overview, including the advantages and disadvantages, of the current methods for predicting concentrations at unsampled locations as a basis for exposure assessment. These models will be broadly explored in the subsequent sections under the following headings: spatial interpolation; stochastic methods; and dispersion modelling. GIS methods are particularly useful for exposure modelling in the context of this thesis as they satisfy several key criteria linked directly to the study aims. In particular, GIS-based methods:

- are suitable for modelling at the small-area (ca. 1km²) scale, across the whole of the EU;
- can provide a basis for modelling different pollutants (i.e. particulates and NO₂);
- can facilitate the use of a combination of monitoring and covariate data to produce models on the basis of a sparse monitoring network;
- can be based on the raster data model which, given the size of the study area, practicably reduces the computer processing demands; and
- can be readily linked to population data to facilitate exposure assessment.

2.3.1 Spatial interpolation

Spatial interpolation involves predicting the concentration at unsampled locations using a sample set of measured points from within the same area or region. The main advantage of interpolation is that it allows the user to create a continuous air pollution surface on the basis of relatively few sample point locations, so long as the underlying spatial structure in the data can be adequately modelled (Burrough & McDonnell 1998c). Spatial interpolation is based on the so-called ‘first law of geography’ (or Tobler’s Law) - that things closer together tend to be more alike than things that are far apart (Tobler 1970). Estimates of concentrations at unsampled locations are, thus, based on information from the nearest available measuring sites, weighted in many cases in terms of their distance.

Interpolation can be based on either deterministic or stochastic (i.e. geostatistical) models. Deterministic methods apply predefined mathematical functions to interpolate a surface from the measured points, while stochastic methods incorporate randomness and use the statistical properties of the measured points to create the surface.

Table 2.3 Air pollution modelling methods

Description	Air Pollution Studies
Spatial Interpolation: Deterministic Methods - Global	
Global methods fit a single surface to the entire dataset (Burrough & McDonnell 1998a)	
Spatial Interpolation: Deterministic Methods - Local	
Local methods include Thiessen polygons, moving window methods (Burrough & McDonnell 1998a) distance weighting (IDW) and splines. In local methods, pollution surfaces in Paris using thin plate splines (Ionescu et al. 1993) local estimates are made on the basis of the nearest monitoring stations.	12 hour mean O_3 at address location for the Adventist cohort participants (Ionescu et al. 1993)
Spatial Interpolation: Geostatistical Methods	
Kriging is an optimal interpolator that produces unbiased estimates (Burrough & McDonnell 1998b) with a known minimum variance. The kriged prediction at a particular point is a weighted average, where the weights are selected to minimize the error variance.	Associations between TSP and mortality in Hamilton, Canada (Jerrett et al. 1996) 12 hour mean O_3 in Toronto using kriging (Liu & Rossini 1996) Methods in environmental epidemiology - PM_{10} example (Cressie 1991) Temporal variation of SO_2 over Istanbul, Turkey (Tayanç 2000) Port associations between income, air pollutants and mortality (Kawachi et al. 2003)
Stochastic methods: Regression Mapping	
A stochastic method that uses least squares regression to develop predictive models of the air pollution surface using monitored pollution data and explanatory variables.	concentrations at the city level in Europe (Briggs et al. 1997; Jerrett et al. 1999) and in the UK (Briggs et al. 2000) and $PM_{2.5}$ in the Netherlands and Sweden (Brauer et al. 2003) Brauer et al. 2003; Hochadel et al. 2006) in Toronto, Canada using a location-allocation and regression approach (Kanakoglou et al. 2005) pollution on the basis of traffic characteristics in Munich (Carr et al. 2003)
Stochastic methods: Bayesian Hierarchical Models	
Statistical method that models variations in air pollution over time and space. The models exploit the dependencies between pollutants and the approach assigns prior distributions to the unknown missing observations at each stage of the hierarchy.	O_3 episodes in the Lombardy Region of Italy (Maffei 1999) temporal and spatial associations between pollutants O_3, NO and SO_2 over 4 years using data from 8 London monitoring sites and between different pollutants monitored at different sites (Wakefield 2002) air pollution models (Koop & Tole 2003)
Dispersion Modelling (DM)	
A dynamic model of dispersion processes that incorporates information on emission sources and the factors that influence dispersion. Models are often developed as a function of distance from the source.	in exposure assessment (Colville & Briggs 2000) DM system to analyse air quality in Espoo, Finland (Valkonen et al. 2001) age population exposure using time activity data and predicted concentrations by DM (Kousa et al. 2002) model to assess school children's exposure in six French cities (Kousa et al. 2006) model to assess exposure in Auckland, Australia (Scoggins et al. 2001) model to assess historic exposure to air pollution in Stockholm (Nyberg et al. 2001; Nyberg et al. 2000) Norwegian men followed up for 27 years (Nafstad et al. 2004; Nafstad et al. 2000)

Table 2.3 Air pollution modelling methods

Description		Advantages		Disadvantages		Reference to Air Pollution Studies	
Spatial Interpolation: Deterministic Methods - Global							
Global methods fit a single surface to the entire data set.		Global methods can be quick assessments useful for the removal of spatial trends.		The meaning of trends may be unclear. Outliers and edge effects may distort the surface. Assessment of errors is limited.		• Review of methods (Burrough & McDonnell 1998a)	
Spatial Interpolation: Deterministic Methods - Local							
Local methods include Thiessen polygons, moving window, inverse distance weighting (IDW) and splines. In local methods a series of local estimates are made on the basis of the nearest sample points.		These methods have quick processing time. Moving window and IDW are good for interpolating from sparse data from regular or irregularly spaced samples.		Most methods provide no, or limited error assessment. Tessellation patterns depend only on the distribution of data points. Moving window results depend on the size of the window and choice of weighting parameter.		• Review of methods (Burrough & McDonnell 1998a) • Estimating NO ₂ pollution surfaces in Paris using thin plate splines (Ionescu et al. 2000) • Estimating NO ₂ at address location for the Adventist cohort participants (Abbey et al. 1993)	
Spatial Interpolation: Geostatistical Methods							
Kriging is an optimal interpolator that produces unbiased estimates with a known minimum variance. The kriged prediction at a particular point is a weighted average, where the weights are selected to minimize the error variance.		The method is easy to implement and has reasonable data requirements and processing time. It is more flexible in its application due to automatic estimation of the semi-variogram at each step.		Kriging performance is dependent on the presence of spatial autocorrelation (i.e. sites close together are more similar than those far apart). Kriging may not be suitable for small areas, such as cities, where the number of monitors is insufficient. A relatively large number of observations are needed to fit an appropriate semi-variogram.		• Review of methods (Burrough & McDonnell 1998b) • Study of associations between TSP and mortality in Hamilton, Canada (Jerrett et al. 2005) • Prediction of 12 hour mean O ₂ in Toronto using kriging (Liu & Rossini 1996) • Geostatistical methods in environmental epidemiology - PM ₁₀ example (Cressie 2000) • Spatial and temporal variation of SO ₂ over Istanbul, Turkey (Tayanc 2000) • Firestone cohort associations between income, air pollutants and mortality (Finkelstein et al. 2003)	
Stochastic methods: Regression Mapping							
A stochastic method that uses least squares regression to generate predictive models of the air pollution surface using a combination of monitored pollution data and explanatory variables.		Computationally simple. Greatly reduces costs and processing times compared to DM. Able to produce high-resolution maps across urban areas over long-term averaging periods (e.g. annual means). Unlike kriging, spatial autocorrelation of the observations is not required to make an accurate surface. Generally low cost - can be adapted to local areas that have existing monitoring.		Models can be case and area specific, and reliability cannot be guaranteed when extrapolating beyond the input data. Recalibration with local monitoring sites may be required to account for differences between areas in topography, meteorology and other factors influencing air pollution. Models cannot be applied to short-term averaging periods (e.g. hourly). This method depends on a large amount of geospatial data, and it also assumes that the predictor variables are free from measurement error.		• Modelling NO ₂ concentrations at the city level in Europe (Briggs et al. 1997; Collins 1998b) and in the UK (Briggs et al. 2000) • Modelling NO ₂ and PM _{2.5} in the Netherlands and Sweden (Brauer et al. 2003) and Germany (Brauer et al. 2003; Hochadel et al. 2006) • Modelling NO ₂ in Toronto, Canada using a location-allocation and regression mapping approach (Kanakoglou et al. 2005) • Modelling air pollution on the basis of traffic characteristics in Munich (Carr et al. 2002)	
Stochastic methods: Bayesian Hierarchical Models							
Statistical method that models variations in air pollution using temporal and spatial associations between pollutants at different sites and between different pollutants monitored at the same site. The models exploit the dependencies between pollutants. Bayesian approach assigns prior distributions to the unknown parameters and missing observations at each stage of the hierarchy.		These models can pick up daily variation in air pollution and can handle time series with missing data. The models address inadequacies in the observed data and allow information from multiple sites on different pollutants to be combined. They can estimate uncertainty at any location (monitored or not), and also allow for the incorporation of prior information about the model parameters.		These models are computationally intensive.		• Forecasting CO episodes in the Lombardy Region of Italy (Maffei 1999) • Model PM ₁₀ , CO, NO and SO ₂ over 4 years using data from 8 London monitoring stations (Shaddick & Wakefield 2002) • Evaluation of air pollution models (Koop & Tole 2003)	
Dispersion Modelling (DM)							
A dynamic model of dispersion processes that incorporates information on emission sources and the factors that influence pollution. Models are often developed as a function of the Gaussian dispersion curve.		Models can be quite easily adapted to new pollutants or areas, without the need for additional monitoring. Can be applied to short-term averaging periods. Can be linked to GIS enabling the joint analysis of information from the monitoring system and population distribution. Have the potential to incorporate both spatial and temporal variation of air pollution within a study area without need for dense monitoring network.		Box models have relatively simple data requirements and processing efficiency, but are only capable of producing broad-scale pollution maps (i.e. high-resolution mapping over large areas is impractical). Gaussian models may gain in spatial accuracy, but are demanding in terms of their input data and processing. DM rely heavily on the dispersion environment and meteorological data. Substantial data requirements and computer processing demands needed to map a city. This method is not suitable when emissions data are sparse. Care is needed when applying DM to exposure assessment, as most models are developed for regulatory purposes, therefore, extensive crossvalidation with monitored data is necessary.		• DM applications in exposure assessment (Colville & Briggs 2000) • Using an urban DM system to analyse air quality in Espoo, Finland (Valkonen et al. 1996) • Evaluate average population exposure using time activity data and predicted NO ₂ concentrations by DM (Kousa et al. 2002) • STREET 5 model to assess school children's exposure in six French cities (Penard-Morand et al. 2006) • Urban airshed model to assess exposure in Auckland, Australia (Scoggins et al. 2004) • AIRVIRO model to assess historic exposure to air pollution in Stockholm (Bellander et al. 2001; Nyberg et al. 2000) • Cohort of Norwegian men followed up for 27 years (Nafstad et al. 2004; Nafstad et al. 2003)	

Deterministic methods

Two types of deterministic method can be usefully identified: global and local interpolators. Global interpolators, such as global polynomial interpolation in ArcGIS, fit a smooth surface defined by a single mathematical function (i.e. a polynomial) to the input data points across the study area. The resulting surface changes gradually to capture the coarse-scale patterns in the data. Global polynomial interpolation is inexact in that the surface does not pass through the actual measured points, instead least squares regression is used to determine the best fit (ESRI 2001). Global interpolators are mainly used for examining, and possibly removing, the effects of global variations or trends in the data (Burrough & McDonnell 1998a).

Local interpolators, on the other hand, calculate predictions only on the basis of the nearest measurement points. This is done by defining and using the input data points within a specified search radius around the pixel (or grid cell) to be predicted. The window is then moved to another unmeasured location and the interpolation repeated until all pixels across the study area are predicted. Examples of local interpolators include Thiessen polygons, inverse distance weighting (IDW) and exact and thin plate splines (Burrough & McDonnell 1998a).

Thiessen or Voronoi polygons, for example, define individual areas of influence around each of the input data points. They are essentially polygons whose boundaries define the area that is closest to each point relative to all other points. This property is mathematically defined by the perpendicular bisectors of the lines between all input data points. IDW is more complex in that it combines the idea of proximity with the gradual change of the trend surface. IDW relies on the assumption that the value (e.g. of air pollution) at an unmeasured location is a distance-weighted average of the input data points within a specified window (Burrough & McDonnell 1998a). The points closest to the prediction location will, therefore, have greater influence on the predicted value than points that are further away. The general formula is given in Equation 2.3 (ESRI 2001).

Equation 2.3
$$\hat{Z}(s_0) = \sum_{i=1}^N \lambda_i Z(s_i)$$

Where: $Z(s_0)$ is the value being predicted for location s_0 ; N is the number of measured data points in the window; λ_i are the weights assigned to each measured point; and $Z(s_i)$ is the observed value at location s_i .

Splines are piece-wise polynomial functions that are fitted exactly to a small number of points while, at the same time, ensuring that the joins between different parts of the curve are continuous. Thin plate splines can be used to remove artefacts (i.e. excessively high or low predicted values) resulting from natural variation and measurement error in the input data. Thin plate splines in effect replace the exact spline surface with a locally smoothed average which passes as close as possible through the data points (Burrough & McDonnell 1998a). Splines are often useful for calculating smooth surfaces from a large number of input data points and often produce good results for gently varying surfaces such as elevation (ESRI 2001).

Several studies have used deterministic methods to predict air pollution concentrations. Abbey et al. (1993), for example, used IDW to derive NO_2 concentrations at the address of each Adventist cohort participant from the nearest three monitoring stations (Section 2.2.3). In a methodological study, Ionescu et al. (2000) proposed the use of thin plate splines to estimate NO_2 concentration fields in Paris. They concluded that thin plate splines produced reliable concentration fields.

One major weakness of deterministic methods is that they do not provide estimates of uncertainties. Geostatistical methods, however, can estimate both the predicted values and their standard errors and are discussed next.

Geostatistical methods

Geostatistical methods, popularly known as kriging, are a range of techniques that utilise the underlying structure of the spatial correlation in the data to inform the interpolation (Burrough & McDonnell 1998c). Kriging models can take several forms which vary in complexity.

Kriging was originally developed for use in the mining industry, and has more recently been applied to other environmental applications. Kriging is an optimal interpolator that uses the spatial configuration and variance of the known data points

to determine the interpolation weights and search radii to provide the best, unbiased estimate at an unsampled point (Burrough & McDonnell 1998c). It is also considered a local interpolator as it uses the nearest data points in making a series of estimates to establish the surface (Briggs et al. 1997). It attempts to divide the spatial variation of a variable into three components: deterministic variation or trend; a random but spatially correlated component (regionalised variable); and uncorrelated noise. Kriging depends on the regionalised variable theory which assumes that the spatially correlated component is homogeneous (i.e. stationary) throughout the surface (Burrough & McDonnell 1998c). The performance of kriging relies on the presence of spatial autocorrelation - meaning that sites which are close together tend to be more similar than those which are further apart (Dien & Comrie 2002) - a further expression of Tobler's 'first law of geography'.

The nature of the regionalised variable is described by the variogram function which is used to optimise the interpolation weights and search radii. The variogram is created from the sample data by plotting the semivariance, $\gamma(h)$, against the distance between the pairs of samples (h) (Equation 2.4). For a stationary spatial process, the variogram would have a distinctive shape as shown in Figure 2.2. Different models (e.g. spherical, Gaussian, exponential and linear) can be fitted to the variogram.

$$\text{Equation 2.4} \quad \hat{\gamma}(h) = \frac{1}{2n} \sum_{i=1}^n \{z(x_i) - z(x_i + h)\}^2$$

Where: n is the number of paired observations of the values of attribute z separated by distance h ; and h is the sample spacing, also called the lag.

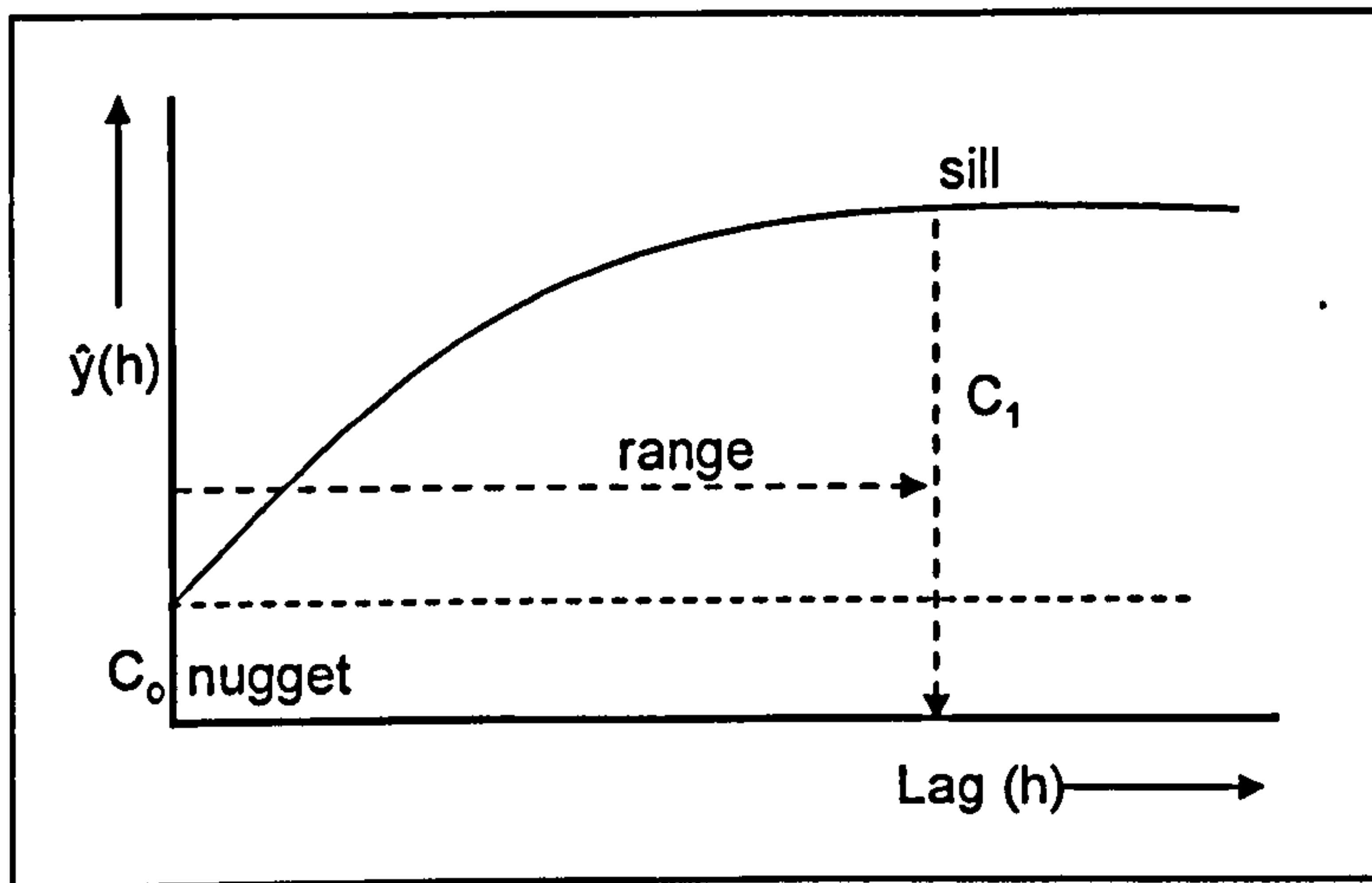


Figure 2.2 The variogram

At large values of lag (h) the variogram tends to level off at what is referred to as the sill. The sill indicates values of lag (i.e. the range) beyond which there is no spatial dependence between the data points. Within the range, spatial autocorrelation exists, so that sites close together have a smaller semivariance. Theoretically, points at the same location would have the same value making the semivariance equal to nil. In practice, however, points that are physically close together can often have quite dissimilar values. This is known as the nugget effect and represents the spatially uncorrelated noise or the variation at a scale finer than the resolution of the measurement grid (i.e. the model does not go through the origin). The data are said to be isotropic if the spatial dependency is only on the distance between the samples and not on direction. When the variation is anisotropic, variograms can be calculated in different orientations to explore the directional effect (Burrough & McDonnell 1998c).

Using the information from the variogram an empirical model can be derived (Equation 2.5). The fitted variogram is used to determine the weights, λ_i , required for local interpolation. The weights are chosen so that the estimate, $z(x_0)$, is unbiased and the estimation variance is less than it would be for any other linear combination of the observations (Burrough & McDonnell 1998c).

$$\text{Equation 2.5} \quad \hat{z}(x_0) = \sum_{i=1}^n \lambda_i \times z(x_i)$$

Where: λ_i are the weights derived from the variogram.

This model (Equation 2.5) applies to ordinary, simple and universal kriging which differ in their assumptions about the mean: ordinary kriging assumes a constant, unknown mean; simple kriging assumes a constant, known mean; and universal kriging assumes a trending mean. Additionally, data are often available for more than one variable at the sampling location. If the secondary variables are correlated with the primary variable being modelled (and are known at the unsampled locations) they may be used to help predict conditions at unsampled locations by co-kriging (Burrough & McDonnell 1998c). Covariates can be included into each of these methods, making ordinary co-kriging, simple co-kriging and universal co-kriging possible.

The use of kriging has been greatly enhanced in recent years through the development of GIS, but as yet they have found only limited application in air pollution modelling. Cressie (2000), for example, used kriging to model PM_{10} exposure. He argues that geostatistical methods are well suited to analysis involving monitored data distributed over an area of interest when the goal is to obtain a map of environmental exposure over the entire region. Recent air pollution studies using kriging tend to use ordinary or simple kriging, but advanced options that allow for more complex structure in the data (e.g. universal kriging and co-kriging) possibly have greater potential for modelling urban air pollution.

Finkelstein et al. (2003), for example, used universal kriging to derive SO_2 and particulate concentrations at the place of residence for the Firestone cohort subjects on the basis of the monitoring stations (19 and 29, respectively) recording over a three year period (Section 2.2.3). In another Canadian study, Jerrett et al. (2005) used kriging to generate a 10 year mean total suspended particulate (TSP) surface for Hamilton, Canada on the basis of 23 monitoring stations. The surfaces were then overlaid onto population-weighted census tract centroids to look for associations between particulate air pollution and mortality, while adjusting for social, demographic and lifestyle confounders. The overall results suggested that intra-urban variations in particulates were significantly associated with premature, all-cause, cardiovascular and cancer mortality. In a study by Liu & Rossini (1996), kriging was used to predict outdoor ozone concentrations in Toronto, Canada. The model performance was evaluated by comparing actual outdoor ozone measurements at 40 homes with the predictions based on 19 ambient monitoring sites. Kriging was generally found to overestimate the outdoor concentrations because the monitoring

stations were not randomly sited and tended to be in areas of higher pollution. When compared to predictions achieved by assigning the concentration from the closest monitoring station, however, the kriging predictions were found to be more accurate.

In the European context, Tayanc (2000) used kriging to examine the spatial and temporal variation in SO₂ concentration in Istanbul to evaluate air pollution during the heating seasons. The spatial patterns were assessed using monthly SO₂ measurements from 16 daily monitoring stations. The kriged surfaces for each heating season were used to identify critical regions in terms of elevated SO₂ levels and also provided insight into the temporal trends. One of their findings, for example, was that SO₂ concentrations during the 1995-96 heating season were considerably lower than the previous two seasons.

As part of the SAVIAH (Small Area Variations in Air Quality and Health) project, associations between SO₂ exposures and respiratory symptoms and disease in children were explored in Prague and Proznán (Pikhart et al. 2001). Kriging was used to generate a 10 x 10m map of long-term winter SO₂ levels in the two cities on the basis of a dense network of passive samplers measuring two 2-week periods. From this map, the mean SO₂ concentrations at the home and school were computed for each child in the study. The prevalence of respiratory outcomes associated with a 50 µg/m³ increase in SO₂ was similar in both cities. After controlling for confounders, the pooled results showed that increased SO₂ was associated with wheezing in the past 12 months (odds ratio 1.32; 95% CI 1.10-1.57) and doctor diagnosed asthma (odds ratio 1.39; 95% CI 1.01-1.92).

The studies mentioned above based their interpolation solely on the monitored data (i.e. the distance and covariance in concentrations between sites). The density of the monitoring network and (often) large distances between sites can be limiting factors in describing the pollution surface with these methods, especially if there is large variation in air pollution. As a result, the interpolated surfaces tend to be overly smoothed. It could be advantageous, therefore, to include covariates as proxy measures of emissions and dispersion to enhance the predictions. In particular, co-kriging and regression mapping may strengthen the interpolation by including variables that reflect the more local variations in air pollution (Briggs 2005). Stochastic methods including regression mapping are reviewed next.

2.3.2 Stochastic methods

Stochastic or probabilistic methods have a foundation in statistical theory. They are mathematical models which take into consideration the presence of randomness in one or more of the model parameters. The geostatistical methods described above could also be classified in this group of methods. This section focuses on two additional stochastic methods, specifically regression mapping and Bayesian hierarchical modelling.

Regression mapping

Regression analysis is a statistical technique used to analyse the relationship between a single dependent variable and several independent variables. Known values of the independent variables can then be used to predict the dependent variable.

Regression mapping uses least squares (or weighted least squares) regression to generate models of the pollution surface based on factors affecting variation in the pollution (e.g. emission sources, land cover, topography) and measured pollutant concentrations. Regression mapping, thus, depends on the availability of predictor variables measured at every spatial location (Diem & Comrie 2002). In regression mapping, the regression equation is developed on a sample set of data, referred to as the training sites, and is then applied to the dependent variables to predict pollution concentrations at the unsampled locations. Model validation is accomplished by comparing the predictions and actual measurements at a reserved set of validation sites.

Regression models have been used to predict pollution surfaces in a number of air pollution and exposure assessment studies. Regression mapping is the foundation of the SAVIAH model which was used to predict NO₂ as a marker for urban traffic-related pollution in Amsterdam, Huddersfield and Prague (Briggs et al. 1997; Collins 1998b). The resulting regression equations, which included traffic volume, land use and altitude as explanatory variables, were used to map NO₂ across the study areas. The accuracy was then evaluated by comparing predicted concentrations with monitored levels at independent validation sites. Regression mapping was also evaluated against dispersion modelling and spatial interpolation methods including contouring, kriging and trend surface analysis. The SAVIAH model was found to provide the best model with an R² of 0.79-0.87 between the predicted and observed concentrations across eight to ten validation sites. Briggs et al. (2000) later recalibrated and applied the SAVIAH model with local data from four cities in the UK.

The R^2 for validation sites was between 0.58-0.76, which was comparable to those achieved by dispersion models.

The SAVIAH methods were extended in the TRAPCA study exploring effects of exposure to traffic-related air pollution in the development of inhalant allergy, asthma and other chronic respiratory conditions in children (Brauer et al. 2003). Regression models were used to relate annual average concentrations measured at approximately 40 sites each in the Netherlands, Germany and Sweden to the explanatory variables calculated in GIS at different spatial scales (i.e. traffic intensity and population density within 50, 250 and 1,000 metre buffers). The authors found that including non-GIS variables, such as whether or not the monitor represented a traffic site or was in a street canyon, improved the model R^2 in all three locations (i.e. from $R^2 = 0.73$ to 0.78 for the Netherlands; $R^2 = 0.56$ to 0.76 for Germany; and $R^2 = 0.50$ to 0.63 for Sweden). In a follow up study, Hochadel et al. (2006) explored further enhancements to the model for Germany by expanding the buffers in which the explanatory variables were calculated to a maximum radii of 10 kilometres. They again used linear regression to assess the dependency of pollutant concentrations on the traffic-related and population based explanatory variables. The models were found to explain a high proportion of the NO_2 and $\text{PM}_{2.5}$ variation with an R^2 equal to 0.90 and 0.82, respectively. The best models contained a combination of the explanatory variables represented at both large and small spatial scales.

A study by Kanaroglou et al. (2005) also used regression mapping to characterise NO_2 in Toronto, Canada for epidemiological research. With 16 fixed monitoring stations, they first used regression modelling to generate an initial pollution surface on the basis of proximate land use and traffic variables. This surface was used as input into a location-allocation approach to create a surface of 'monitoring demand' to optimally site 100 NO_2 passive samplers in areas exhibiting greater population and pollution variability. Following a monitoring campaign with the optimally located passive samplers (95 of which produced valid results), a second land use regression model and final pollution surface (R^2 of 0.63) were developed.

In another study to characterise traffic-related air pollution, Carr et al. (2002) used regression analysis to determine whether traffic characteristics alone could adequately predict benzene, NO_2 and soot at homes in Munich, Germany for an



epidemiological study. Pollution data were collected from 18 city and 16 school monitoring sites and mean annual concentrations modelled using road characteristics as explanatory variables. The best model accounted for 77% of the variation in air pollution concentrations on the basis of traffic counts within 300m of the site and the percentage of 'stop-go' or jammed traffic.

Bayesian hierarchical modelling

Bayesian hierarchical modelling is a statistical method that uses the temporal and spatial associations between pollutants at different sites, and between different pollutants monitored at the same site, to model variations in air pollution. The Bayesian approach assigns prior distributions to the unknown parameters and missing observations at each stage of the hierarchy. Bayesian methods are generally applied to time-series data to address inadequacies or missing values in the observed data by allowing information from multiple sites on different pollutants to be combined (Shaddick & Wakefield 2002). In the APMoSPHERE project, for example, detailed geographic covariate information was incorporated into the hierarchy to model the spatial effect of air pollutants. A major advantage to the Bayesian approach, similar to kriging, is that it can estimate uncertainty at any location. The main drawback, however, is that the models are computationally intensive.

2.3.3 Dispersion modelling

Dispersion modelling involves constructing a dynamic model of dispersion processes, taking account of the main processes that influence the pollution concentration field (Briggs et al. 1997). These models use mathematical equations representing atmospheric dispersion and transport of pollutants from source to sink in a particular geographic area (Colville & Briggs 2000). A wide range of different models, based on different principles, are available including Eulerian, Gaussian and Lagrangian models. Eulerian models, for example, numerically solve the atmospheric diffusion equation, Gaussian models are built on the normal probability distribution of wind vector fluctuations, and Lagrangian models represent the process by the dispersion of fictitious particles in a moving airmass (Colls 2002). For short-range applications (i.e. where high resolution data are needed at a local scale), Gaussian models tend to be preferred (Colville & Briggs 2000).

A wide range of Gaussian dispersion models (and derivations from these) have been developed over recent years (Colvile & Briggs 2000). Most of these have been motivated by the need to model concentrations to inform air pollution action plans and remedial strategies (CERC 2006; Colvile et al. 2002) for regulatory purposes. The use of dispersion models in epidemiological studies is less common, partly perhaps because of their substantial data requirements and/or computer processing demands, but partly also because of lack of knowledge (or suspicion about their validity) amongst researchers. In a study conducted in 2000, however, Penard-Morand et al. (2006) used the STREET 5 dispersion model to estimate ambient traffic-related air pollution at a major road outside 108 schools in six French cities. They found considerable differences in annual mean concentrations in street segments outside the schools, as well as in the six cities, implying large contrasts in the children's exposure to traffic-related pollution. Modelled concentrations in excess of the European guideline levels for annual mean NO₂ and PM₁₀ were found respectively at 64% and 29% of the schools.

Other studies have used dispersion modelling as a basis for area-wide pollution modelling typically at the city level. Scoggins et al. (2004), for example, applied the CALGRID model to derive 3km grids of hourly NO₂ concentrations for Auckland, New Zealand. Model performance, determined by calculating the index of agreement between the modelled surface and monitored data at validation sites, was found to be good. The modelled NO₂ concentrations were thus averaged over 1999 and used in an epidemiological study that found a positive association between air pollution and mortality at the census area unit level.

In a separate study, the UDM-FMI urban dispersion modelling system (from the Finnish Meteorological Institute), consisting of a multiple source Gaussian plume model and meteorological pre-processor, was used to model air pollution in the city of Espoo, Finland (Valkonen et al. 1996). Estimates of hourly time-series of SO₂, NO₂ and CO concentrations in 1990 were calculated and used in analysing the influence of urban air pollution on children's health in the city. Research by Kousa et al. (2002) involved the combined use of a line dispersion model (CAR-FMI) and the UDM-FMI to predict hourly NO₂ concentrations in Helsinki for 1996 and 1997. Agreement between predicted and measured hourly concentrations at ten monitoring sites was shown to be significant, with a Pearson correlation of $r = 0.71$. The resulting surfaces were used with time-activity data from the EXPOLIS study to evaluate average exposure of

the Helsinki population to NO₂ pollution. Average exposures to NO₂ at home and in the workplace were shown to be more important than exposures associated with other activities.

Using a case-control study design, Bellander et al. (2001) and Nyberg et al. (2000) assessed the association between historical exposure to air pollution and lung cancer in Stockholm, Sweden. The AIRVIRO dispersion model with area-wide emissions of traffic-related pollution was used to create surfaces for 1960, 1970 and 1980. The NO₂ estimates for 1980 were validated against five independent monitoring sites, indicating that predictions were within $\pm 20\%$ of the measured value (validation was not possible for SO₂). The annual levels of SO₂ and NO_x/NO₂ were then computed for each year between 1950 and 1990 by linear extrapolation and interpolation to complete the historical record. In a separate study already described in Section 2.2.3, exposures to air pollution were assigned to the participants of a large cohort of Norwegian men on the basis of SO₂ and NO_x pollutant concentration fields, also derived from dispersion models (Nafstad et al. 2004; Nafstad et al. 2003).

2.4 Conclusion

The studies outlined in this chapter have shown that the relative risks of exposure to air pollution are small. Due to the large numbers of people exposed, however, these risks nevertheless give rise to large public health burdens. There is, thus, continuing need for population-based studies which are large enough to detect the small relative risks against a lot of uncertainty. Given the spatial variability of air pollution, it is necessary to assign exposures at the local or individual level. This requires methods of exposure assessment that: (1) are accurate and applicable at the individual or small-area scale; and (2) can also be consistently applied to large numbers of people across large geographic areas. Because the monitoring data alone cannot meet both of these requirements there is a need for modelling to estimate exposures at unsampled locations. Various modelling methods are available, but for reasons discussed above, GIS techniques have particular potential. Exposure modelling, however, requires more than simply the pollution surface; it also requires information on where people live or spend their time. GIS-based methods are potentially useful for modelling and mapping the population distribution, and also enable population data to be readily linked to data on pollution. This thesis will,

thus, explore use of GIS for population (Chapter 4) and pollution (Chapters 5 and 6) modelling, and exposure assessment (Chapter 7), at a large population scale (i.e. for the European Union).

3 GIS Development

This chapter describes the development of the GIS used in this study. First, the geographical context in terms of the study area extent and the spatial resolution of the GIS are outlined. Then the attribute data are described, explaining why particular input data were selected, how they were incorporated into the GIS to create the derived data, and the data enhancements used to overcome incompleteness in some data sets.

The two main components of the GIS include: the derived data used as predictor variables in the spatial modelling (1 x 1km gridded data); and the concentration data from the air pollution monitoring network (point data).

3.1 Geographical context

3.1.1 Study period and area

The study period was originally defined as 1997-2001 while the original study area included the 15 Member States of the European Union (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden and the United Kingdom) plus Norway. As will be detailed in subsequent sections, however, the temporal and spatial scopes were reduced due to limitations of the input data. Specifically, pollution modelling was limited to 2001 to take advantage of the improved spatial coverage of the monitoring sites for that year (Section 3.3.4). In terms of the geographic extent of the pollution and population modelling, Norway and Sweden were excluded on the basis of inadequate land cover data (Section 3.2.2). The geographical extent of the redefined study area is shown in Figure 3.1.

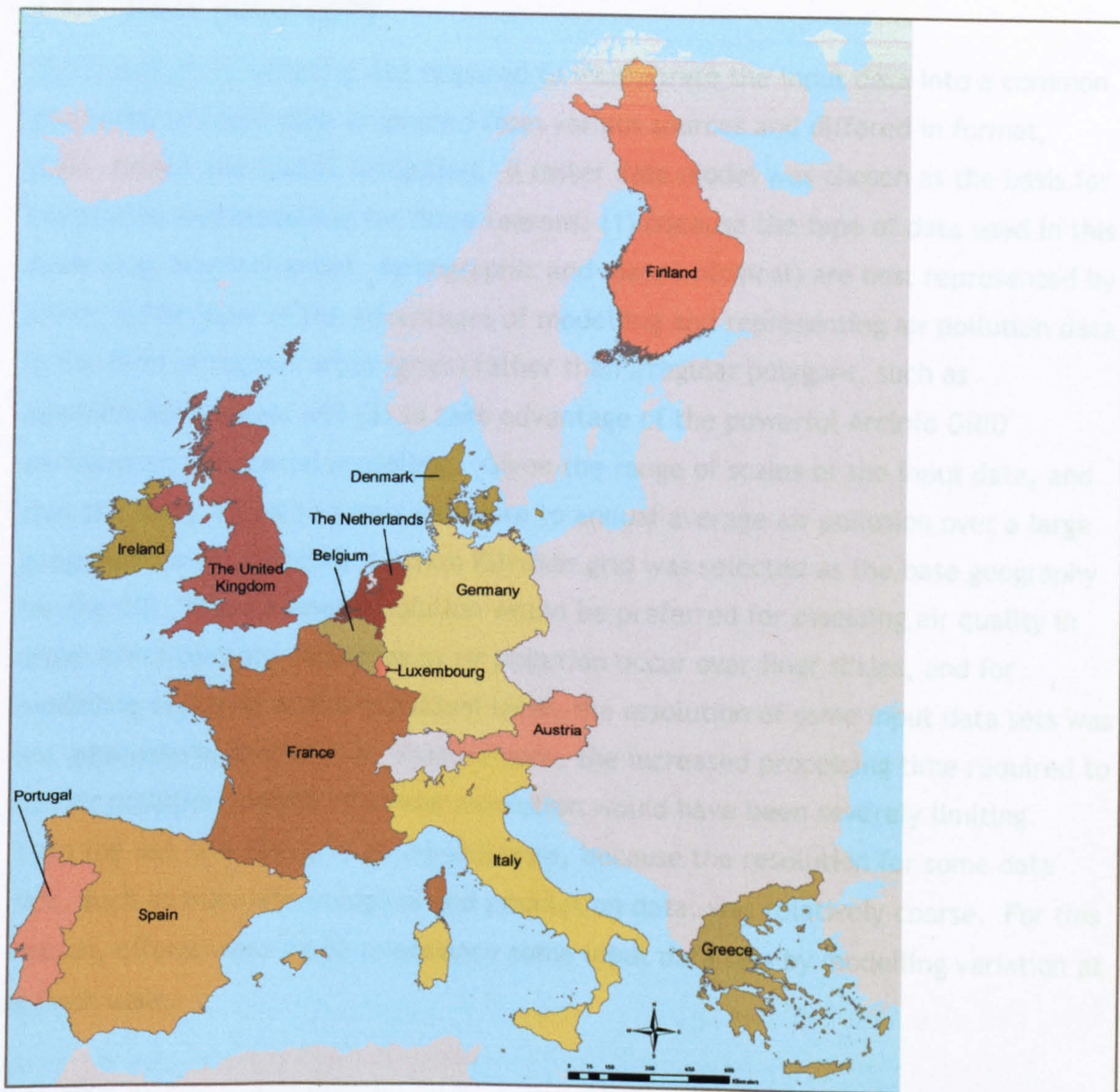


Figure 3.1 Study area extent

3.1.2 Base geography

Significant pre-processing was required to incorporate the input data into a common geography as these data originated from various sources and differed in format, scale, extent and spatial resolution. A raster data model was chosen as the basis for integration and modelling for three reasons: (1) because the type of data used in this study (e.g. environmental, demographic and meteorological) are best represented by areas; (2) because of the advantages of modelling and representing air pollution data in the form of regular areas (grids) rather than irregular polygons, such as administrative areas; and (3) to take advantage of the powerful ArcInfo GRID environment for spatial modelling. Given the range of scales of the input data, and that the study aimed to assess exposure to annual average air pollution over a large geographic area, a regular 1 x 1km EU-wide grid was selected as the base geography for the GIS. While a finer resolution would be preferred for assessing air quality in urban areas because variations in air pollution occur over finer scales, and for modelling exposure at the individual level, the resolution of some input data sets was not adequate to enable this. Furthermore, the increased processing time required to run air pollution models at a finer resolution would have been severely limiting. Even the use of a 1km grid posed problems, because the resolution for some data sets, such as the meteorological and population data, was relatively coarse. For this reason, efforts were made to enhance some input data sets by modelling variation at a finer scale.

ArcGIS 9.0 was used to store, analyse and manipulate the GIS data. The raster base geography was created based on the specification of the CORINE land cover data set, a fundamental data set in this study, using the ARC GENERATE command. Because this data set was in Lambert Azimuthal Equal Area projection, this was used as the base projection for the study, and all other data sets were reprojected accordingly (parameters given Appendix A1). The vector geography used to define the study area extent (i.e. coastline and land boundaries) was derived from the AND Roads data set (1:300,000).

3.2 Derived 1 x 1km data

As stated above, the raw data sets were converted to a common 1 x 1km grid format to facilitate modelling. In selecting the input data sets, every attempt was made to

satisfy several criteria, and deviations from the criteria are noted in the relevant sections. The criteria were:

- A single data source for each variable covering the full spatial extent of the study area.
- A spatial resolution of 1 x 1km or better for local phenomena (e.g. land cover) and the highest possible resolution for regional variables (e.g. meteorology).
- Data had to be up-to-date and readily available.
- Data had to relate to the same period as the pollution concentration data (i.e. 2001).

The input data sets (Table 3.1) were obtained mainly via download from the Internet or delivery via CD-Rom from data holders. Although many of the data sets are widely used and taken to be of good quality, each was carefully checked for errors prior to and during integration into the GIS. Checks included mapping all data to identify errors in geographical coordinates and gaps in the attribute data; tests for internal consistency (e.g. in names, coordinates, attribute values) from different data sets; and statistical analysis of attribute data to identify outliers. Where errors were identified, the values were, if possible, replaced with estimates based on the available information and the specific data items were flagged. A code for missing values was entered where estimation was not possible.

Various spatial analysis techniques were used to manipulate the input data sets in order to calculate the derived variables listed in Table 3.2 for each 1 x 1km cell. The subsequent sections outline the data pre-processing involved in creating each of the derived variables.

Table 3.1 Input data sets

Data set	Source	Attribute/ description	Scale/ geography	Time period	Geographic extent
Monitored air pollution	AIRBASE database	PM ₁₀ , NO ₂ , SO ₂ , CO, O ₃ Daily / Hourly average. Includes data on site type and altitude	Point	1997- 2001	Various
	EMEP concentrations	PM ₁₀ , NO ₂ , SO ₂ , O ₃ Daily / Hourly average	Point	1997- 2001	All countries
	Sub National data	NO ₂ , SO ₂	Point	1997- 2001	United Kingdom
Emissions	Modelled Emissions (APMoSPHERE project)	PM ₁₀ , NO _x , SO _x , CO	1km grid	2001	All countries
Land cover	CORINE land cover 100m grid - Version 12/2000 (CLC90) from the EEA	44 land cover classes	1:100,000; 100m grid	1990	Missing Sweden and Norway
	Global Land Cover 2000 (GLC2000)	22 land cover classes	1km grid	2002	All countries
Altitude	GTOPO30	Digital elevation model - mean altitude (m)	30 arc seconds (~1km grid)	1993- 1996	All countries
Transport networks	AND Data Ireland Ltd.	Road and rail lines, 6 levels of classification	1:300,000 Vectors	2002	All countries
Meteorology	The European Commission Joint Research Centre (JRC). MARS daily data (derived from ECMWF)	Atmospheric stability, temperature, precipitation	50km grid	1997- 2001	All countries
	ECMWF, Operational Archive	10m U and V wind components (to calculate wind speed and direction)	0.5 degree grid, lat/long	1997- 2001	All countries
Administrative boundaries	GISCO	NUTS 3 boundaries	1:100,000 Vectors	2001	All countries
	SABE	NUTS 5 boundaries	1:100,000 Vectors	2001	All countries
Population	REGIO	Number of resident population	NUTS 3 statistics	1997	Germany
	SIRE database	Number of resident population	NUTS 5 statistics	1991/ 2003	All countries
Light emissions	MANTLE Study	Light emissions	200m grid	1999- 2000	Excludes Norway and Sweden

Table 3.2 Derived variables

Variable	Description	Neighbourhood
Topography		
Altitude	Metres above sea level	1 x 1km
Topex	Height difference between 1km ² centroid and centroids of 24 surrounding 1km squares	5 x 5km
Slope/Aspect	Mean slope (percent) between 1km ² centroid and centroids of 4 compass directions (NW, NE, SW, SE)	3 x 3km
Climate		
Temperature	Mean annual, summer and winter temperatures (°C)	1 x 1km
Temperature range	Annual mean temperature range (°C) (Max - Min)	
Wind speed	Mean annual, summer and winter wind speed (m/s)	
Wind direction	Mean annual, summer and winter wind direction (0-360°)	
Days with calm	Percentage of days with wind speed <1 m/s (annual, summer and winter)	
Rainfall	Annual mean rainfall (mm)	
Vapour pressure	Annual mean vapour pressure (hPa)	
Radiation	Annual mean radiation (Kj/m ²)	
Distance to sea	Distance (km) to nearest coastline, by compass quadrant (NE, NW, SE, SW)	
Land cover		
HDRES	Continuous urban fabric (% percent area of grid square)	1 x 1km; 3 x 3km; 5 x 5km; 11 x 11km; 21 x 21km
LDRES	Discontinuous urban fabric (%)	
IND	Industrial/commercial land (%)	
TRANS	Road/rail and associated land (%)	
PORT	Port areas (%)	
AIRP	Airport areas (%)	
WECS	Waste, construction & urban institutions (%)	
URBGR	Urban green spaces (%)	
OBU	Other built-up: TRANS, PORT, AIRP and WECS (%)	
NRBU	Non-residential built-up: OBU plus IND (%)	
TBU	Total built-up: NRBU plus HDRES and LDRES (%)	
FOR	Forest: Deciduous, coniferous and mixed woodland; plantations (%)	
TILL	Tilled agricultural land (%)	
AGR	Agriculture: TILL plus other arable land (%)	
Other		
Population density	Number of residents per km ²	1 x 1km; 3 x 3km; 5 x 5km; 11 x 11km; 21 x 21km
Road length	Road length by road type (km/neighbourhood)	
Rail length	Rail length (km/neighbourhood)	
Light emissions	Light emissions (by CORINE land cover class)	

3.2.1 Regional variables

Each of the derived variables listed in Table 3.2 was initially calculated for every 1 x 1km cell across the study area for use as predictors in the population or pollution modelling. As both local and regional trends influence air pollution, focal functions were applied to some variables to create additional regionalised variables. Focal functions, such as focalsum, calculate new values for each cell in a grid using some function of the input value at that location and the values of the surrounding cells in a specified neighbourhood or window. This is also known as moving window analysis.

The window configuration is determined by a kernel file which specifies the cells around the target cell that are to be included in the calculation of each output value. A 3x3 cell window is the default configuration; however other sizes and irregular shapes can be defined by the kernel. Focal functions, for example, can be used to calculate the sum, mean, standard deviation, or range of values within the window. Specifically, focalsum functions were extensively used in this study (Figure 3.2).

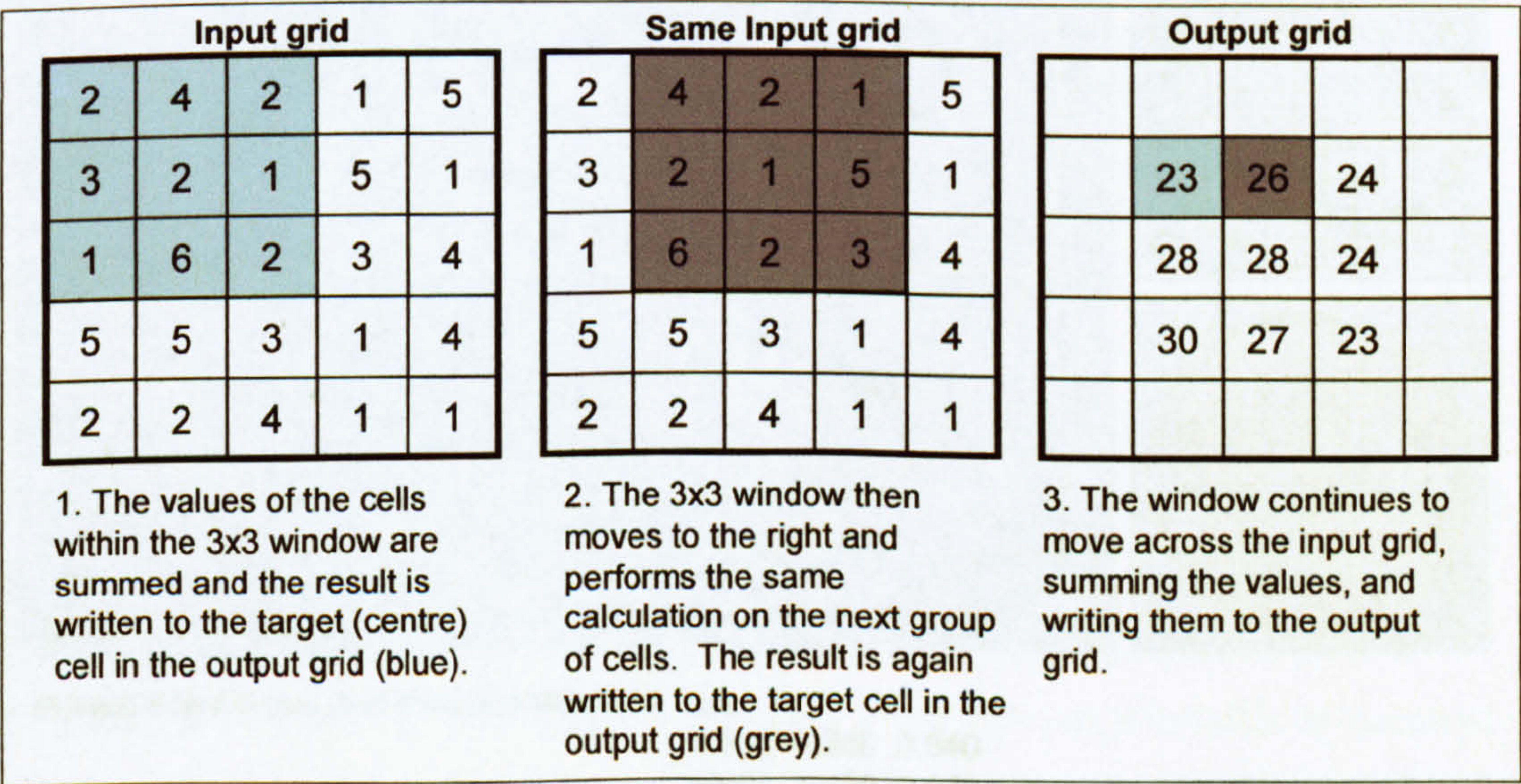


Figure 3.2 Moving window example: Focalsum

A kernel file is basically a text file used to define the window size and shape, as well as the weights assigned to each cell. The first line specifies the number of cells in the x and y directions (i.e. the width and height of the window). The subsequent lines give the values for each position in the window. To define an irregular shaped kernel, the window is represented by ones and zeros to identify cells that are

included and excluded, respectively, from processing. Alternatively, a value between 0 and 1 can be assigned to weight the cells under the window.

To calculate the regional variables, circular windows with a radius of 1.5, 2.5, 5.5 and 10.5km, corresponding respectively to a 3x3, 5x5, 11x11 and 21x21 kernel, were used (Figure 3.3). The circular windows were created by specifying a kernel containing weights, where the weights represent the proportion of each 1 x 1km cell that falls inside the window. Multiplying each cell value by its associated weight will result in an area-weighted focal function.

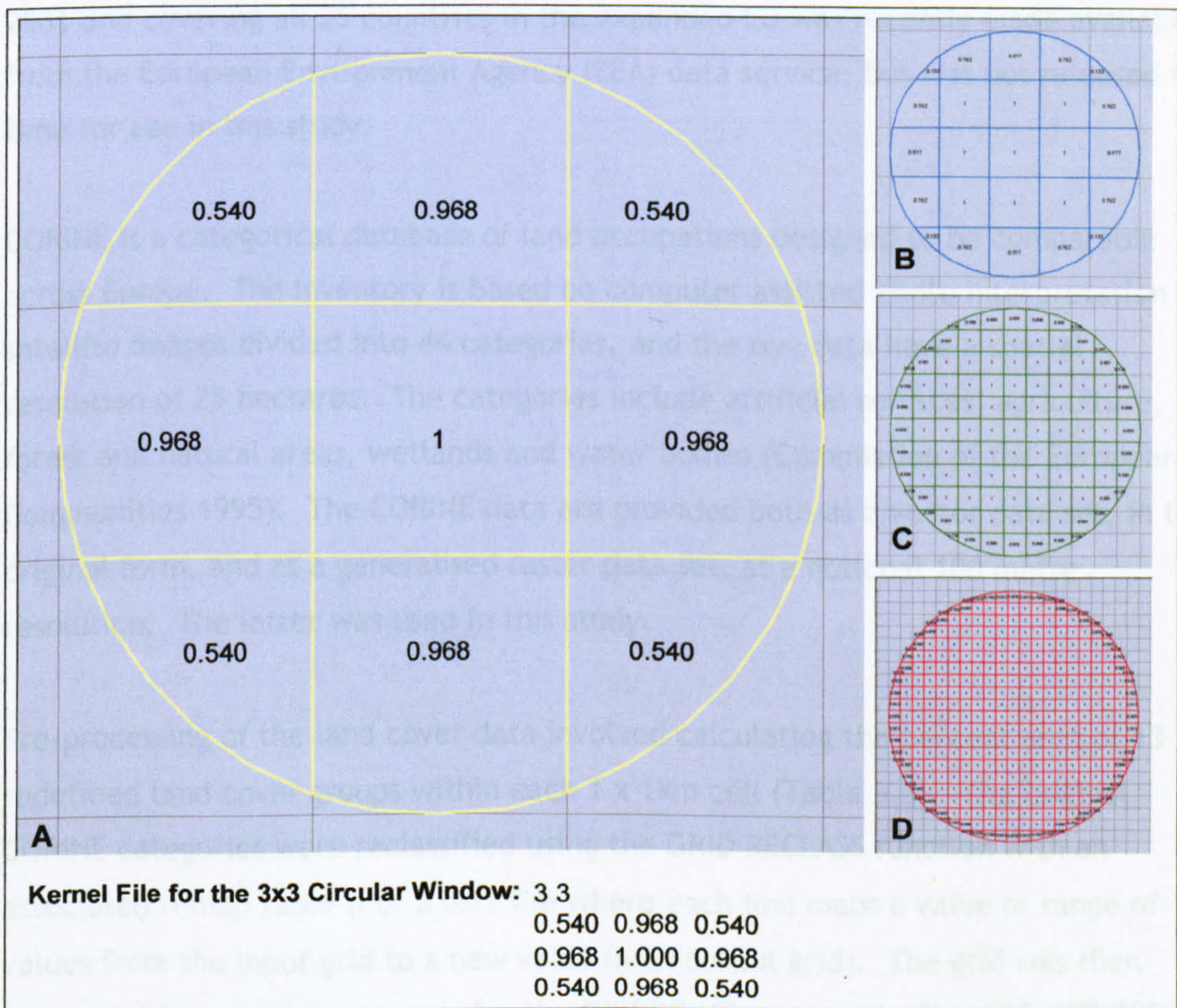


Figure 3.3 Circular window weights used to calculate the regional variables
(A) 3x3 window and kernel file; (B) 5x5 window; (C) 11x11 window; and
(D) 21x21 window

3.2.2 Land cover

Land cover data are important inputs to air pollution modelling because they provide information on the distribution both of potential emission sources and of factors

(such as vegetation or buildings) that might affect local dispersion. Land cover is also a useful proxy for population distribution, and is used in this study to help model population density at the small-area scale (see Chapter 4).

Land cover data were obtained from the CORINE Land Cover (CLC90) map, version 12/2000. This data set was compiled on the basis of Landsat and SPOT satellite imagery, using a combination of semi-automated and manual interpretation techniques (Commission of the European Communities 1995). The source imagery derives from the late 1980s and early 1990s, but the full digital map coverage of the EU was not completed until the mid 1990s, and coverage for Sweden has never been available. A new data set, CLC2000 (version 5/2005), based on imagery from 1999 to 2001 and covering all 25 countries in the expanded EU was recently made available from the European Environment Agency (EEA) data service, but was not released in time for use in this study.

CORINE is a categorical database of land occupations designed to be comparable across Europe. The inventory is based on computer assisted photo interpretation of satellite images divided into 44 categories, and the raw data have a spatial resolution of 25 hectares. The categories include artificial surfaces, agriculture, forest and natural areas, wetlands and water bodies (Commission of the European Communities 1995). The CORINE data are provided both as a vector data set, in their original form, and as a generalised raster data set, at a notional 100 metre resolution. The latter was used in this study.

Pre-processing of the land cover data involved calculating the percent area of 13 redefined land cover groups within each 1 x 1km cell (Table 3.3). The original CORINE categories were reclassified using the GRID RECLASS function with an associated remap table (i.e. a text file where each line maps a value or range of values from the input grid to a new value in an output grid). The grid was then converted to a point coverage using the GRIDPOINT command. The ARC INTERSECT command was then used to compute the geometric intersection of the point coverage (land cover) and the 1 x 1km grid to identify the cell within which each point was contained.

For each of the 13 land cover groups the following steps were performed. First, the number of points within each 1 x 1km cell was summed to derive the percent land

cover area by group (i.e. at the 1 x 1km level, the number of land cover points is equivalent to the percent area as there are 100 points per 1 x 1km cell). A new grid was created using the POLYGRID command, and the GRID FOCALSUM function with a weighted kernel was applied to sum the land cover area in four circular windows (3 x 3km, 5 x 5km, 11 x 11km and 21 x 21km). For each neighbourhood, the percent land cover area was calculated by dividing the cell value by the area of the corresponding window. In total, 65 land cover variables were calculated and assigned to every 1 x 1km cell in the study area.

Table 3.3 Land cover classification

Original CORINE class (CLC90)	Description	Redefined group
1. Continuous urban fabric	high density residential	HDRES (1)
2. Discontinuous urban fabric	low density residential	LDRES (2)
3. Industrial or commercial units	industry	IND (3)
4. Road and rail networks	road and rail infrastructure	TRANS (4)
5. Port areas	sea ports	PORT (5)
6. Airports	airports	AIRP (6)
7. Mineral extraction sites	waste and construction	WECS (7a)
8. Dump sites		
9. Construction sites		
10. Green urban areas	urban greenery	URBGR (7b)
11. Sport and leisure facilities		
22. Agro-forestry areas	forest	FOR (8)
23. Broad-leaved forest		
24. Coniferous forest		
25. Mixed forest		
12. Non-irrigated arable land	agriculture	AGR (9)
13. Permanently irrigated land		
14. Rice fields		
15. Vineyards		
16. Fruit trees and berry plantations		
17. Olive groves		
18. Pastures		
19. Annual crops		
20. Complex cultivation patterns		
21. Land mainly occupied by agriculture		
Classes 1-9	total built-up	TBU (10)
Classes 3-9	non-residential built-up	NRBU (11)
Classes 12-17	tilled agriculture	TILL (12)
Classes 4-9	other built-up	OBU (13)

The CLC90 version of CORINE does not include data for Norway and Sweden; therefore, other data sources were sought. Of these, potentially the most useful was the Global Land Cover 2000 (GLC2000) data set which was produced by 30 institutions world-wide through a coordinated effort led by the Joint Research Centre (JRC)

Institute for Environmental Sustainability. GLC2000 was derived from 14 months of pre-processed daily global data acquired by the VEGETATION instrument on the SPOT 4 satellite. It has a resolution of 1km² and contains 22 land classifications, 18 of which are vegetation classes and only one which represents artificial surface (JRC 2006b). Preliminary analysis indicated that these data were inadequate for population and pollution modelling as the resolution and available land cover classes were too coarse. As a result, the study area was reduced to the extent of the CORINE land cover data (i.e. without Sweden and Norway). Thus limited, it comprised approximately 2.8 million 1 x 1km cells. Figure 3.4 shows the percent total built-up land at the 1 x 1km level derived from the CORINE land cover map.

3.2.3 Population

At the time of this study, population data from the 2001 census had not been released across the EU (and, indeed, no plans existed to provide small-area census data from the 2001 census for the whole EU). In their absence, the most recent and detailed data set covering the EU were based on the 1991 census (or its nearest equivalent). These were available for Nationales Unites Territoriales Statistiques level 5 (i.e. NUTS 5, now referred to as Local Administrative Unit or LAU 2) administrative regions, and held in the SIRE database maintained by Eurostat.

The NUTS system is a hierarchical classification that subdivides the Member States into five levels - the first three of which are legislated by the EU to facilitate the collection, compilation and dissemination of regional statistics (Official Journal of the European Communities 2003). NUTS levels 1 to 3 represent the country, province/state and county areas, respectively, while NUTS 5 comprise the smallest set of administrative regions for which data are consistently available. The NUTS 5 units are represented, for example, by wards in Great Britain and communes in France. Administrative boundaries for these regions are presented as vectors in the SABE database. The SABE boundaries, dated 2001, were obtained in digital form from GISCO, which is a warehouse for geographical data in the EU.

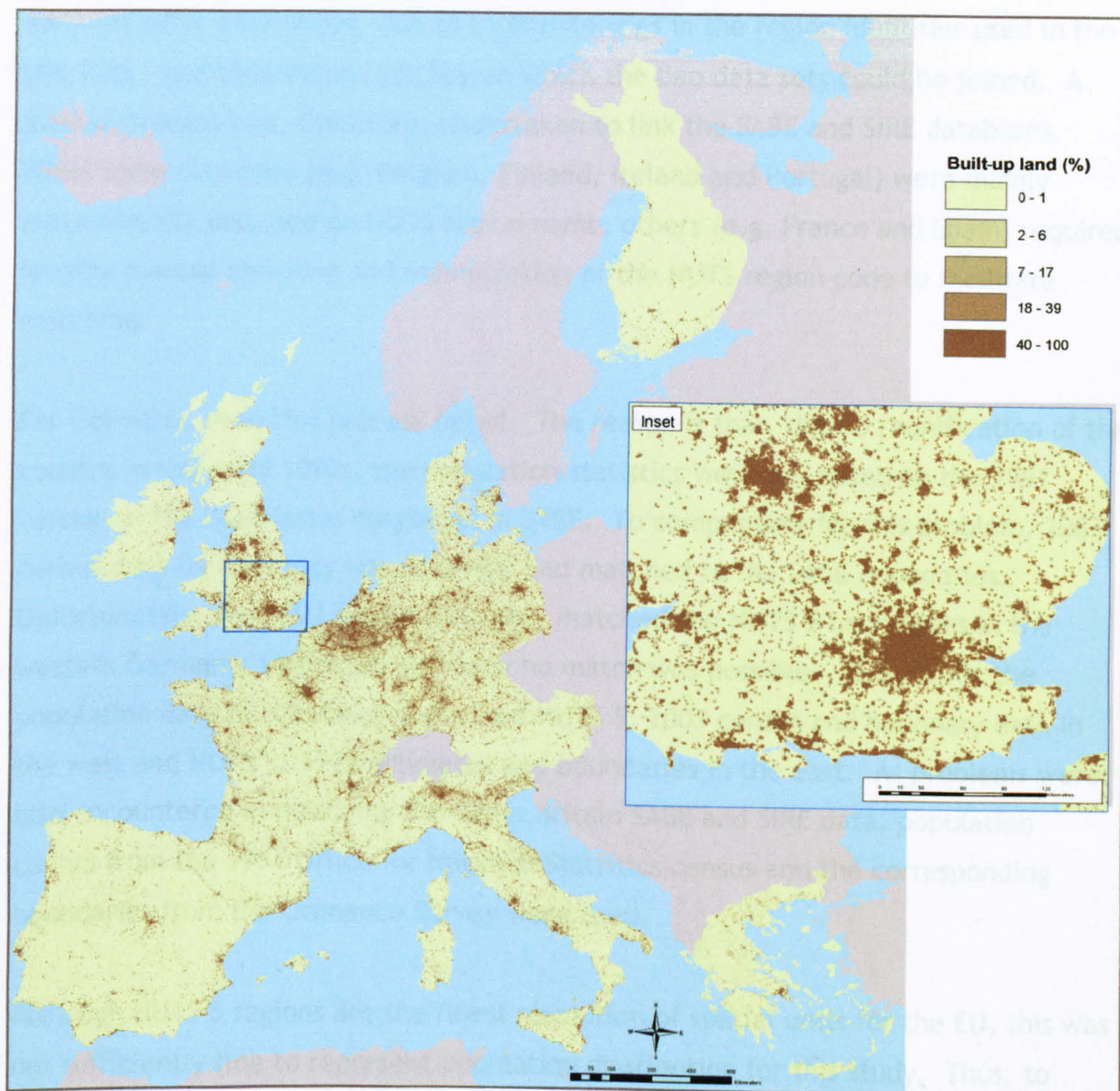


Figure 3.4 Total built-up land (% area)
Derived from CORINE land cover

3.2.4 Transport networks

Road transport is clearly an important exposure source for both particulates and nitrogen dioxide. A highly detailed digital transportation network, equivalent to the Ordnance Survey data in the UK, is not available for the entire study area at an affordable price. The AAS data for Europe were, therefore, used for the transportation network in this study. The data are vector-based shapefiles with a spatial resolution of 1:300,000, representing railways and all major road classification.

Some problems were encountered during matching the population data to their administrative boundaries, due to inconsistencies in the region identifier used in the two files - and thus the unique key on which the two data sets could be joined. A manual process was, therefore, undertaken to link the SABE and SIRE databases. While some countries (e.g. Belgium, Finland, Ireland and Portugal) were mainly automatically matched on NUTS region name, others (e.g. France and Spain) required lengthy manual checking and manipulation of the NUTS region code to facilitate matching.

For Germany, even this process failed. The reason is that, due to reunification of the country in the early 1990s, the population statistics were not based on the 1991 version of the boundaries contained in SABE. To compensate for this problem, 2003 census data for Germany was obtained and matched to the SABE boundaries. Unfortunately, the 2003 census data only matched the NUTS 5 boundaries in the western Germany; in eastern Germany no match was possible. As a result, the population data for Germany comprised NUTS 5, 2003 census and boundary data in the west and NUTS 3, 1997 estimates and boundaries in the east. As problems were also encountered in matching the Great Britain SABE and SIRE data, population counts from the 1991 Office for National Statistics census and the corresponding boundaries from the Ordnance Survey were used.

Although NUTS 5 regions are the finest resolution of spatial units for the EU, this was not sufficiently fine to represent population distribution for this study. Thus, to enable more precise estimates of exposure, the population data were converted to the 1 x 1km grid by spatial modelling as described in Chapter 4.

3.2.4 Transport networks

Road transport is clearly an important emission source for both particulates and nitrogen dioxide. A highly resolved digital transportation network, equivalent to the Ordnance Survey data in the UK, is not available for the entire study area at an affordable price. The AND Data for Europe were, therefore, used for the transportation network in this study. The data are vector-based shapefiles with a spatial resolution of 1:300,000, comprising railways and six major road classifications.

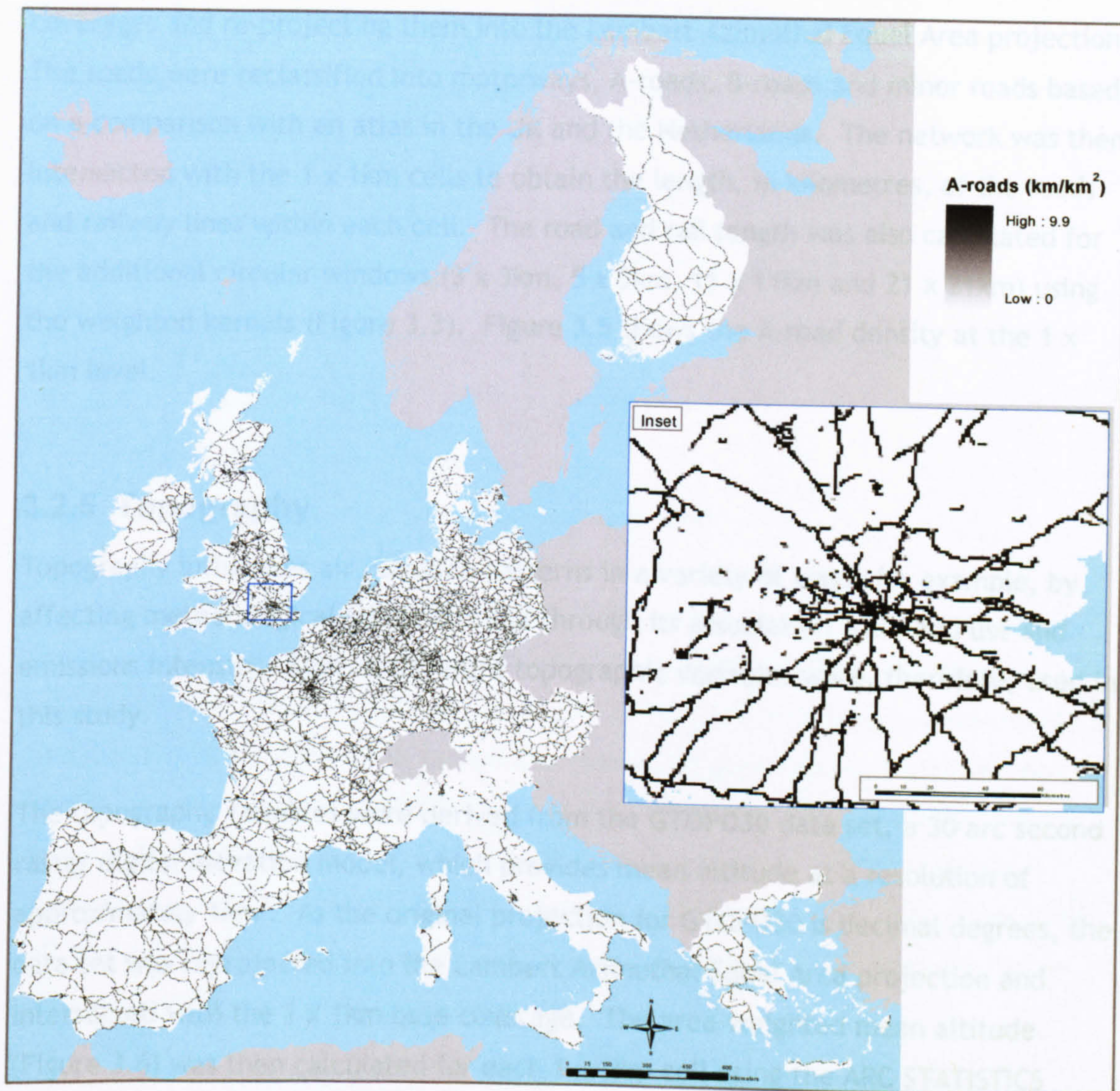


Figure 3.5 **A-road density (km/km²)**
 Derived from the AND Data for Europe

Pre-processing of this data initially involved converting the shapefiles to ArcInfo coverages and re-projecting them into the Lambert Azimuthal Equal Area projection. The roads were reclassified into motorways, A-roads, B-roads and minor roads based on a comparison with an atlas in the UK and the Netherlands. The network was then intersected with the 1 x 1km cells to obtain the length, in kilometres, of the roads and railway lines within each cell. The road and rail length was also calculated for the additional circular windows (3 x 3km, 5 x 5km, 11 x 11km and 21 x 21km) using the weighted kernels (Figure 3.3). Figure 3.5 shows the A-road density at the 1 x 1km level.

3.2.5 Topography

Topography influences air pollution patterns in a variety of ways: for example, by affecting meteorological conditions and through its association with land use and emissions intensity. Several different topographic variables were, therefore, used in this study.

The topography variables were derived from the GTOPO30 data set, a 30 arc second raster digital elevation model, which provides mean altitude at a resolution of approximately 1km². As the original projection for GTOPO30 is decimal degrees, the data set was reprojected into the Lambert Azimuthal Equal Area projection and intersected with the 1 x 1km base coverage. The area-weighted mean altitude (Figure 3.6) was then calculated for each 1 x 1km cell using the ARC STATISTICS command, and a new altitude grid (Figure 3.7) was created from this attribute using the POLYGRID command. Two additional topography variables, Topex and slope-aspect, were then derived from this grid.

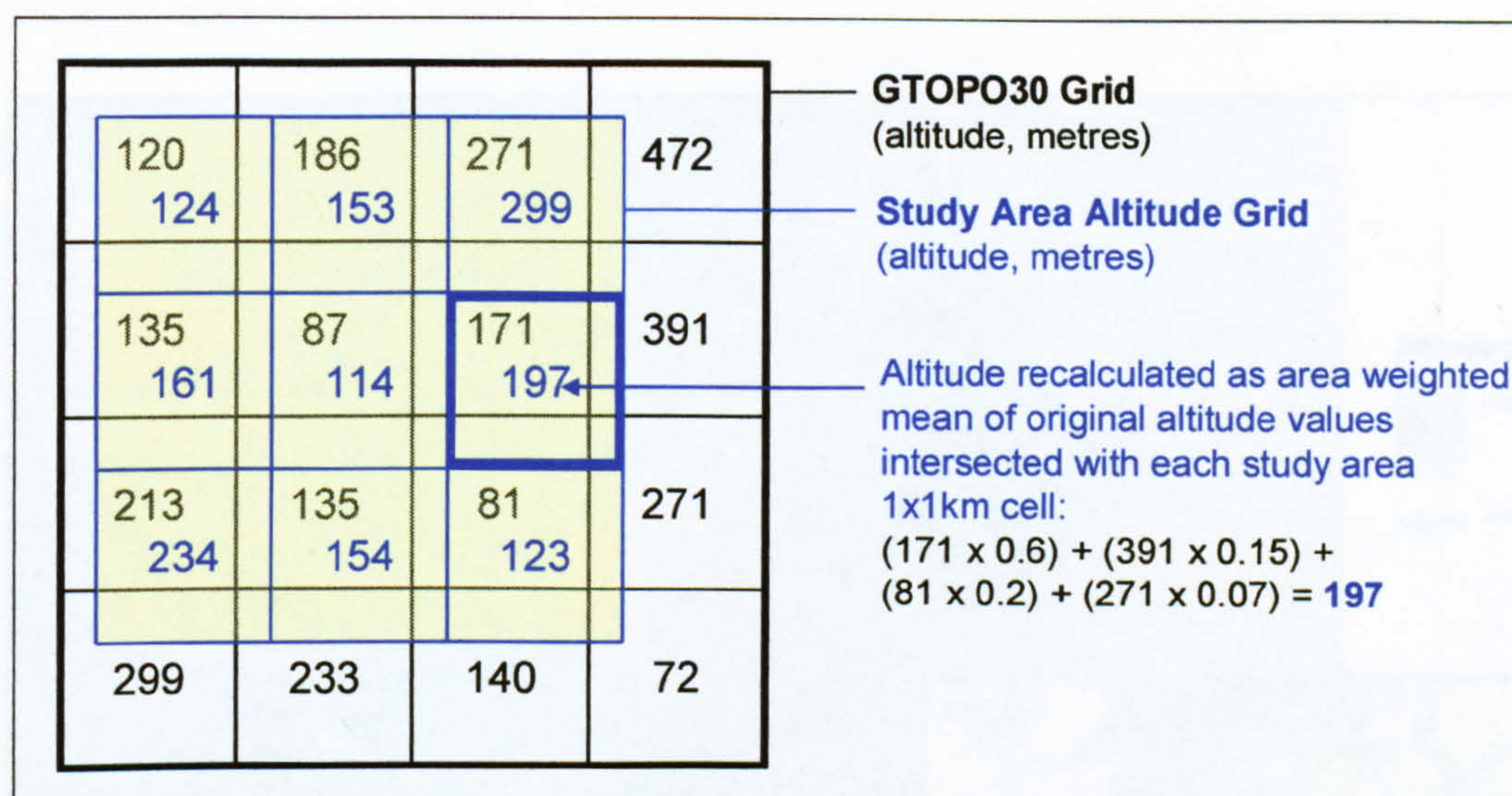


Figure 3.6 Reprojected, area-weighted altitude grid

Topex is a measure of the degree of topographic exposure, or openness, of the terrain. This variable provides an index of the capability for free dispersion of pollutants away from local emission sources. In general, high Topex values should be associated with lower levels of pollution. Topex was computed as the difference in altitude between a 1km² centroid and the mean altitude of the 24 surrounding 1 x 1km cells (Figure 3.8). A Topex value near to zero indicates that the regional topography is flat or uniformly sloping. A large positive value means that the 1km² cell is peaked compared to the surrounding area, whereas a large negative value indicates a depression in the topography.

The steepness of hills and mountains, and the direction that they face in relation to emission sources, can influence pollution patterns and act as barriers to dispersion. A measure of slope and aspect was, therefore, also derived from the altitude grid. The slope-aspect variables indicate the height difference between the target cell and adjacent cells in four compass directions (NW - northwest, NE - northeast, SE - southeast and SW - southwest) as shown in Figure 3.9. GRID FOCALSUM functions with associated 3x3 kernels were used to assign the adjacent cell value in each direction to the target cell in the output grid. Each directional altitude grid was then subtracted from the altitude grid, resulting in four variables indicating the height difference in each compass direction.

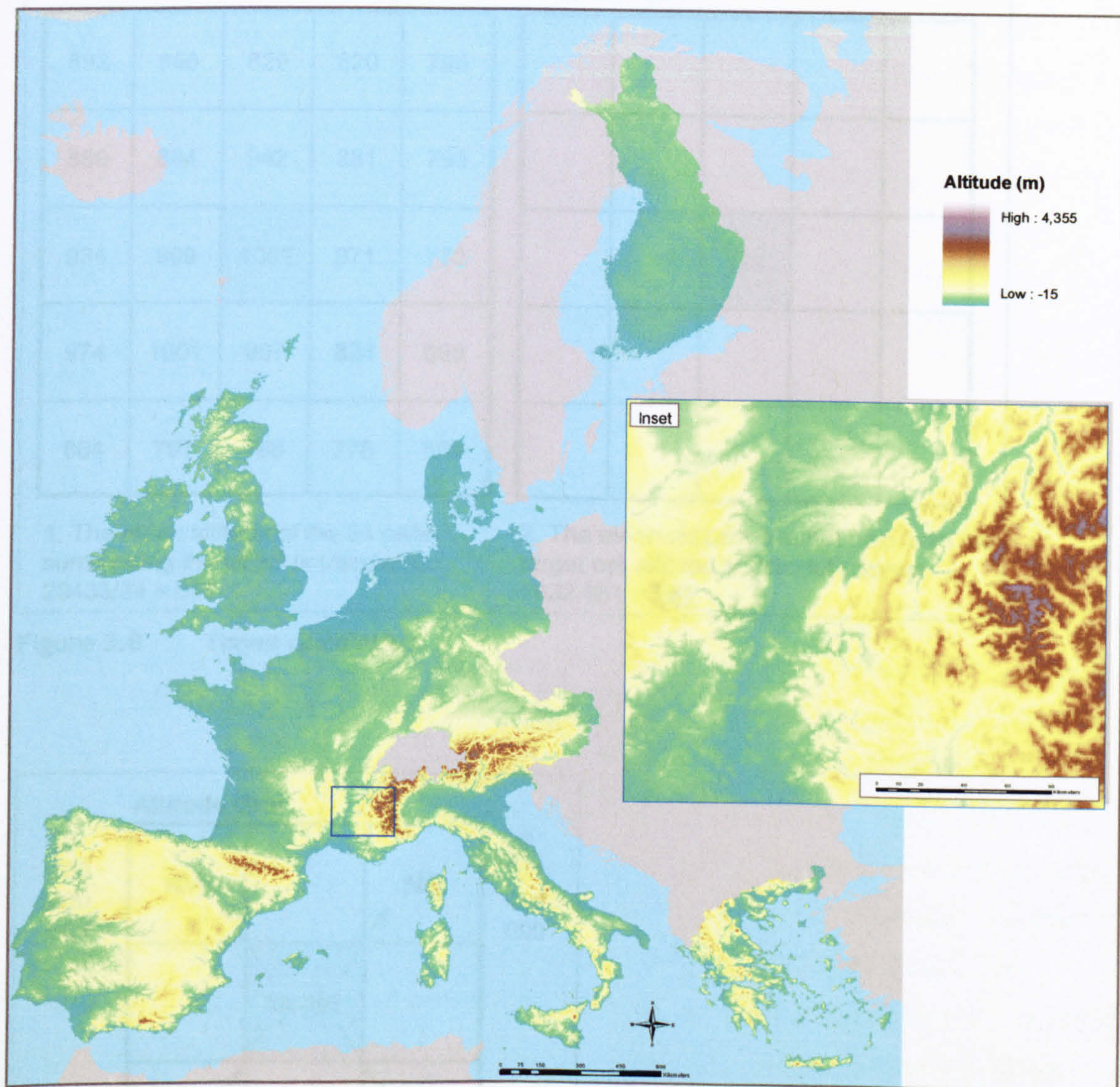


Figure 3.7 Altitude (m)
Derived from GTOPO30

Figure 3.7 : Slope aspect calculations

Maps for each of the four directions are shown in Part

3.2.6 Light Intensity

Since light emissions are primarily associated with roadways, industrial, commercial and residential sources, night time light intensity provides a potentially useful measure of such source activity. In a previous EU-funded project called NANTLE (Mapping Night-Time Light Emissions), light was also shown to be correlated with atmospheric variations in the UK, and (weakly) with NO_x and SO_2 concentrations.

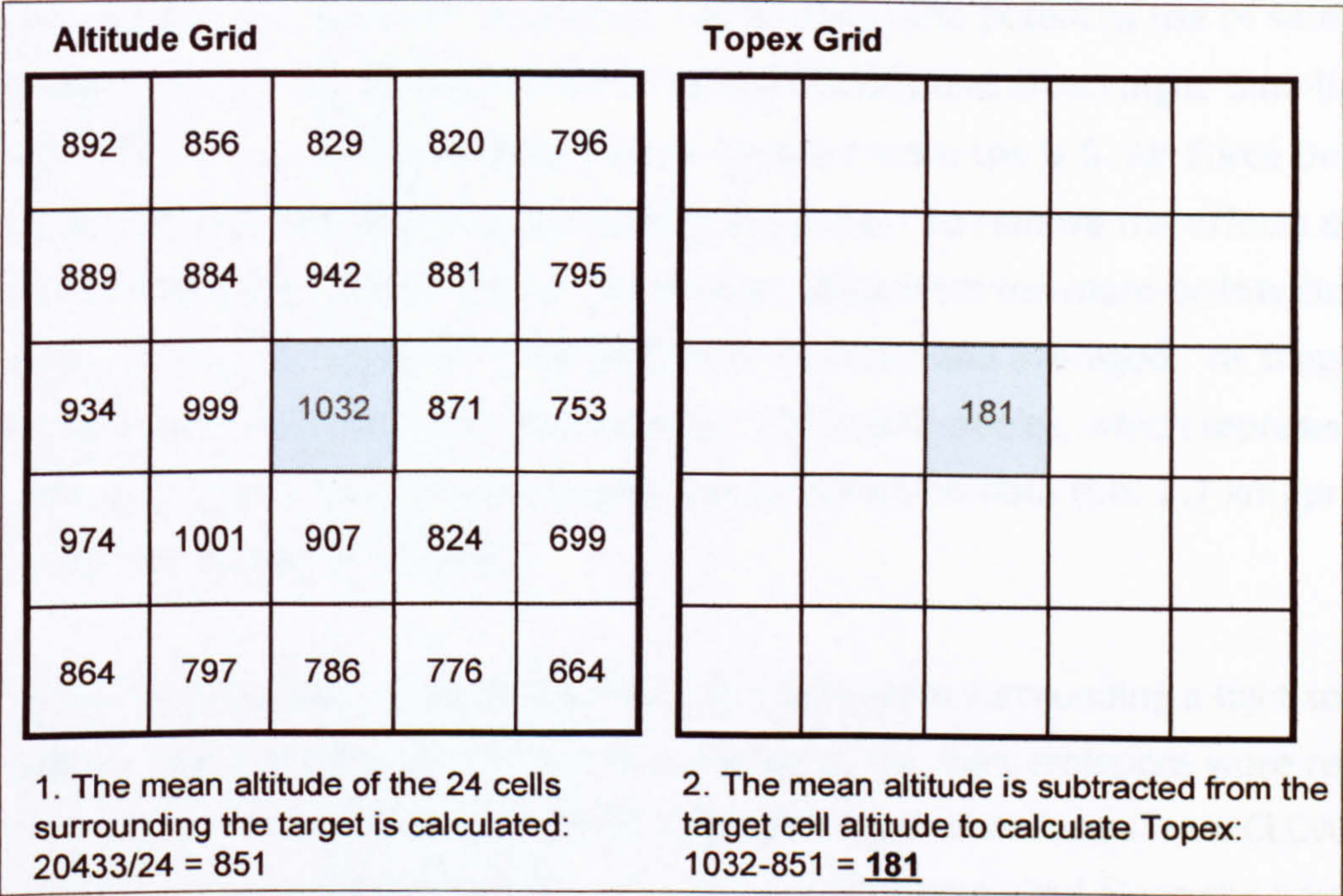


Figure 3.8 Topex calculation

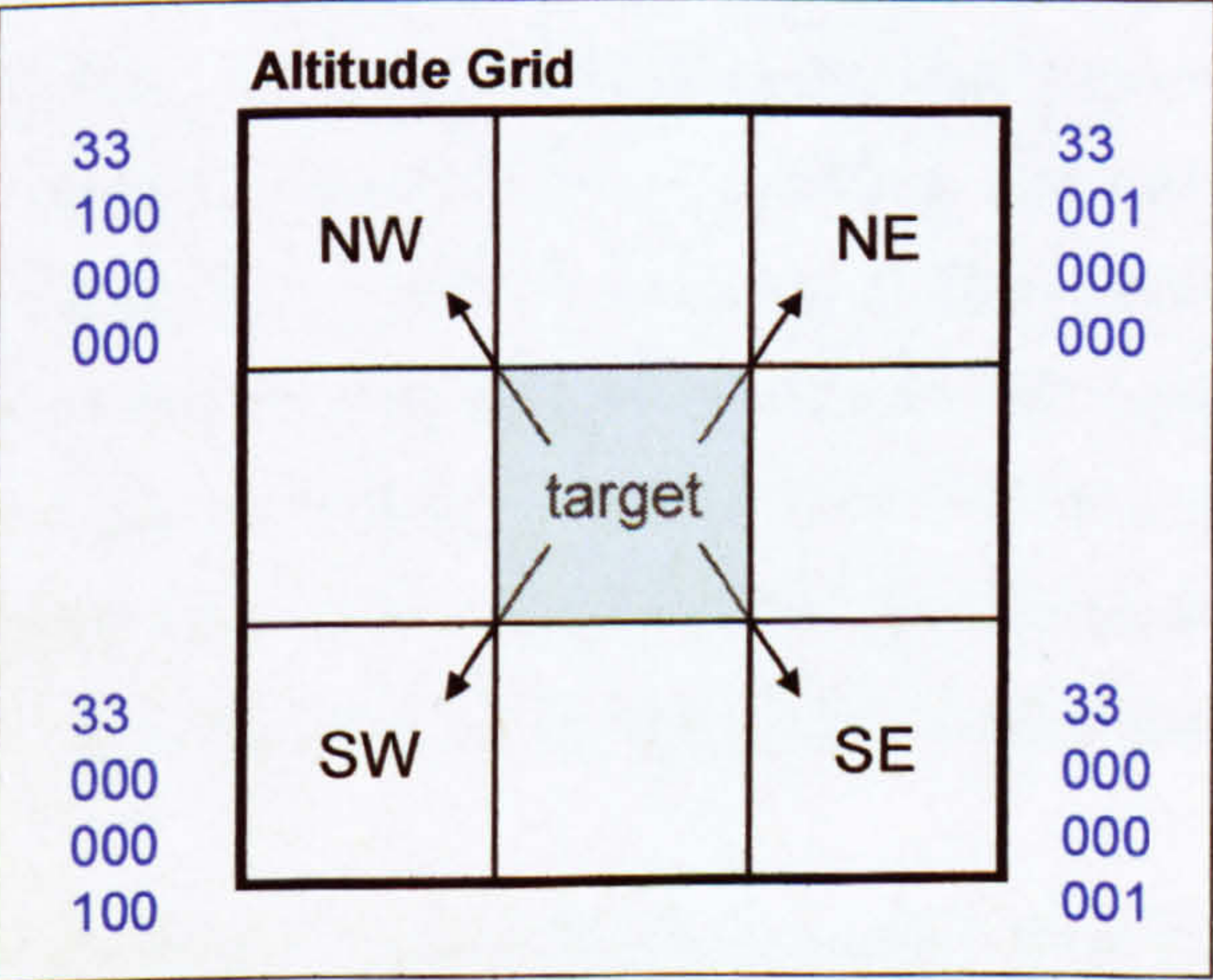


Figure 3.9 Slope-aspect calculations
Kernels for each of the four directions are shown in blue.

3.2.6 Light intensity

Since light emissions are primarily associated with roadways, industrial, commercial and residential sources, night-time light intensity provides a potentially useful measure of such source activity. In a previous EU-funded project called MANTLE (Mapping Night-Time Light Emissions), light was also shown to be correlated with atmospheric emissions in the UK, and (weakly) with NO₂ and SO₂ concentrations.

The objective of the MANTLE project was to assess the potential use of satellite imagery to produce EU-wide maps of light emissions and urban night-time light intensity. Data on light emissions were obtained from the U.S. Air Force Defence Meteorological Satellite Program (DMSP) database. To remove the effects of short-term variations in emissions and cloud cover, data from ten more-or-less cloud-free nights during the winter of 1999-2000 were selected and averaged. As supplied, these data had a nominal resolution of ca. 750 x 600 metres, which represented a notional enhancement of the original coarse resolution data (i.e. 2.7 km) produced by sensors (Briggs et al. 2006).

To assess the extent of blooming (i.e. the bright glare surrounding a light source against a dark background) inherent in the data, the light emissions were reprojected into Lambert Azimuthal and overlaid onto land cover boundaries from CLC90. The match between the two data sets was then visually inspected along the boundaries of lit and unlit land cover classes (e.g. where urban areas were adjacent to forest or open grassland). Based on this, a buffer of 400 metres was selected to represent the limit of blooming. To facilitate processing, the gridded data were converted to points, each representing a 200 metre pixel centroid. To identify points for which the attached DMSP value could be considered unaffected by blooming from adjacent classes (i.e. 'pure'), a moving window was used to locate all points within a single land cover class that were at least 400 metres from any other land class. These points were then randomly stratified into two subsets - 90% of the selected points for each land cover class were designated as a training set, and the remaining 10% of selected points were reserved for validation.

Estimates of light emissions at each unselected point were then made using inverse distance weighting (IDW) and universal kriging in ArcGIS. Results from all models were assessed by comparing the actual and predicted light emissions at the 10% 'pure' points, retained for validation purposes. Universal kriging slightly outperformed IDW in terms of the resulting regression statistics ($R^2 = 0.97$ and 0.96 , respectively, $SEE=113$ and 125 respectively, for $n = 1,130,467$ pixels). Kriging, however, occasionally produced extreme values in areas with few training points. Results from the IDW method, therefore, were used to generate the MANTLE light intensity map.

For use in this research, the mean value from the 200m MANTLE light intensity map was simply aggregated to 1km² to match the 1 x 1km cells in the study area. Weighted kernels were then passed across the 1 x 1km grid to sum the light intensity in the four circular windows (3 x 3km, 5 x 5km, 11 x 11km and 21 x 21km diameter), and the resulting attributes were assigned to each 1 x 1km cell. The 1 x 1km aggregated map is presented in Figure 3.10.

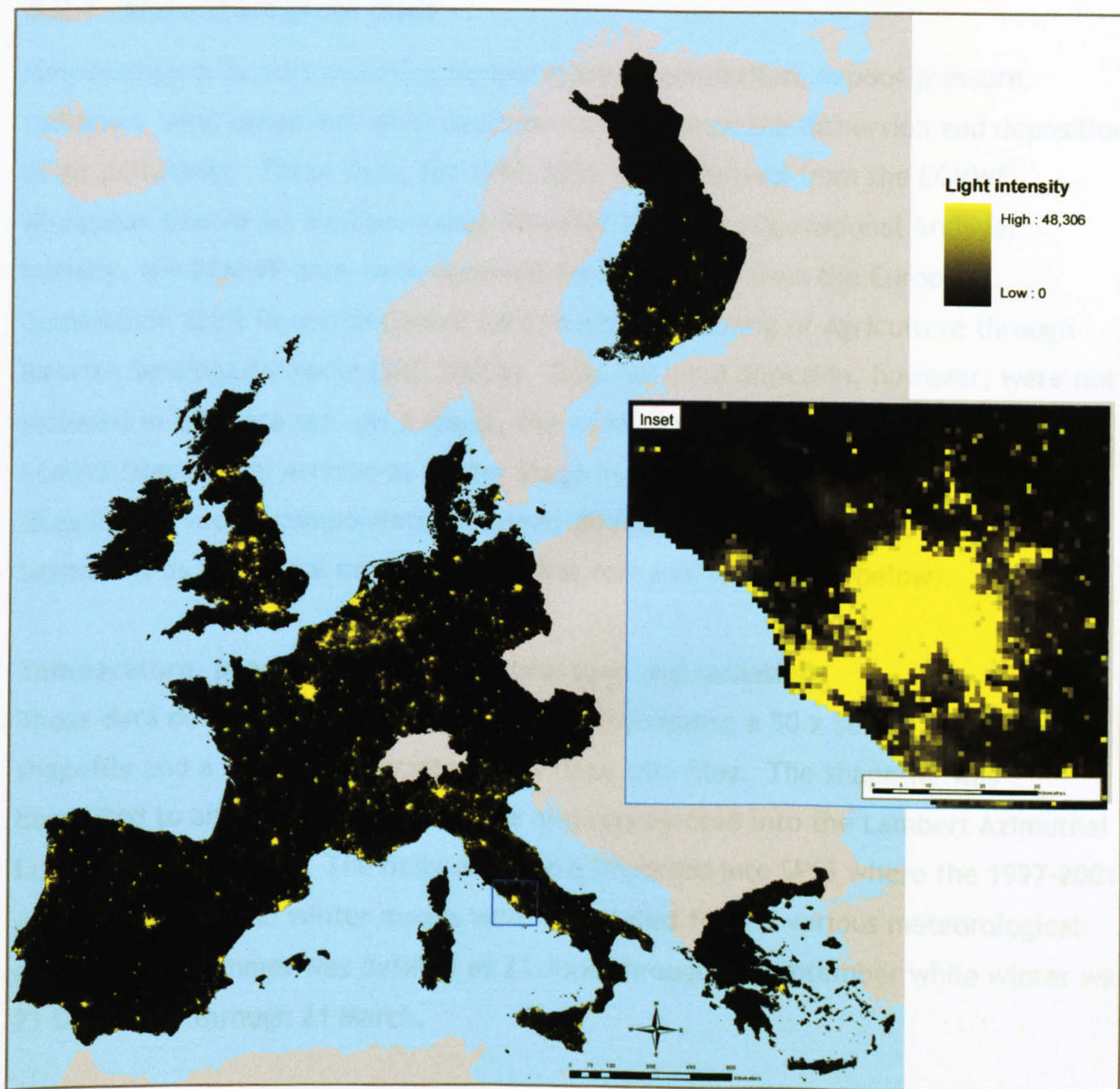


Figure 3.10 1 x 1km light intensity
Derived from the 200m MANTLE map

3.2.7 Meteorological data

Meteorological factors including temperature, precipitation, vapour pressure, radiation, wind speed and wind direction can influence the dispersion and deposition of air pollutants. These data, for 1997-2001, were derived from the ECMWF (European Centre for Medium-Range Weather Forecast) Operational Archive. Initially, the ECMWF data were obtained free of charge from the European Commission Joint Research Centre (JRC) MARS (Monitoring of Agriculture through Remote Sensing) database (JRC 2006a). Data for wind direction, however, were not included in this data set. As a result, the wind data were sourced directly from the ECMWF Operational Archive at a later stage in the study. The wind variables (supplied as vector components) obtained directly from ECMWF were treated separately as additional pre-processing was required (described below).

Temperature, precipitation, vapour pressure and radiation

These data downloaded from the JRC website comprise a 50 x 50km resolution shapefile and a series of associated daily data text files. The shapefile was converted to an ArcInfo point coverage and reprojected into the Lambert Azimuthal Equal Area projection. The daily data were imported into SPSS where the 1997-2001 annual, summer and winter means were calculated for the various meteorological parameters. Summer was defined as 21 June through 21 September while winter was 21 December through 21 March.

The mean meteorology data were then exported to ArcInfo where they were joined to the 50 x 50km point coverage using the ARC JOINITEM command. Kriging, with the ArcMap GeoProcessing Wizard, was applied to this coverage to interpolate a continuous surface across the EU for each of the meteorology variables. Each surface was then converted to a polygon coverage and intersected with the base coverage to assign the meteorological variables to every 1 x 1km cell. A final 1 x 1km coverage containing all meteorological variables was created for each year. A map of the 2001 annual mean temperature is shown as an example in Figure 3.11.

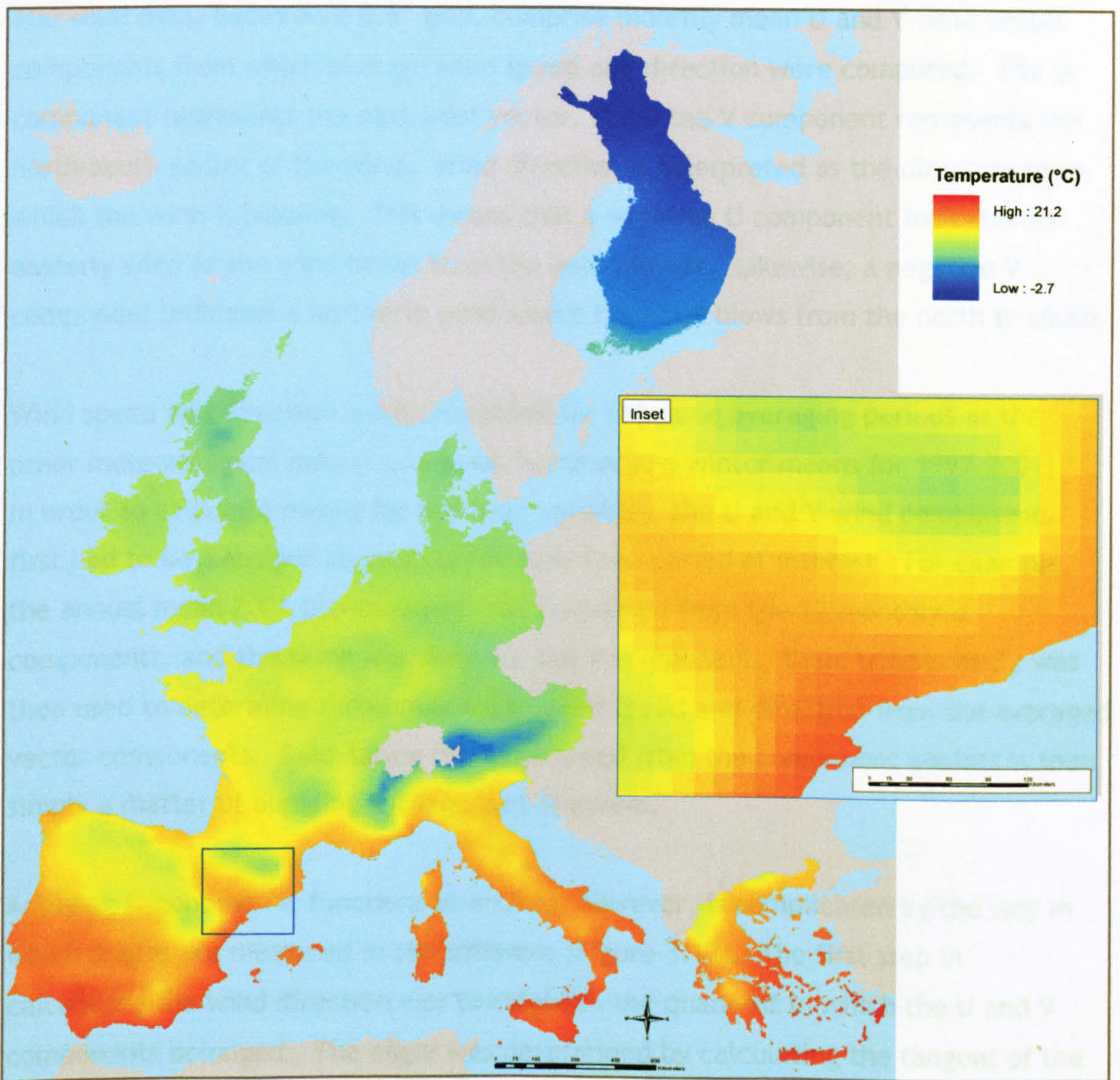


Figure 3.11 2001 Annual mean temperature (°C)
Derived from the JRC MARS database

Wind speed and direction

The wind data, based on a 0.5° grid, comprise monthly mean U and V wind vector components from which average wind speed and direction were computed. The U component represents the east-west vector, while the V component represents the north-south vector of the wind. Wind direction is interpreted as the direction from which the wind is blowing. This means that a negative U component indicates an easterly wind as the wind blows from the east to west. Likewise, a negative V component indicates a northerly wind where the wind blows from the north to south.

Wind speed and direction were calculated for the same averaging periods as the other meteorological data (i.e. annual, summer and winter means for 1997-2001). In order to calculate means for the wind variables, the U and V wind components first had to be averaged separately for each time period of interest. For example, the annual mean 2001 U component was calculated from the 12 monthly U components, and the same was done for the V component. Basic trigonometry was then used to determine the annual mean wind speed and direction from the averaged vector components. Calculating the wind speed from the component vectors is then simply a matter of applying Pythagorean Theorem.

Applying trigonometric functions in ArcInfo, however, is complicated by the way in which angles are measured in the software (Figure 3.12). The first step in calculating the wind direction was to establish the quadrant in which the U and V components belonged. The angle was determined by calculating the tangent of the two components, and the resulting angle was adjusted accordingly. For example, if the U component was positive and the V component was negative, the vector falls into quadrant 2 (Q2). The tangent of the two components was, therefore, subtracted from 180° to give the resultant wind direction.

After the annual, summer and winter mean wind speed and direction had been determined for the 0.5° grid, the grid was reprojected into the Lambert Azimuthal Equal Area projection and intersected with the 1 x 1km base coverage to assign a wind values to every cell in the study area. The steps and AMLs (Arc Macro Language scripts) used to process the component wind vectors are presented in Appendix A2.

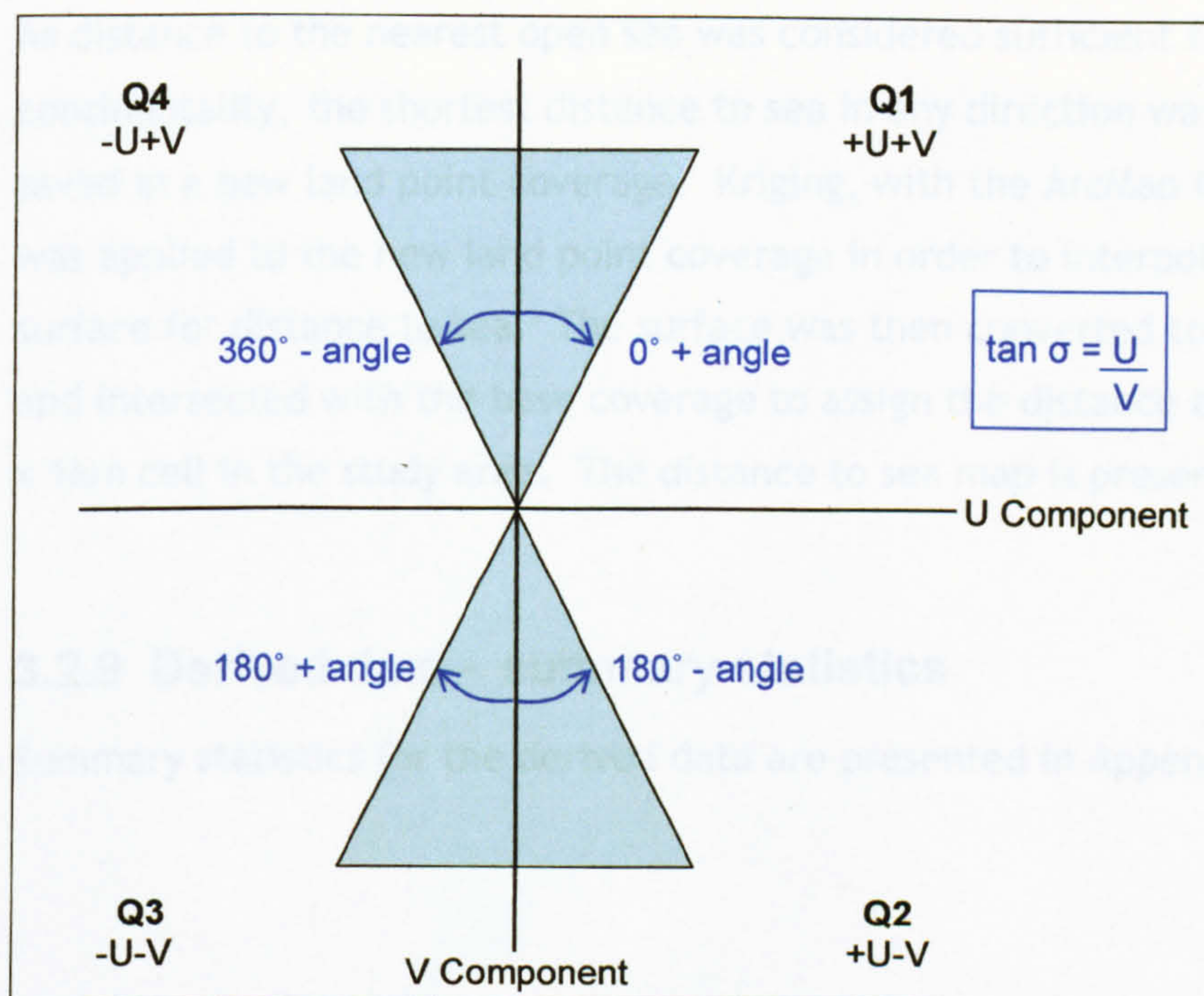


Figure 3.12 Measuring angles to calculate wind direction in ArcInfo

3.2.8 Distance to sea

Distance to sea provides information on continentality. It is included as a predictor variable for two reasons: 1) because of the effect of the sea on the local meteorology; and 2) since marine sources may be important contributors to particulate concentrations and to chemical reactions involving particulates and nitrogen dioxide in coastal areas.

The distance to the nearest sea was determined using two input point coverages: a land coverage derived directly from the base coverage, and a generalised sea coverage created by buffering 20km around the EU coastline. Generalisation was performed in order to identify areas of open sea, and avoid influence on the measure by small but deep inlets, which would be expected to have minimal meteorological relevance. Both input coverages were re-sampled to a resolution of 50km to achieve a manageable processing time (i.e. at 50km the AML took approximately 7 days of processing time). The AML ("inside_land.aml" written by Dr. L. Beale, given in Appendix A3) determines the distance between every combination of land and sea points, then selects the closest sea point for each land point in four directions (NE, NW, SE and SW).

As distance to the nearest open sea was considered sufficient for determining continentality, the shortest distance to sea in any direction was then selected and saved in a new land point coverage. Kriging, with the ArcMap GeoProcessing Wizard, was applied to the new land point coverage in order to interpolate a continuous surface for distance to sea. The surface was then converted to a polygon coverage and intersected with the base coverage to assign the distance to sea value to every 1 x 1km cell in the study area. The distance to sea map is presented in Figure 3.13.

3.2.9 Derived data – summary statistics

Summary statistics for the derived data are presented in Appendix A4.

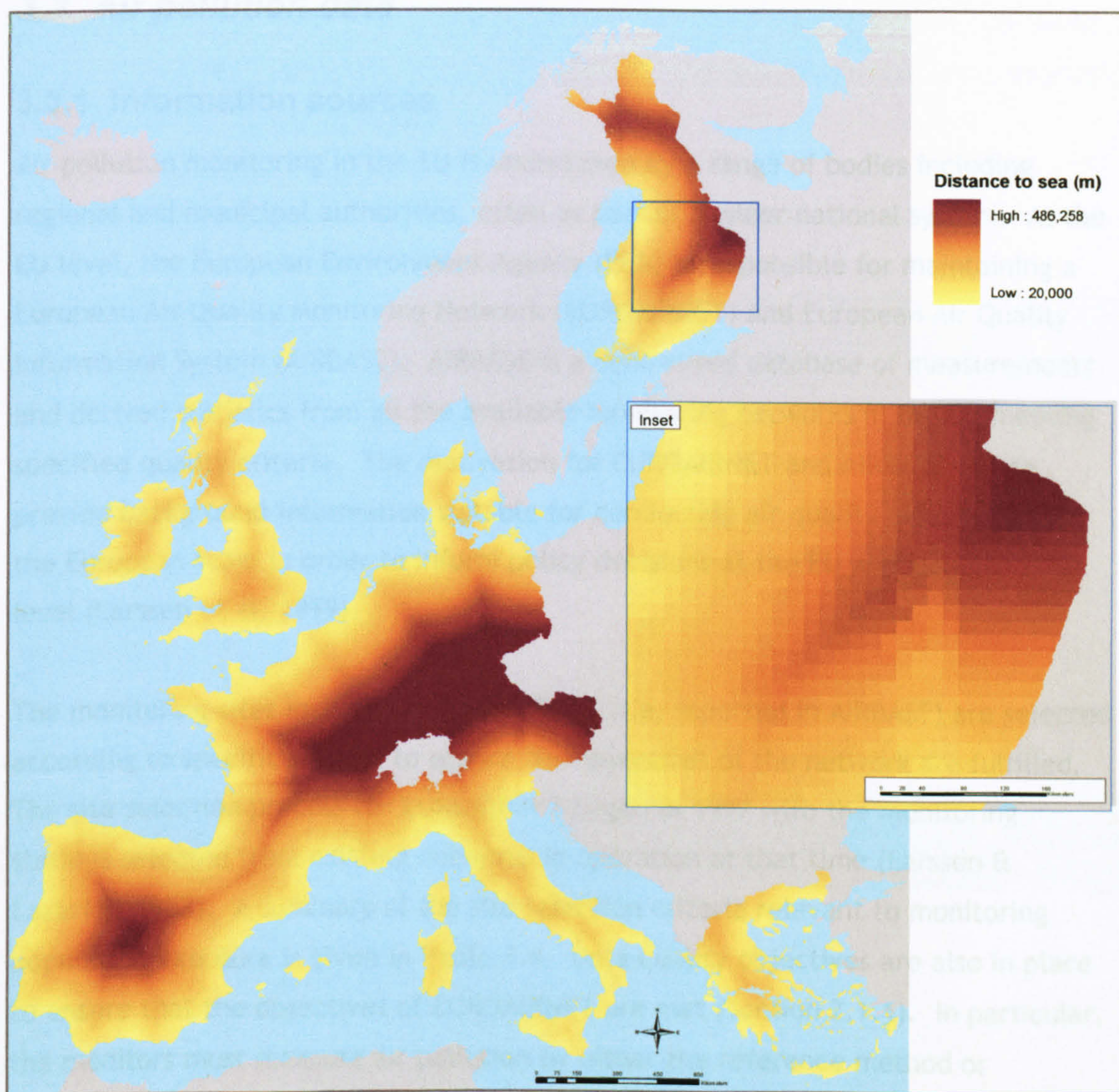


Figure 3.13 Distance to sea (m)

equivalent to the ISO 7794:1975) is the reference method for NO_x (ISO 2206) and EN 12341 is the European reference method for PM_{10} (CEN 1998). Additionally, the measuring devices should be continuously operating and report 1-hour (e.g. for SO_2 , NO_x , O_3) or 24-hour (e.g. for PM_{10}) averages to fully evaluate the measured levels relative to the air quality guideline limits (Larsen et al. 1999).

The air pollution risk assessment data used in this study were primarily taken from AIRBASE. Hourly and annual statistics for NO_x and PM_{10} between 1997 and 2001 were downloaded via the Airbase online Air quality Visualisation Instrument for Europe on the Web from the European Topic Centre on Air and Climate Change website (ETC 2004). The data were downloaded several times in the course of the study as new or

3.3 Air pollution data

3.3.1 Information sources

Air pollution monitoring in the EU is undertaken by a range of bodies including regional and municipal authorities, often as part of a wider national system. At the EU level, the European Environment Agency (EEA) is responsible for maintaining a European Air Quality Monitoring Network (EUROAIRNET) and European Air Quality Information System (AIRBASE). AIRBASE is a centralised database of measurements and derived statistics from all the available monitoring networks in the EU meeting specified quality criteria. The motivation for EUROAIRNET and AIRBASE was to provide background information suitable for conducting air quality assessments at the European scale in order to inform policy decisions at the EU and Member State level (Larssen et al. 1999).

The monitoring sites included in EUROAIRNET (i.e. reported in AIRBASE) are selected according to specific criteria to ensure the objectives of the network are fulfilled. The site selection process for EUROAIRNET began in 1997 with the monitoring stations selected from existing networks in operation at that time (Larssen & Lazaridis 1998). A summary of the site selection criteria relevant to monitoring population exposure is given in Table 3.4. Data quality objectives are also in place to ensure that the objectives of EUROAIRNET are met (Section 2.1.1). In particular, the monitors must measure air pollution by either the reference method or equivalent according to EU legislation and/or internationally accepted standards (Larssen et al. 1999). Chemiluminescence (ISO 7996: 1985) is the reference method for NO₂ (ISO 2006) and EN-12341 is the European reference method for PM₁₀ (CEN 1998). Additionally, the measuring devices should be continuously operating and report 1-hour (e.g. for SO₂, NO₂, O₃) or 24-hour (e.g. for PM₁₀) averages to fully evaluate the measured levels relative to the air quality guideline limits (Larssen et al. 1999).

The air pollution concentration data used in this study were primarily taken from AIRBASE. Hourly and annual statistics for NO₂ and PM₁₀ between 1997 and 2001 were downloaded via the AirView applet (Air quality Visualisation Instrument for Europe on the Web) from the European Topic Centre on Air and Climate Change website (ETC 2004). The data were downloaded several times in the course of the study as new or

improved data were posted. A third and final download, conducted in autumn 2003, provided the air pollution concentration data used in the study. It should be noted that no data for Luxembourg, for the pollutants and dates of interest, were available at the time of the last download.

Table 3.4 EUROAIRNET site selection for assessment of population exposure
(Larssen et al. 1999)

Selection of areas to be monitored (assessment of population exposure)		
Type of area	Area selection	Station selection
>0.5 million persons	All cities	All stations up to 20 in the agglomeration. When >20, all station categories in the city must be represented and be spatially distributed to cover the whole population
0.25-0.5 million persons	At least 25% of the cities	Selected cities must represent high, medium and low industrialisation levels.
0.05-0.25 million persons	At least 10% of the cities	Selected cities must represent high, medium and low industrialisation levels.
Rural areas		To be selected in consultation with European Topic Centre on Air Quality. At least 50% of the population should be well represented for relevant compounds (e.g. O ₃ , PM ₁₀ , PM _{2.5}).
Classification of monitoring stations		
Type of station	Type of zone	Characterisation of zone
Traffic	Urban	Residential
Industrial	Suburban	Commercial
Background	Rural	Industrial
		Agricultural
		Natural
		Residential/Commercial
		Commercial/Industrial
		Industrial/Residential
		Residential/Commercial/Industrial
		Agricultural/Natural
Selection of compounds and materials (assessment of population exposure)		
Priority 1	Priority 2	Priority 3
1h (24h): SO ₂ , NO ₂ , NO _x , O ₃	1h: CO	Other compounds
1h or 24h: PM ₁₀ , PM _{2.5}	1h or 24h: SMP (for TSP), black smoke	
24h or longer: Pb	24h or longer: benzene, PAH, Cd, As, Ni, Hg	

Notes: Relevant station types for population exposure include: traffic, industrial, urban/suburban background, near-city background and regional background stations.

In order to enhance the AIRBASE data, additional data from national networks were sought through an extensive Internet search. In general, however, relatively few additional sites were found as in most cases data from national sources were already contained within AIRBASE. Although many municipal and regional networks do exist, access to the data is usually possible only through direct consultation with the local network managers - a lengthy and often bureaucratically complex process, which was not feasible in this study. Additional concentration data were, thus, limited to the UK National Air Quality Information Archive non-automatic network (approximately 1,100 NO₂ diffusion tubes and 223 SO₂ monitors), and the EMEP (European Monitoring Evaluation Programme) measurement network.

EMEP is an acronym for the Co-operative Programme for Monitoring and Evaluation of the Long-Range Transmission of Air Pollutants in Europe, which is a policy driven programme designed to address issues related to transboundary air pollution. Launched in 1977, the EMEP programme initially focused on the transboundary transport of acidification and eutrophication then expanded to address the formation of ground level ozone. More recently assessment of persistent organic pollutants, heavy metals and particulate matter has been included. The EMEP network currently comprises more than 250 stations situated in rural areas across 35 countries covering the European continent (EMEP 2003). After removing duplicate sites, the EMEP data were integrated into the core air quality data set while the additional UK data were reserved for specific model development and testing in the UK.

3.3.2 Data limitations

During the initial development of AIRBASE, the metadata and time series information was obtained from the GIRAFE and APIS databases, respectively. GIRAFE was created in 1990 and contained information for 12 Member States. Unfortunately, the metadata in GIRAFE were often outdated, non-standardised and contained obvious errors (Sluyter et al. 1997). Many of the hindrances encountered with AIRBASE during this study are said by the European Topic Centre to have originated from this stage of its development.

Checking and cleaning of AIRBASE and incorporation of the EMEP data posed several problems. The first of these derived from the lack of a unique key in AIRBASE to link the site information (e.g. station coordinates, names and metadata) to the files

containing the air pollution data. As a result, a combination of four attributes in the metadata file had to be used to recreate the station name found in the data files. Inconsistencies in naming of sites and in the formats used (e.g. abbreviations, order or spelling of sites) also meant that this process could not be fully automated. For the same reasons, links to sites in the EMEP database had to be manually checked. It should be noted, however, that since extracting the AIRBASE data significant improvements have been made to the data and AirView applet. Using AirView Version 3.1, released 20 December 2004 (ETC 2004), both the data and metadata can now be downloaded and directly linked via a single unique key.

Other problems included errors in data formats (e.g. concatenation of data records) and gaps or inconsistencies in the reported concentration data. Missing or incorrect geographical coordinates, as depicted in Figure 3.14, have seemingly arisen due to the complex history of AIRBASE derived, in part, from GIRAFE and APIS, both of which contained coordinate errors. Sluyter et al. (1997) report that coordinate errors in AIRBASE were resolved through reference to an atlas. Lack of detail in the information taken from an atlas, however, has meant that all stations with missing coordinates in a particular city were assigned the same coordinates. In preparing the data for this study, sites with common coordinates thus had to be located and checked to ensure that they were not genuine duplicates (sometimes with different names) and then reallocated to their correct coordinates. This process was undertaken using a digital gazetteer and successfully located the large majority of ambiguous sites. Consequently, less than 1% of sites had to be deleted from the database due to missing or incomplete coordinates.

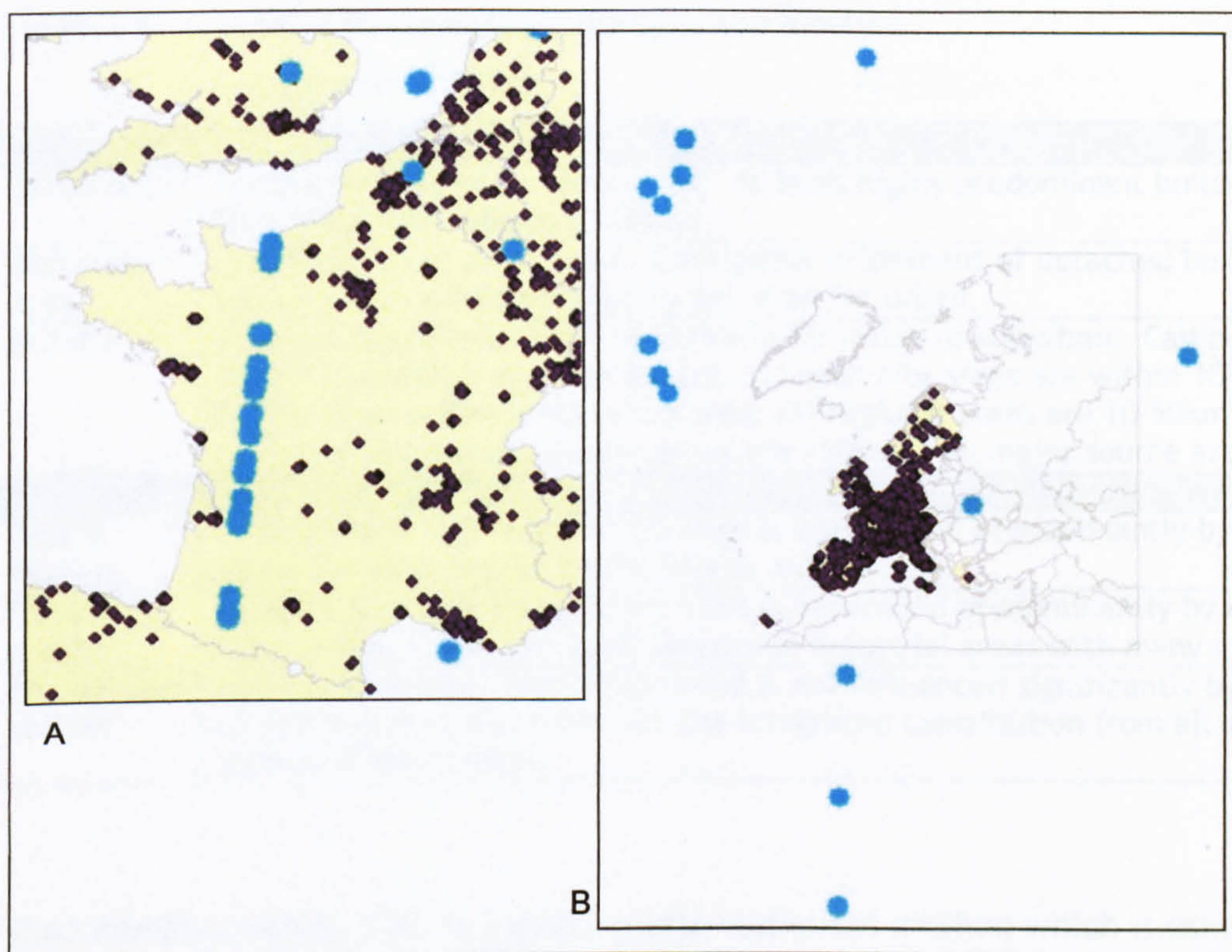


Figure 3.14 AIRBASE sites with incorrect coordinates
Incorrect coordinates are highlighted in blue for (A) sites along 0° longitude in France and those falling outside the designated country; and (B) EU sites falling completely outside the study area.

3.3.3 Classification of site location

In AIRBASE, the monitoring sites are classified by location (urban, suburban, rural, unknown) and type (traffic, industrial, background and unknown) according to the definitions listed in Table 3.5. Initially, the database included a substantial number of sites with an unknown location and/or type. Mapping of the sites in relation to the land cover data also suggested that the locational classification was not always accurate, and had been differentially applied in different countries, especially in the case of suburban sites. For these reasons, the locations of the monitoring sites were reclassified on the basis of land cover data, using discriminant analysis in SPSS 12. This also enabled the classification of the 1 x 1km grid cells into urban and rural subsets.

Table 3.5 AIRBASE definitions: site type and location
(Garber et al. 2002)

Site type	
Urban area	Continuously built-up urban area. At least highly predominant building-up of the street front side by buildings.
Suburban area	Largely built-up urban area. Contiguous settlement of detached buildings of any size with a building density less than for urban.
Rural area	All areas that do not fulfil the criteria for urban or suburban. Can be subdivided based on distance to major source: (1) near-city areas are within 10km from the border of an urban or suburban area; (2) regional areas are 10-50km from major source areas; and (3) remote areas are >50km from major source areas.
Location	
Traffic station	Located such that its pollution level is determined predominantly by the emissions from nearby traffic (roads, motorways).
Industrial station	Located such that its pollution level is influenced predominantly by emissions from nearby single industrial sources or industrial areas with many sources.
Background station	Located such that its pollution level is not influenced significantly by any single source or street, but rather by the integrated contribution from all sources upwind of the station.

Discriminant analysis (DA) is a multivariate statistical method which is used to estimate the relationship between a single categorical dependent variable (in this case site location) and a set of metric independent variables. The aim of DA is to maximise the between-group variance while minimising the within-group variance, in relation to the independent variables. The discriminant function, a linear combination of the weighted variables (Equation 3.1), is said to separate the groups well when the between-group variance is large compared to the within-group variance. When good separation is achieved the discriminant function can then be used to predict group membership for unassigned objects.

Equation 3.1 $D = B_0 + B_1X_1 + B_2X_2 + ... + B_pX_p$

Where: X's are the values of the independent variables; and B's are coefficients estimated from the data (Norusis 1988).

If the dependant variable is made up of N groups, N-1 discriminant functions can be determined. The estimation of more than one discriminant function is beneficial as it may provide additional information on which combinations best discriminate among the groups. The first discriminant function has the largest ratio of between-group to within-group variance, and the second function is uncorrelated to the first and has the next largest between to within group variance, and so on. For each object, the discriminant function is used to calculate a predicted value called the discriminant Z score. When there are two or more discriminant functions, each is

tested for significance and the significant functions are used to calculate individual discriminant Z scores for all objects. To assess the overall model fit, Z scores can be averaged on a group basis to calculate the centroid. A successful DA is one that has well-separated groups and will result in significantly different group centroids.

An SPSS file for the 2.8 million 1 x 1km cells and land cover variables was created, and those cells containing monitoring sites were flagged. It should be noted that this analysis was based on all monitoring sites used in the APMoSPHERE project (including NO₂, PM₁₀, SO₂, O₃ and CO sites) for 2001; thus, there were approximately 540 rural, 1,200 urban and 940 suburban/unknown sites. The AIRBASE site location was recoded into rural = 1, urban = 2 and suburban/unknown (i.e. other) = 3. Only the urban and rural sites were included in the discriminant analysis as grouping variables since the aim was also to assign all monitoring sites and 1 x 1km cells to either the urban or rural group.

Four independent variables (urban1, urban3, rural1 and rural3) were used in the discrimination. The urban variables represented built-up land (TBU + URBGR; see Table 3.3) at the 1 x 1 and 3 x 3km level, respectively, while the rural variables represented agricultural and forested areas (FOR + AGR; see Table 3.3) at the 1 x 1 and 3 x 3km level, respectively. The independent variables were assumed to be equally important to the discrimination and were, therefore, entered together, rather than by a stepwise process. Equal prior probabilities for all groups were also assigned as there is no relationship between the proportion of urban/rural monitoring sites and urban/rural land area.

The analysis resulted in the discriminant function provided in Equation 3.2. The canonical correlation coefficient for the discriminant function was 0.75. The classification results are presented in Table 3.6 and indicate that 89.3% of the rural monitoring sites and 86.4% of the urban monitoring sites were correctly classified in relation to the AIRBASE location class. This table also shows that 36% of the suburban/unknown sites were classified as rural and the remaining 64% were designated as urban.

Equation 3.2

$$D = -2.691 + rural1(0.002) + rural3(0.003) + urban1(0.028) + urban3(0.016)$$

Table 3.6 Discrimination of monitoring sites and 1km cells into urban and rural

Original site type		Predicted group membership		Total
		1 (rural)	2 (urban)	
Count	1 (rural sites)	484	58	542
	2 (urban sites)	164	1,042	1206
	3 (other sites)	335	609	944
	Unclassified 1km cells	2,787,815	64,627	2,852,442
%	1 (rural sites)	89.3	10.7	100.0
	2 (urban sites)	13.6	86.4	100.0
	3 (other sites)	35.5	64.5	100.0
	Unclassified 1km cells	97.7	2.3	100.0

Notes: 87.3% of original grouped cases correctly classified.

The box plots in Figure 3.15 provide further evidence that good separation was achieved on the basis of the selected land cover variables. The pattern of the box plots is identical for the original and predicted site locations for the AIRBASE monitors, with the rural groups being the inverse of the urban groups. As expected, the amount of rural land is greater in the rural group (group 1) and likewise the amount of urban land is greater in the urban group (group 2). The interquartile ranges for the predicted site locations are also slightly smaller than those of the original site locations, indicating that this analysis marginally increases concordance between the monitoring site location group and the land cover data in the GIS.

Given these results, the discriminant function (Equation 3.2) can be said to have achieved good separation of monitoring sites and was, thus, considered adequate for classifying the 1 x 1km cells into either urban or rural. Cases with $D \leq -0.466$ (the maximum discriminant score for group 1) were assigned to the rural group and all others were assigned to the urban group. The percentage of urban and rural area, by country, achieved through this analysis is presented in Table 3.7.

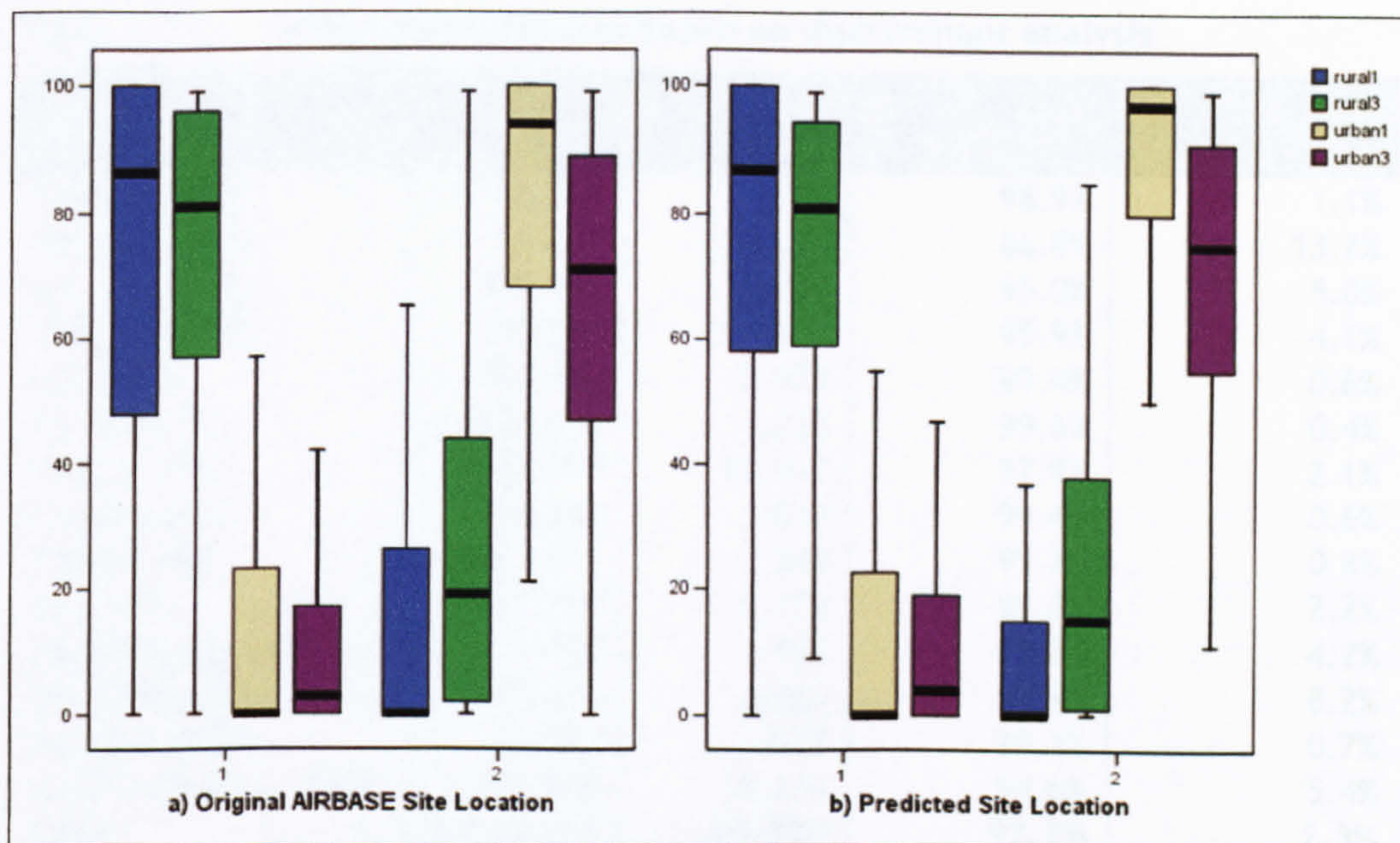


Figure 3.15 Discrimination of site location by land cover
 (A) Original Reported AIRBASE Site Location; and (B) Predicted Site Location.
 Notes: 1 = Rural and 2 = Urban

Unlike site location, the AIRBASE site type classifications were maintained as suitable predictors were unavailable in the GIS to facilitate reclassification by discriminant analysis. As shown in Table 3.8, approximately 2% of the 2001 NO₂ and PM₁₀ sites were classified as unknown.

Table 3.7 Urban and rural area based on discriminant analysis

Country	Area (1km ²)		Percent Area (%)	
	Rural	Urban	Rural	Urban
Austria (AT)	82,956	925	98.9%	1.1%
Belgium (BE)	26,587	4,033	86.8%	13.2%
Germany (DE)	339,449	17,747	95.0%	5.0%
Denmark (DK)	43,189	1,834	95.9%	4.1%
Spain (ES)	496,153	2,928	99.4%	0.6%
Finland (FI)	333,678	1,225	99.6%	0.4%
France (FR)	537,699	11,412	97.9%	2.1%
Greece (GR)	131,222	810	99.4%	0.6%
Ireland (IE)	69,591	543	99.2%	0.8%
Italy (IT)	294,561	6,779	97.8%	2.2%
Luxembourg (LU)	2,487	108	95.8%	4.2%
the Netherlands (NL)	32,435	2,880	91.8%	8.2%
Portugal (PT)	88,322	622	99.3%	0.7%
United Kingdom (UK)	231,840	13,234	94.6%	5.4%
Total	2,710,168	65,080	97.7%	2.3%

Table 3.8 Number of NO₂ and PM₁₀ monitoring sites by year and site type

Year	Reporting countries	Total # stations	Traffic	Background		Industrial	Unknown
				Urban	Rural		
NO ₂							
1997	14	924	279	353	220	56	16
1998	12	1,098	392	380	226	76	24
1999	15	1,505	453	596	279	143	34
2000	15	1,578	470	630	305	142	31
2001	15	1,781	544	670	348	188	31
PM ₁₀							
1997	8	115	42	50	14	9	-
1998	9	200	64	106	19	11	-
1999	12	256	108	90	36	22	-
2000	12	419	153	159	71	31	5
2001	15	911	302	387	134	78	10

Notes: Total number of reporting countries was 15 (including Norway and excluding Luxembourg).

3.3.4 Spatial and temporal coverage

Once the concentration data were compiled, it was obvious that the 2001 data provided much better coverage compared to the earlier years. The number of monitoring sites by year and type for the enhanced AIRBASE (i.e. AIRBASE plus EMEP, referred to as AIRBASE-e) database are given in Table 3.8. This table indicates a two-fold increase in the number of NO₂ monitoring sites and eight-fold increase in

PM₁₀ sites between 1997 and 2001. For this reason, only the 2001 sites were used in pollution modelling.

The Exchange of Information Decision (Decision 97/101/EC amended by Decision 2001/752/EC) states that a minimum data capture of 75% is required for calculating 1-hour values for pollutant concentrations (Garber et al. 2002). Only sites meeting this requirement, therefore, were used in the pollution modelling. Table 3.9 shows both the total number of recording sites, and those with greater than or equal to 75 percent data capture, by country for 2001. When only sites achieving the data capture criteria were included, there was a 9 and 16% reduction in the number of NO₂ and PM₁₀ sites, respectively, available for pollution modelling.

Table 3.9 **Number of sites in 2001 with >= 75% data capture by country**
All site types, AIRBASE-e

Country	Number of NO ₂ monitoring sites		Number of PM ₁₀ monitoring sites	
	Total	With >=75% data capture	Total	With >=75% data capture
AT	147	143	53	42
BE	42	39	23	21
DE	463	443	305	272
DK	10	8	5	2
ES	271	236	129	107
FI	22	21	25	25
FR	436	390	197	145
GR	27	25	14	12
IE	6	6	6	5
IT	166	137	54	36
NL	44	44	19	19
PT	23	23	11	10
UK	97	89	62	60
Total	1,754	1,604	903	756

The concentration data were incorporated into the GIS as point locations, represented by the geographic coordinates for the monitoring sites. The 2001 annual mean NO₂ and PM₁₀ concentrations at the monitoring sites are depicted in Figures 3.16 and 3.17, respectively. Even in 2001, and with the addition of the EMEP sites, the spatial coverage of monitoring stations is far from uniform across the EU. This unevenness in site distribution is an important consideration when using the data for pollution modelling, and was also explored in more detail as part of the affinity zone stratification analysis, described in Chapter 5.

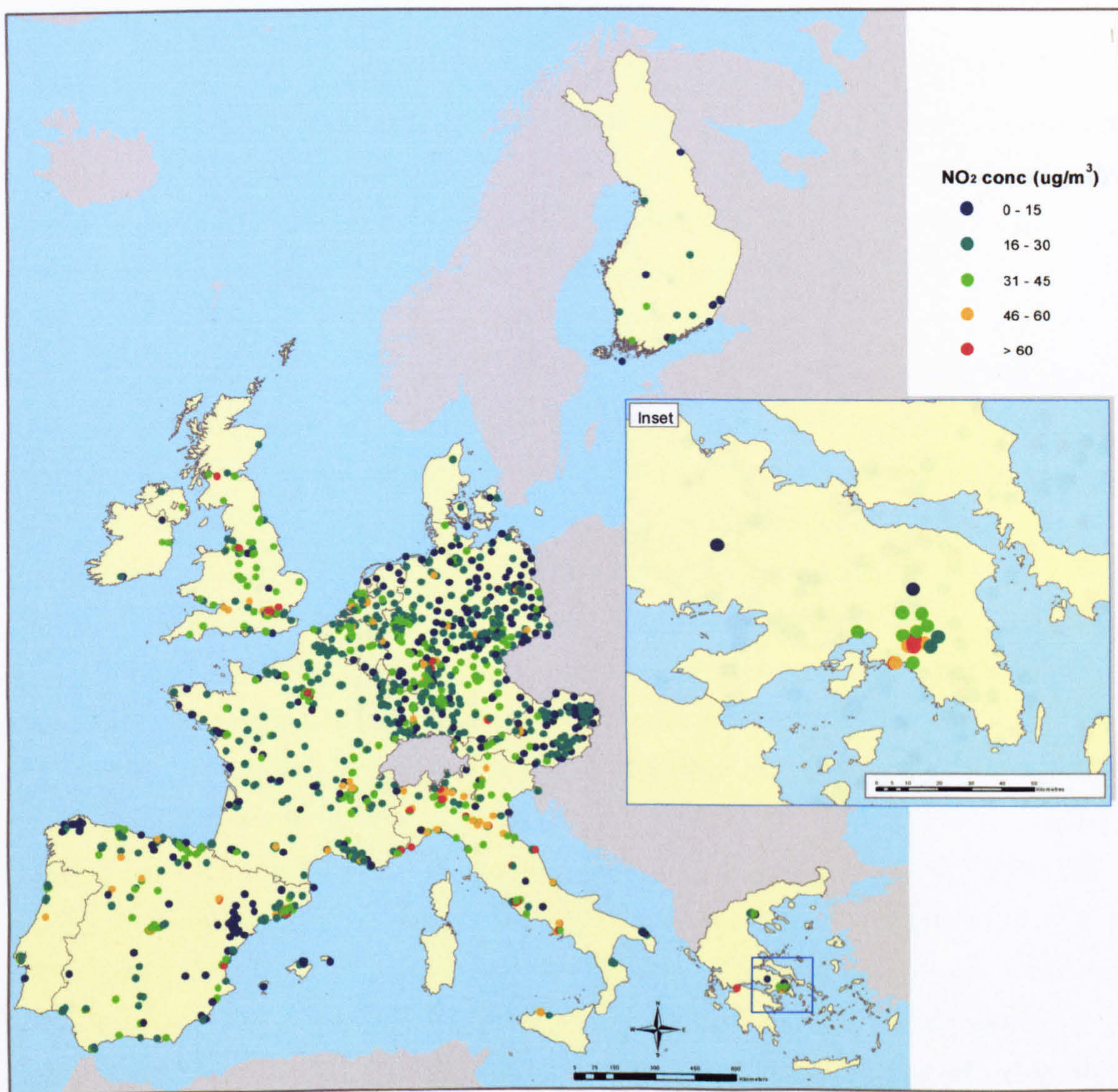


Figure 3.16 2001 annual mean NO₂ concentration (µg/m³)
AIRBASE-e monitoring sites with ≥75% data capture

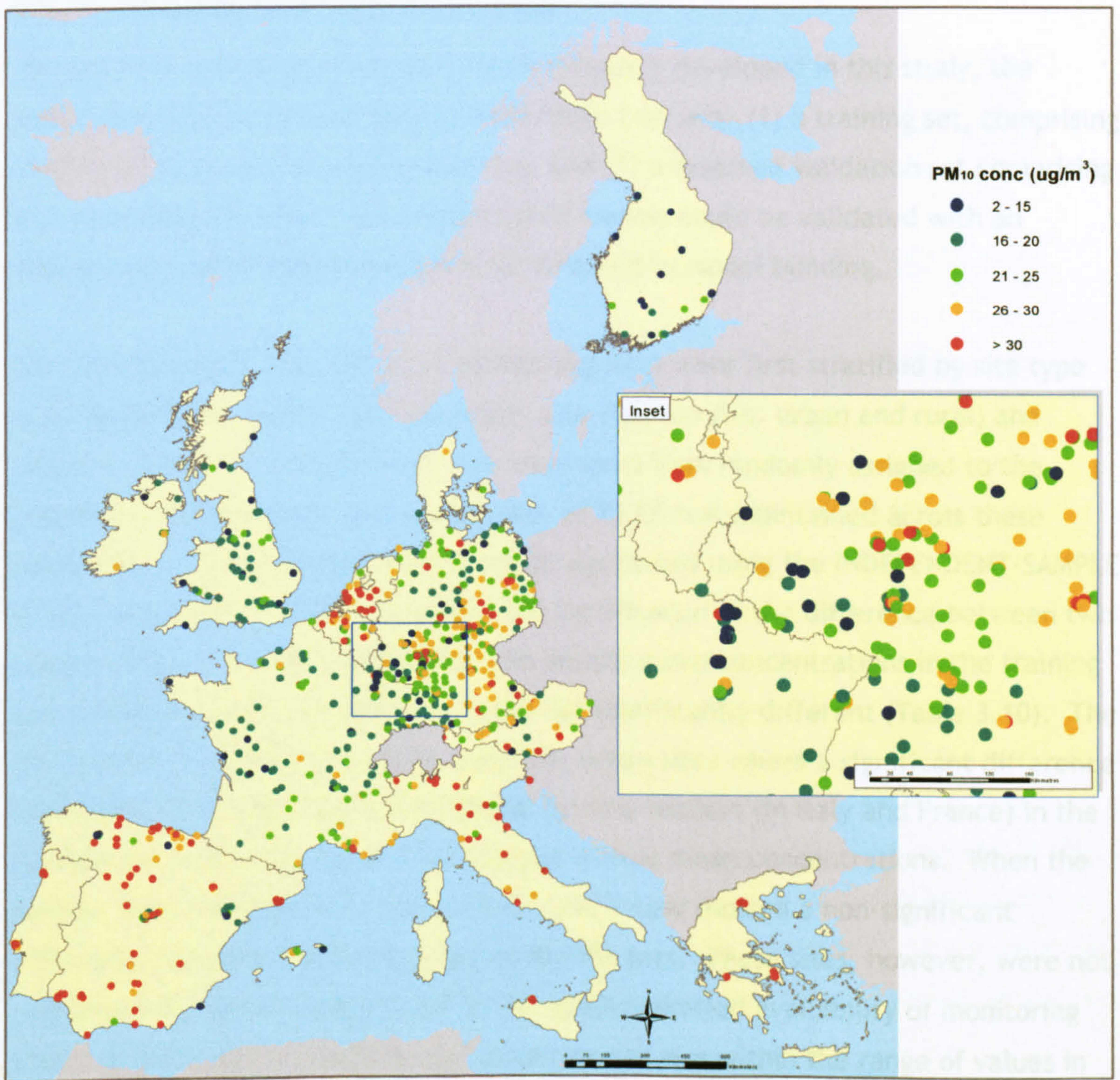


Figure 3.17 2001 annual mean PM₁₀ concentration (µg/m³)
AIRBASE-e monitoring sites with ≥75% data capture

3.3.5 Training and validation data

To facilitate validation of the air pollution models developed in this study, the monitoring sites were randomly stratified into two sets: (1) a training set, comprising 75% of the sites used for model building; and (2) a reserved validation set comprising the remaining 25% sites. This ensured that models could be validated with an independent set of monitoring sites not involved in model building.

For each pollutant, the full set of monitoring sites were first stratified by site type (i.e. background, traffic and industrial), site location (i.e. urban and rural) and country. Using this stratification, the sites were then randomly assigned to the training or validation set so that the ratio of 75:25 was maintained across these parameters. The training/validation split was tested using the INDEPENDENT-SAMPLE T-TEST procedure in SPSS which tests the significance of the difference between two sample means. For most pollutants, the annual mean concentrations in the training and validation strata, by site type, were not significantly different (Table 3.10). The only exception was for the background NO₂ urban sites where a significant difference was found. This result can be explained by nine outliers (in Italy and France) in the training set which had higher than average annual mean concentrations. When the outliers were removed from this analysis, the T-test showed a non-significant difference between the training and validation sets. These sites, however, were not excluded from model building due to the already limited availability of monitoring sites. As the range of values in the validation set was within the range of values in the training set, this decision was considered justified.

Once defined, the training and validation sets were unchanged throughout the duration of the study. The final stratification applied to the monitoring sites, by type and location, is shown in Table 3.10.

Table 3.10 Monitoring sites by type and location
 (Sites recording in 2001 with $\geq 75\%$ data capture)

Site Type	Training N		Validation N		Total N	T-test Significance ¹		
	Rural	Urban	Rural	Urban		Rural sites	Urban sites	Combined sites
NO ₂								
Background	232	482	77	151	942	0.787	0.025	0.097
Traffic	41	330	10	119	500	0.813	0.148	0.106
PM ₁₀								
Background	98	228	31	73	430	0.639	0.950	0.945
Traffic	14	179	5	53	251	0.539	0.224	0.165
SO ₂								
Background	251	414	81	139	885	0.605	0.487	0.420
Traffic	33	225	13	83	354	0.445	0.823	0.066

Notes: (1) $\alpha = 0.05$
 Only sites with $\geq 75\%$ data capture were used in air pollution modelling. SO₂ was explored in focalsum modelling (Chapter 6).

3.4 Conclusion

Database development plays an important role in any GIS-based analysis. The importance of a reliable, well-built GIS database cannot be over-emphasised in this particular study where it was the foundation for developing estimates of population exposure to air pollution. Particular attention, therefore, was given to the selection of appropriate input data to facilitate the spatial modelling of population density and air pollution surfaces.

As was demonstrated in this chapter, GIS development involved significant pre-processing to incorporate the input data which originated from various sources into a common geography. During GIS development several issues related to differing geographies (e.g. projection, intersection and area-weighting), data quality (e.g. lack of unique codes to link attribute and boundary data) and data availability were encountered. This chapter clearly details how these issues were addressed to ensure the construction of a suitable database to facilitate a GIS-based exposure assessment to air pollution across the EU.

4 Population Modelling

A consistent, high resolution population density map is essential to facilitate a GIS-based assessment of human exposure to air pollution - and indeed for many other applications in environmental epidemiology and risk assessment. The area of research described in this chapter explores a GIS-based approach to redistribute census-tract population to a 1km grid across the study area. This is achieved by disaggregating NUTS 5 population totals using land cover and night-time light intensity variables to reflect small-area variations in population density. The population map thus produced is validated using independent population data, where available, prior to using it as a basis for an EU-wide exposure assessment (Chapter 7) to identify potential locations where populations may be exposed to above guideline concentrations of traffic-related air pollutants.

4.1 Introduction

4.1.1 Need for high resolution population data

Reliable information on the location and distribution of the population is essential in many epidemiological studies and, in particular, for exposure and risk assessment where it is needed to facilitate the calculation of population-level exposures. Census data provide regularly updated information on population counts and demographics in the EU and elsewhere around the world. The way in which census tracts are defined, however, can limit the value of these data for assessing exposure.

Census tracts are designed to be all-inclusive to facilitate enumeration, rather than to reflect the underlying geographic distributions of population. As such, they are rarely uniform in size or population numbers, and typically cover large geographic areas including unpopulated lands (e.g. rural land, parks and industrial developments). The smallest geographic unit for the EU administrative regions, for example, is the Nationales Unites Territoriales Statistiques level 5 (i.e. NUTS 5, now referred to as Local Administrative Unit or LAU 2). Outside densely built-up urban centres most NUTS 5 regions are around 5-10 km². In rural areas, the NUTS 5 regions may be in excess of 100km². Where the NUTS 5 regions are this large, the population tends to be dispersed or clustered in small towns with inhabitants typically residing in only a small proportion of the region.

NUTS 5 regions also vary greatly in area and population size from one country to another. Table 4.1 shows summary statistics for 1991; Figure 4.1 shows population density mapped on the basis of these data. In addition, census tracts (and other administrative regions) are usually not stable over time, but are adjusted at intervals to reflect changes in population pattern and administrative structures. An example of this was quoted in Section 3.2.3, in relation to population data: changes in the census geography between 1991 and 2001, for example, made it difficult to match population data to the available administrative boundaries. This was a particular problem in Germany where reunification of the country in the early 1990s led to a mismatch in the population statistics and administrative boundary data.

Table 4.1 Summary statistics for NUTS 5 census data (1991)

Country	N	Area (km ²)		Population	
		Mean	SD	Mean	SD
Austria	2,358	35.6	37.5	3,312.8	32,738.1
Belgium/Luxembourg	708	46.9	36.8	14,636.5	26,367.4
Denmark	277	156.2	101.3	18,579.3	36,070.5
Finland	434	721.5	832.5	11,190.3	30,265.2
France	36,570	15.0	15.3	1,548.5	13,938.0
Great Britain ¹	10,528	21.8	61.9	5,204.1	3,894.9
Germany (East) ²	114	953.3	736.3	153,948.4	317,426.4
Germany (West) ³	8,789	28.3	34.1	7,458.4	35,468.9
Greece	1,034	127.9	107.7	9,938.0	28,216.5
Ireland/N. Ireland	3,990	21.2	18.8	1,262.9	1,550.5
Italy	8,109	37.2	50.8	6,998.1	42,428.6
Netherlands	505	82.1	91.2	29,447.9	51,724.1
Portugal ⁴	4,007	22.3	36.1	2,335.5	4,833.3
Spain ⁵	8,094	61.7	93.0	4,603.9	43,768.9
EU⁶	85,403	31.0	91.9	3,965.9	25,480.9

Notes: (1) Based on 1991 census and ward data; (2) Based on NUTS 3 regions; (3) Based on 2003 population data; (4) Excludes Madeira; (5) Excludes Canary Islands; (6) Excludes eastern Germany

In response to these problems, several modelling methods have emerged to estimate small-area population for various applications. Following a brief review of current population modelling techniques, this chapter presents the methods and results for a GIS-based approach using remotely sensed land cover and night-time light intensity data.

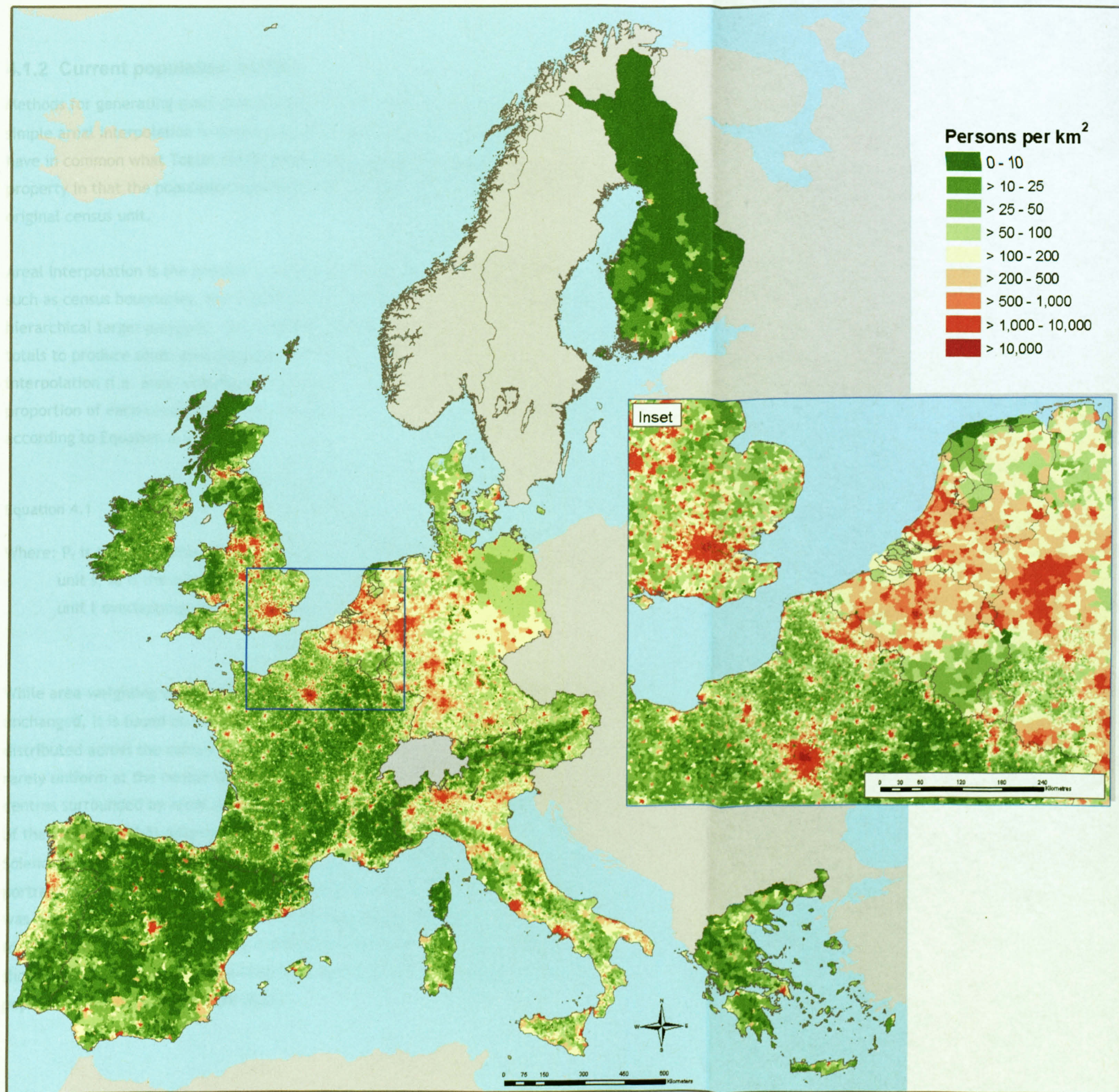


Figure 4.1
EU population density based on NUTS 5 regions

4.1.2 Current population models

Methods for generating small-area population estimates range in complexity from simple areal interpolation to dasymetric mapping techniques. The approaches all have in common what Tobler (1979) termed the *pycnophylactic*, or mass-preserving, property in that the population estimates are conditioned to sum to that of the original census unit.

Areal interpolation is the process by which data from one set of source polygons, such as census boundaries, are re-districted onto a set of overlapping but non-hierarchical target polygons. This method is most often used to redistribute census totals to produce small-area population estimates. Mass-preserving areal interpolation (i.e. area-weighting) redistributes the population based on the proportion of each census tract that overlaps with the target map unit (Figure 4.2) according to Equation 4.1.

$$\text{Equation 4.1} \quad P_t = \sum_{s=1}^S \frac{P_s \times A_{ts}}{A_s}$$

Where: P_t is the population in target map unit t ; P_s is the population in source map unit s ; A_s is the area of source map unit s ; and A_{ts} is the area of target map unit t overlapping source map unit s .

While area-weighting ensures that the total population from the source data remains unchanged, it is based on the often incorrect assumption that population is evenly distributed across the census tracts. Most populations, including that of the EU, are rarely uniform at the census level, and instead tend to be highly clustered in urban centres surrounded by areas of dispersed rural homesteads. The Gridded Population of the World (GPW-3) database, developed by the Center for International Earth Science Information Network (CIESIN), claims to be “the latest development in the portrayal of global human populations in a common geo-referenced framework,” and was designed to provide information on the population distribution across the world for use in research, policy making and communications (CIESIN 2005). Despite its shortcomings, however, an area-weighting approach has been used to construct the popular GPW-3 database (CIESIN 2004).

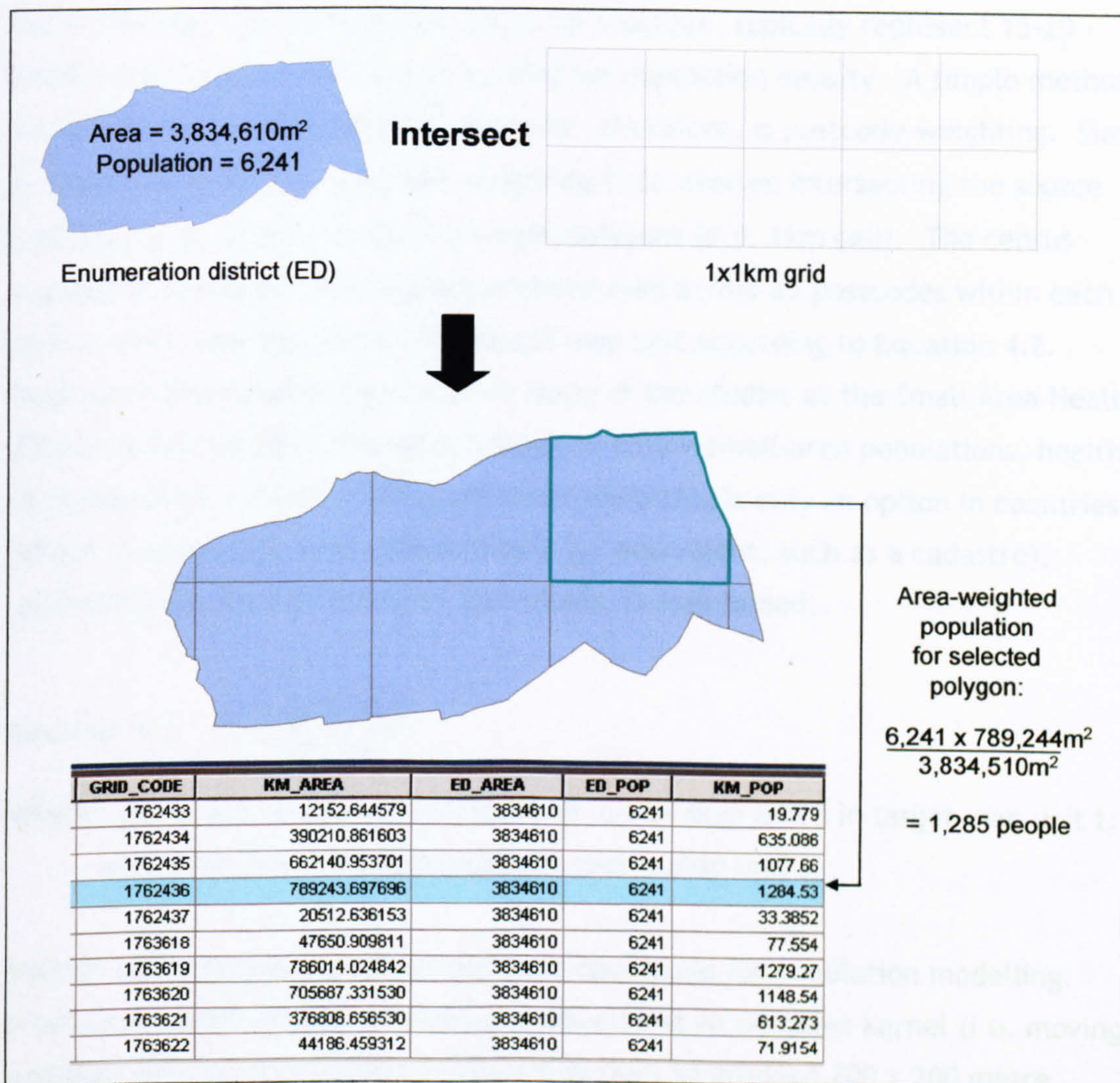


Figure 4.2 Area-weighting example

An alternative method to impute small-area population density on the basis of postcode geography has evolved in the UK where, since the late 1970s, the Office for National Statistics (ONS) has used postcodes as their main geographic reference when collecting statistical data. Disadvantages of using postcode geography include the fact that postcodes do not take account of administrative boundaries, and that they can move around, be discarded or reassigned as the Royal Mail maintains their mail service (ONS 2003). There are, however, two main advantages to using postcodes as the geographic reference for statistical data. The first is that the majority of postcodes are precise to the ca. 1 metre level (ONS 2005), and second, most people know their postcode which may ease the compilation of statistical data. Additionally, headcounts for each postcode are also now available in the UK based on the 2001 census data.

Given that the 1991 census postcodes, for example, typically represent 15-20 households, they can be used as a proxy for population density. A simple method to compute small-area population estimates, therefore, is postcode-weighting. Similar to areal interpolation, postcode-weighting first involves intersecting the source polygons (e.g. census tracts) and target polygons (e.g. 1km cell). The census population totals are then equally proportioned across all postcodes within each census tract, and summed to the target map unit according to Equation 4.2. Postcode-weighting has been used in many of the studies at the Small Area Health Statistics Unit (SAHSU), Imperial College to obtain small-area populations, health or socioeconomic indices. Clearly postcode-weighting is only an option in countries where an up-to-date postcode database (or equivalent, such as a cadastre), representing a known density of individuals, is maintained.

Equation 4.2
$$P_t = \sum_{s=1}^S \frac{P_s \cdot pc_{ts}}{pc_s}$$

Where: pc_{ts} is the number of postcodes in source map unit s in target map unit t ; and pc_s is the number of postcodes in source map unit s .

Various other approaches have also been developed for population modelling. Bracken & Martin (1989), for example, developed an adaptive kernel (i.e. moving window) method using distance-decay functions to produce 200 x 200 metre estimates from population-weighted enumeration district (ED) centroids. The size of the kernel, which influences the possible number of cells across which the centroid population may be attributed, was determined locally based on the centroid density. Cells closest to the ED centroid, therefore, received the highest contribution, while those falling outside the kernel remained unpopulated. Although this method was an improvement over simple areal interpolation, it was based on the assumption that the ED centroids represent 'high information points' (i.e. the highest population density for each census tract). Further research by Harris & Longley (2000) discussed the questionable accuracy of the population-weighted centroids in the Bracken & Martin (1989) study. The centroids apparently were determined visually at the census agency, and had a respective resolution of ca. 10 and 100 metres in urban and rural areas. Using higher resolution postcode data, therefore, Harris & Longley (2000) were able to achieve improved estimates and generated a 48 x 48 metre population grid using the same density estimation method. Outside the UK, however,

census information at the postcode level is not necessarily accessible. This has led to the emergence of alternative methods including dasymetric mapping.

Dasymetric mapping is a type of areal interpolation that incorporates ancillary data to facilitate the areal interpolation process. It differs from choropleth mapping (e.g. the NUTS 5 population density map shown in Figure 4.1) in that, rather than defining areas on the basis of administrative or other arbitrary units, the boundaries derive from the actual spatial distribution of the variable being mapped. This technique has gained popularity with the increasing availability of satellite imagery and improved methods of utilising Earth observed (EO) data in a GIS. The simplest approach is binary dasymetric mapping (Equation 4.3) in which each source unit is divided into two sub-regions - populated and unpopulated - and the census information is then allocated only to the populated areas (Langford 2006).

$$\text{Equation 4.3} \quad P_t = \sum_{s=1}^S \frac{P_s \cdot A_{tsp}}{A_{sp}}$$

Where: A_{tsp} is the area of populated land that overlaps between the target map unit t and source map unit s ; and A_{sp} is the area of target map unit s that is populated land.

Land cover data, in particular, offer a means by which residential areas can be distinguished from non-residential areas. These data also allow more sophisticated approaches to mapping, using more than two classes - polycategorical dasymetric mapping. In this, the challenge becomes to devise an appropriate set of weights that can be applied to the land parcels (or other ancillary data) to reflect population density. According to Langford (2006), when regression analysis rather than selective sampling is used to determine the weights, the approach should be termed a *hybrid model* as it has elements of both dasymetric mapping and statistical modelling.

Several research studies have focused on applying polycategorical dasymetric mapping techniques to calculate small-area population estimates. Gallego & Peedell (2001) describe an iterative process by which weights were devised for disaggregating population totals on the basis of the CORINE land cover data. Communes (i.e. NUTS 5 regions) were first stratified, by comparing the commune population density to the average density of the surrounding NUTS 2 region, into one of three levels reflecting population density (i.e. dense, less dense and no urban). The method then involved

disaggregating the NUTS 2 totals using an initial set of weights, re-aggregating the population to the commune level and comparing it to the known total, computing a disagreement indicator, and adjusting the weights to reduce the disagreement. The iterative nature of this method most likely involved considerable effort in model tuning to achieve suitable weights. Research by Mennis (2002 and 2003) might offer an easier approach to implement. He first classified Landsat TM imagery for southeast Pennsylvania into three urbanization levels: high-density, low-density and non-urban. Weighting was then achieved by proportioning the census population to each grid cell based on (1) the relative difference in population densities (determined by sampling) among the three urbanization levels, and (2) the percentage of the total census block occupied by each of the three classes. This approach, however, is often limited by the lack of 'pure' source map units (i.e. those entirely filled by a single urbanization level) that are sampled to establish the weight for each class.

Using similar ancillary data, regression analysis has also been explored as a means to determine weights by which to disaggregate population to a smaller spatial unit. In a study across four counties in Arkansas, Yuan et al. (1997) used regression analysis to determine the correlation between a reclassified land cover map and population counts. They developed area-wide as well as regional regression equations using seven land cover predictor variables, and ultimately selected the regional equations as a basis for mapping. The method also involved applying an adjustment to the regression coefficients prior to mapping to remove negative estimates and to maintain the census population total across the larger area. Langford & Harvey (2001) provide another regional scale example, whereby they used regression analysis to disaggregate ward (i.e. NUTS 5) population to the ED level on the basis of the spectral signature of pixels in a Landsat TM image classified as residential land. A similar method was applied at the more local scale by Chen (2002), who defined three levels of residential density from a Landsat TM image and assessed the correlation between census dwelling counts and the residential densities in a Sydney, Australia suburb. The general consensus from all these investigations was that the ancillary data improved the accuracy of redistributing population to a resolution below that of the census tract.

The particular research presented here describes a hybrid model that incorporates ancillary data, specifically EO night-time light intensity data and CORINE land cover,

in a GIS-based regression approach to disaggregate NUTS 5 census totals to a resolution of 1km². The necessary data sets as well as the statistical and mapping approach are described in Section 4.2. The results are presented in Section 4.3 and the validation studies conducted in Great Britain are detailed in Section 4.4.

4.2 Methods

4.2.1 Data requirements

Three key data inputs were used in generating the high resolution population map for the EU: census information; land cover; and night-time light intensity. These data sets were already introduced in Section 3.2 where the pre-processing undertaken to calculate the derived variables was described. Additional comments and pre-processing details directly relevant to population modelling are included in this section.

Population data

As previously stated, census data provide the best available baseline on population, and thus give reference data against which any population model can be calibrated and validated. As described in Section 3.2.3, the most recent population data for the EU (at the time of this study) related to the 1991 census. The census information is stored in the SIRE database and, at the finest resolution, corresponds to 2001 boundaries for NUTS 5 regions maintained in the SABE database. Not all population counts could be matched to the boundaries, however, necessitating the use of 2003 NUTS 5 population data in western Germany and 1997 NUTS 3 population data and boundaries in eastern Germany. For Great Britain, the 1991 boundaries and population counts obtained from national sources were substituted for those from SABE and SIRE in order to facilitate validation against national data (Section 4.4).

Land cover data

The land cover data were obtained from the CORINE Land Cover (CLC90) data set, described in Section 3.2.2. For this application, the land cover data were re-sampled to a 200 metre cell size, in order to reduce computational demands and in recognition of underlying accuracy limits of the data (i.e. 25 hectare). All of the 44 original land cover classes were maintained for use in the population modelling.

Light intensity data

The MANTLE project, from which the light intensity data derived, was described in Section 3.2.6. These data in their original form (i.e. 200 m point estimates) comprised the light intensity information used in population modelling.

4.2.2 Modelling approach

Figure 4.3 outlines the population modelling approach. To prepare the data sets for population modelling, the predictor variables had to be computed at the NUTS 5 level (facilitated by the *data_prep.aml* in Appendix B1). This was achieved through a series of geometric intersections. The 44 light intensity point coverages (i.e. one for each land cover class) from the MANTLE project were first intersected with the 1 x 1km base coverage and the NUTS 5 boundaries so that each point was tagged according to the target grid cell (t) and the source NUTS region (s) by which it was contained. Each point was also assigned a binary 'lit' and 'unlit' code, where points with a light intensity greater than zero were considered 'lit'. Three variables (referred to as the land class-light variables) could, thus, be computed for each NUTS 5 region: Al, the number of lit points (equivalent to area of lit land); Au, the number of unlit points (equivalent to area of unlit land); and Li, the total light intensity. NUTS 5 totals for the land class-light predictors were then exported from ArcGIS as a comma separated text file and imported into SPSS 12 for statistical analysis.

As will be described in the subsequent section, a supervised stepwise regression analysis was used to develop the population models. It should be noted that this approach preserves what Tobler (1979) termed the pycnophylactic property, in that the population estimates are always conditioned to sum to that of the original NUTS region. Other than errors in the original census data, therefore, any error introduced by the hybrid model is confined to the NUTS regions (i.e. any modelling error generated in a particular NUTS region remains within that unit). Thus, to enable this condition, coefficients from the regression models were used to calculate weights for each land cover class (according to its light intensity status), and these were then used to disaggregate the NUTS 5 census totals to the 200 metre points. The point populations were then aggregated to the target 1km grid, and divided by the target cell land area, to produce the population density map. Given the availability of independent census data, validation was only possible in Great Britain.

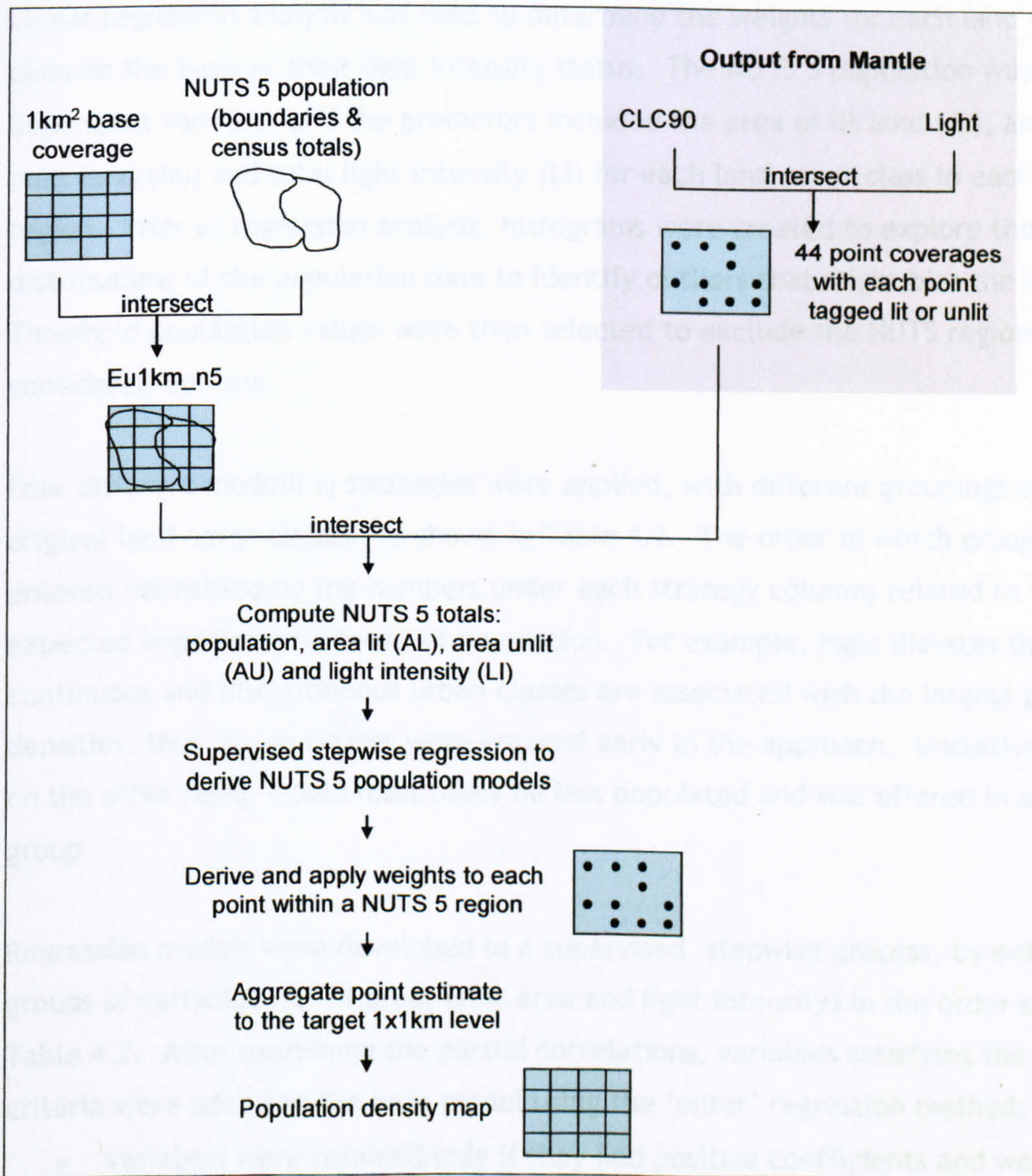


Figure 4.3 Population modelling procedure

- Within each land cover class, the lit area and light intensity variables were given priority over unlit area.
- To maintain the mediate hierarchy of importance of land cover groups, new variables were allowed to enter the model only if they did not cause the removal of an entire land cover group entered earlier in the analysis, though individual variables (e.g. AL, AU or LI) could be removed if their significance fell below 0.05.
- The final model not only had to be significant at the 0.05 confidence level, it also had to have a positive constant in order to avoid negative population predictions.

Regression modelling

Linear regression analysis was used to determine the weights for each land cover class on the basis of their light intensity status. The NUTS 5 population was the dependent variable, and the predictors included the area of lit land (Al), area of unlit land (Au) and total light intensity (Li) for each land cover class in each NUTS 5 region. Prior to regression analysis, histograms were created to explore the distributions of the population data to identify outliers that might bias the models. Threshold population values were then selected to exclude the NUTS regions considered outliers.

Four different modelling strategies were applied, with different groupings of the original land cover classes, as shown in Table 4.2. The order in which groups were entered (identified by the numbers under each strategy column) related to their expected importance in terms of population. For example, logic dictates that the continuous and discontinuous urban classes are associated with the largest population densities, thus, these classes were entered early in the approach. Uncultivated land, on the other hand, would most likely be less populated and was offered in a later group.

Regression models were developed in a supervised, stepwise process, by entering groups of variables (for lit area, unlit area and light intensity) in the order shown in Table 4.2. After examining the partial correlations, variables satisfying the following criteria were added to the base model using the 'enter' regression method:

- Variables were retained only if they had positive coefficients and were significant at the 0.05 confidence level.
- Within each land cover class, the lit area and light intensity variables were given priority over unlit area.
- To maintain the predefined hierarchy of importance of land cover groups, new variables were allowed to enter the model only if they did not cause the removal of an entire land cover group entered earlier in the analysis, though individual variables (e.g. Al, Au or Li) could be removed if their significance fell below 0.05.
- The final model not only had to be significant at the 0.05 confidence level, it also had to have a positive constant (in order to avoid negative population predictions).

Table 4.2 Regression modelling strategy

Land Cover Class(es)	Strategy 1		Strategy 2		Strategy 3		Strategy 4	
1	2	Continuous urban	2	Continuous urban	2	Continuous urban	1	Residential (res)
2	1	Discontinuous urban	1	Discontinuous urban	1	Discontinuous urban		
49	3	Unclassified surfaces	3	Unclassified surfaces	3	Unclassified surfaces	2	Unclassified surfaces
3	5	Industrial / commercial	5	Industrial / commercial	5	Non-residential urban (nrurb)	3	Non-residential urban (nrurb)
10	8	Institutions	7	Open urban (opurb)				
11	9	Urban green space						
4-6	10	Transport (trans)	8	Other urban (oturb)				
7-9	13	Waste, extraction and construction sites (wecs)						
12-17, 19	4	Arable and permanent crops (arab)	4	Arable and permanent crops (arab)	4	Cultivated (cul)	4	Rural (rur)
18	6	Pasture	6	Pasture				
20	11	Complex cultivation patterns	10	Other cultivated (otcul)				
21	12	Land principally occupied by agriculture						
23	7	Broad-leaved woodland	9	Woodland (wood)	6	Uncultivated (ucul)		
22, 24, 25	14	Other woodland (otfor)						
26-29	15	Unimproved grass and scrub (unimp)	11	Unimproved grass and scrub (unimp)				
30-38	16	Other uncultivated (otunc)	12	Other uncultivated (otunc)				

Notes: For each strategy, the numbers represent the order in which groups were entered into the regression analysis.

Upon establishing the base model, the Pearson’s correlation between the residual and the excluded variables was calculated. All land class-light variables that were positively, and significantly, correlated with the residual were grouped and reoffered to the model using the ‘stepwise’ regression method. Additional variables were only accepted into the final model if they did not cause any of the base model variables to become non-significant or negative, and if the first and last criteria (above) held true.

By applying these criteria, the final models were not necessarily those that gave the best prediction of population at the NUTS 5 level, but instead were constrained to avoid negative population predictions and counter-intuitive relationships (e.g. negative associations with light intensity) and to provide estimates of ‘background’ population densities - i.e. across areas covered by land cover types that did not explicitly enter the model. The resulting models took the general form shown in Equation 4.4.

$$\text{Equation 4.4} \quad PE_s = PB + \sum_{j=1}^J [(l \times Al) + (u \times Au) + (i \times Li)]_{js}$$

Where: PE_s is the estimated population in NUTS region s ; PB is the ‘background’ population (as defined by the constant); Al_{js} is the number of lit points in land cover group j in NUTS region s ; Au_{js} is the number of unlit points in land cover group j in NUTS region s ; Li_{js} is the total light intensity from land cover group j in NUTS region s ; and l , u and i are the regression coefficients (for lit, unlit and light intensity, respectively).

Regression models were developed at the national level as well as for the whole EU (excluding eastern Germany which was based on NUTS 3 level population data). In an attempt to reduce computational demands, several adjacent countries considered to have similar landscape characteristics were grouped for modelling. The combined countries included Ireland and Northern Ireland, Belgium and Luxembourg, and Portugal and Spain.

To evaluate the potential added value of using the light intensity variables, models were also developed for each strategy and geographic area on the basis of land cover area alone. These are presented in the results as the ‘no-light’ models.

The results of the various models were compared on the basis of the regression statistics (R^2 , SEE and SEE/mean). These can be interpreted as measures of ‘stress’ which indicate the degree of mismatch between the predicted and actual population. Because the population modelling approach used the pycnophylactic principle to rescale the predicted population to that of the true NUTS 5 total, models which maximise R^2 and minimise SEE signify that the dasymetric redistribution of the population was subject to minimal stress. In other words, areas with low stress are those within which the model works well.

Calculating point population

The models were then used to calculate the population for each 200 metre point by way of a mass-preserving approach. Essentially, the census population was redistributed to the points within a particular NUTS 5 region to ensure that the population estimates summed to those of the original NUTS regions. To facilitate this approach, each point was assigned a ‘background’ and land class-light specific contribution according to Equation 4.5. The ‘background’ population derives from the model constant and is distributed evenly across all points within a NUTS region. The land class-light contribution, based on whether the point is lit or unlit, derives from the lit area (Al), unlit area (Au) and light intensity (Li) value and coefficients for each land cover class.

Equation 4.5

$$PP_{ijs} = (PC_s - PB) \times \frac{[(l \times pAl) + (u \times pAu) + (i \times pLi)]_{ijs}}{\sum_{j=1}^J \sum_{i=1}^I [(l \times pAl) + (u \times pAu) + (i \times pLi)]_{ijs}} + \frac{PB}{(Al_s + Au_s)}$$

Where: PP_{ijs} is the population for point i belonging to land cover class j , in NUTS region s ; PC_s is the census population in NUTS region s ; pAl_{ijs} is the light status (1 = lit, 0 = unlit) for point i in land cover group j in NUTS region s ; pAu_{ijs} is the light status (1 = unlit, 0 = l) for point i in land cover group j in NUTS region s ; and pLi_{ijs} is the light intensity (integer if lit, 0 if unlit) for point i in land cover group j in NUTS region s .

The 200 metre point populations were calculated using the *pop_model.aml* and *lut_2_point.aml* presented in Appendix B1. The point populations were then summed to the 1km grid, and divided by land area, to derive a 1km² population density map across the study area.

4.3 Results

As was mentioned above, two sets of models were developed for each strategy and geographic area. The ‘no-light’ models were developed on the basis of land cover area alone, while the ‘light intensity models’ included land cover in association with the three light variables (area lit, area unlit and light intensity). The regression statistics for the no-light and light intensity models, both national and EU for each strategy, are presented in Tables 4.3 and 4.4, respectively.

4.3.1 No-light versus light intensity models

The statistics presented in Tables 4.3 and 4.4 show that stress associated with both the light and no-light models tended to increase for most countries as the models were simplified (i.e. from strategy 1 to strategy 4, as indicated by the decrease in R^2 and increase in SEE). This trend, however, was more pronounced in the no-light models, with clear increases in stress for Spain/Portugal, Finland, France, Italy and eastern Germany. The no-light models for Denmark and Belgium/Luxembourg also exhibited a lack of stability as the R^2 increased by approximately 0.2 between strategies 3 and 4. Apart from models for western Germany, Austria, Finland and the Netherlands, with R^2 exceeding 0.9 and small standard errors relative to the mean NUTS 5 population, the no-light models (Table 4.3) did not perform as well as the light models (Table 4.4). This was particularly evident for Spain/Portugal, Greece and France which all had poor strategy 1 models with R^2 values ranging from 0.40-0.56, compared to 0.81-0.83 for the associated light intensity models. Given the clear disadvantage of the no-light models, they were neither mapped nor validated and the remaining discussion relates only to the models in which the light variables were incorporated.

4.3.2 Light intensity models

Table 4.4 presents the stress statistics associated with the national and EU-level light intensity models. The R^2 was generally in the order of 0.8-0.9 and, unlike the no-light models, differences between the strategies were only slight. Variation in model stress between the countries was noted, with the Great Britain and Ireland/Northern Ireland models having the lowest R^2 values (-0.7) across all strategies. These particular models also had the smallest SEE, however, which may reflect the way administrative areas in these countries are designed to limit variations in population totals. The NUTS populations on which the models were built, therefore, have little variance as indicated by the small standard deviations in Table 4.1.

The strategy 3 models, selected to demonstrate general trends in the model predictors, are presented in Table 4.5 (see Appendix B2 for all strategies). The structure of these models was broadly similar across the geographic areas. In most cases, the continuous and discontinuous urban classes were represented by one or both of the area lit and light intensity variables. Unlit areas of these land classes also appeared in many of the models. The continuous and discontinuous land classes

tended to dominate the models, and accounted for the majority of the total R^2 in each model (i.e. between 88 and 99%). The non-residential urban class was also an important predictor in the EU model, as well as in some of the national models. The role of the rural classes was more variable across the countries, however, and generally accounted for only a small percent of the total model R^2 . The model coefficients were also by and large logical for all strategies, and followed the expected variations in population density. The coefficients for continuous urban areas, for example, tended to be higher than those for the discontinuous and non-residential urban classes, and much greater compared to the rural classes. Similarly, the coefficients for lit areas tended to be greater than those for unlit areas. The constant, which represented the 'background' population, varied from country to country, and tended to be higher in models having limited rural classes.

Table 4.3 Regression statistics for land cover area (no-light) models

Area	Strategy 1			Strategy 2			Strategy 3			Strategy 4			Pop threshold
	R ²	SSE	SSE/mean	R ²	SSE	SSE/mean	R ²	SSE	SSE/mean	R ²	SSE	SSE/mean	
AT	0.95	1,954	0.59	0.92	2,328	0.70	0.91	2,502	0.76	0.87	2,970	0.90	1,500,000
BE-LU	0.75	10,160	0.69	0.73	10,387	0.71	0.39	15,703	1.07	0.61	12,633	0.86	400,000
east DE ¹	0.77	30,410	0.20	0.74	32,227	0.21	0.73	32,913	0.21	0.45	46,974	0.31	3,000,000
west DE ²	0.92	7,461	1.00	0.90	8,186	1.10	0.86	9,790	1.31	0.82	10,872	1.46	900,000
DK	0.81	8,210	0.44	0.81	8,303	0.45	0.64	11,274	0.61	0.84	7,607	0.41	260,000
ES-PT	0.40	13,954	3.62	0.25	15,529	4.03	0.23	15,755	4.09	0.05	17,558	4.56	1,600,000
FI	0.96	6,258	0.56	0.95	6,555	0.59	0.97	5,084	0.45	0.60	19,208	1.72	
FR	0.56	4,486	2.90	0.55	4,556	2.94	0.43	5,122	3.31	0.01	6,729	4.35	400,000
GB	0.71	2,095	0.40	0.71	2,107	0.40	0.71	2,107	0.40	0.69	2,156	0.41	
GR	0.42	11,977	1.21	0.42	12,020	1.21	0.38	12,369	1.24	0.37	12,508	1.26	300,000
IE-NI	0.72	822	0.65	0.72	824	0.65	0.72	827	0.65	0.71	832	0.66	
IT	0.71	10,421	1.49	0.69	10,837	1.55	0.56	12,828	1.83	0.45	14,367	2.05	900,000
NL	0.92	7,777	0.26	0.91	8,352	0.28	0.90	8,478	0.29	0.89	9,012	0.31	200,000
EU ³	0.77	5,038	1.27	0.77	5,057	1.28	0.76	5,165	1.30	0.72	5,560	1.40	300,000

Notes: (1) Based on 1997 NUTS 3 census data and boundaries; (2) Based on 2003 census data; (3) Model excludes east DE

Table 4.4 Regression statistics for light intensity models

Area	Strategy 1			Strategy 2			Strategy 3			Strategy 4			Pop threshold
	R ²	SSE	SSE/mean	R ²	SSE	SSE/mean	R ²	SSE	SSE/mean	R ²	SSE	SSE/mean	
AT	0.90	2,581	0.78	0.90	2,702	0.82	0.92	2,362	0.71	0.89	2,740	0.83	1,500,000
BE-LU	0.87	7,255	0.50	0.86	7,451	0.51	0.81	8,708	0.59	0.83	8,267	0.56	400,000
east DE ¹	0.84	25,414	0.17	0.83	26,200	0.17	0.83	26,002	0.17	0.82	26,712	0.17	3,000,000
west DE ²	0.97	4,246	0.57	0.97	4,670	0.63	0.97	4,611	0.62	0.96	5,160	0.69	900,000
DK	0.95	4,071	0.22	0.95	4,097	0.22	0.95	4,367	0.24	0.91	5,606	0.30	260,000
ES-PT	0.81	7,886	2.05	0.80	8,117	2.11	0.80	8,140	2.11	0.74	9,252	2.40	1,600,000
FI	0.97	5,459	0.49	0.97	5,555	0.50	0.97	5,367	0.48	0.94	7,146	0.64	
FR	0.83	2,794	1.80	0.82	2,848	1.84	0.82	2,862	1.85	0.80	3,060	1.98	400,000
GB	0.73	2,033	0.39	0.73	2,032	0.39	0.73	2,037	0.39	0.72	2,050	0.39	
GR	0.82	6,644	0.67	0.81	6,835	0.69	0.81	6,889	0.69	0.70	8,660	0.87	300,000
IE-NI	0.73	802	0.63	0.73	801	0.63	0.73	813	0.64	0.72	815	0.65	
IT	0.86	7,386	1.06	0.85	7,449	1.06	0.85	7,611	1.09	0.81	8,387	1.20	900,000
NL	0.92	7,946	0.27	0.92	7,576	0.26	0.92	7,632	0.26	0.92	7,968	0.27	200,000
EU ³	0.82	4,518	1.14	0.81	4,551	1.15	0.81	4,634	1.17	0.77	5,063	1.28	300,000

Notes: (1) Based on 1997 NUTS 3 census data and boundaries; (2) Based on 2003 census data; (3) Model excludes east DE

Table 4.5 Strategy 3 models (with light intensity)

Land Cover Class		EU	AT	BELU	west DE	east DE	DK	ESPT	FI	FR	GB	GR	IENI	IT	NL
	Adj. R ²	0.81	0.92	0.81	0.97	0.83	0.95	0.80	0.97	0.82	0.73	0.81	0.73	0.85	0.92
Constant	Std Error	4634	2362	8708	4611	26002	4367	8140	5367	2862	2037	6889	813	7611	7632
1: Continuous urban		60.6	374.6	1,804.6	27.8	44,722.7	2,819.8	58.3	2,989.4	32.1	1,602.3	2,017.5	432.9	84.6	1,036.3
	lc1_al	191.9	662.1	616.7	927.1	694.3	579.4		385.9	402.0	167.4	667.1	114.0	196.3	195.6
	lc1_li	145.2	1,404.3		910.7			253.0			161.1	20.5			175.9
2: Discontinuous urban	lc1_li	0.023						0.047		0.013	0.003		0.002	0.042	0.006
	lc2_al	104.9	148.7		106.5	47.2	50.7	114.3			115.2	64.5	100.7	124.6	48.4
	lc2_au	89.3	64.2	25.6		33.1	29.5	88.7		67.4	70.8	99.4	100.2	112.8	107.1
3-11: Non-residential urban	lc2_li	0.007		0.033	0.027	0.029	0.011		0.002	0.028	0.009	0.012	0.006	0.016	
	nrurb_al	91.2			66.0		56.2	236.1		116.4				65.3	62.2
	nrurb_au	22.3		140.0	69.1	15.5			52.5				12.5	78.9	
12-21: Cultivated	nrurb_li	0.002	0.017	0.011	0.018	0.046							0.003	0.009	0.020
	cul_al	0.7		3.2					1.5	0.7		8.6			
	cul_au	0.2	1.5		3.4		0.4	0.3		0.5			0.3		
22-38: Uncultivated	cul_li	0.007			0.014			0.012		0.004		0.013	0.001	0.012	
	ucul_al						7.0	2.7	0.5		1.1			1.6	
	ucul_au	0.1	0.3		0.8										
49: Unclassified surfaces	ucul_li			0.031	0.071				0.006	0.003	0.005		0.014	0.005	
	lc49_al	20.1								49.1		27.8	59.9		
	lc49_au	1.0									12.3			51.2	
	lc49_li							0.528							

4.3.3 Model selection and mapping

The stress statistics associated with the EU models were assessed by examining the R^2 and SEE in relation to the median of the national models (Table 4.4). Note that eastern Germany was excluded from this comparison as it was based on NUTS 3 totals. The EU models across the four strategies showed more stress than the respective national models ($R^2 = 0.77-0.82$ compared to $0.82-0.87$), but the SEE was slightly smaller than the national level median (SEE = 4,518-5,063 compared to 4,852-6,376 people).

Both the national and EU models (for strategy 1 and 2) were applied to the land class-light points using Equation 4.5 to calculate the estimated 200 metre point population. The point populations were then aggregated to the 1km level for mapping. When mapped together across the EU, the national models resulted in apparent discontinuities at some national borders. This is depicted in Figure 4.4 which shows the national models for France, Belgium, the Netherlands, Luxembourg and western Germany. Although legitimate difference between countries in urban structure or land use might account for these step changes, it is also possible that they are the consequence of differential application of the land cover classification across countries, during development of the CORINE land cover map. In any event, the discontinuities were less apparent in the EU model map. As indicated in Figure 4.5, there was little apparent difference between the 1km predictions for the EU strategy 1 and strategy 2 models. Based on the SEE presented in Table 4.4, however, the stress associated with the strategy 1 model was slightly reduced compared to the strategy 2 model (additionally, many of the strategy 1 national models outperformed, although only marginally, the respective strategy 2 models). On this basis, the EU strategy 1 model was thus selected as the preferred model for mapping population density (Figure 4.6).

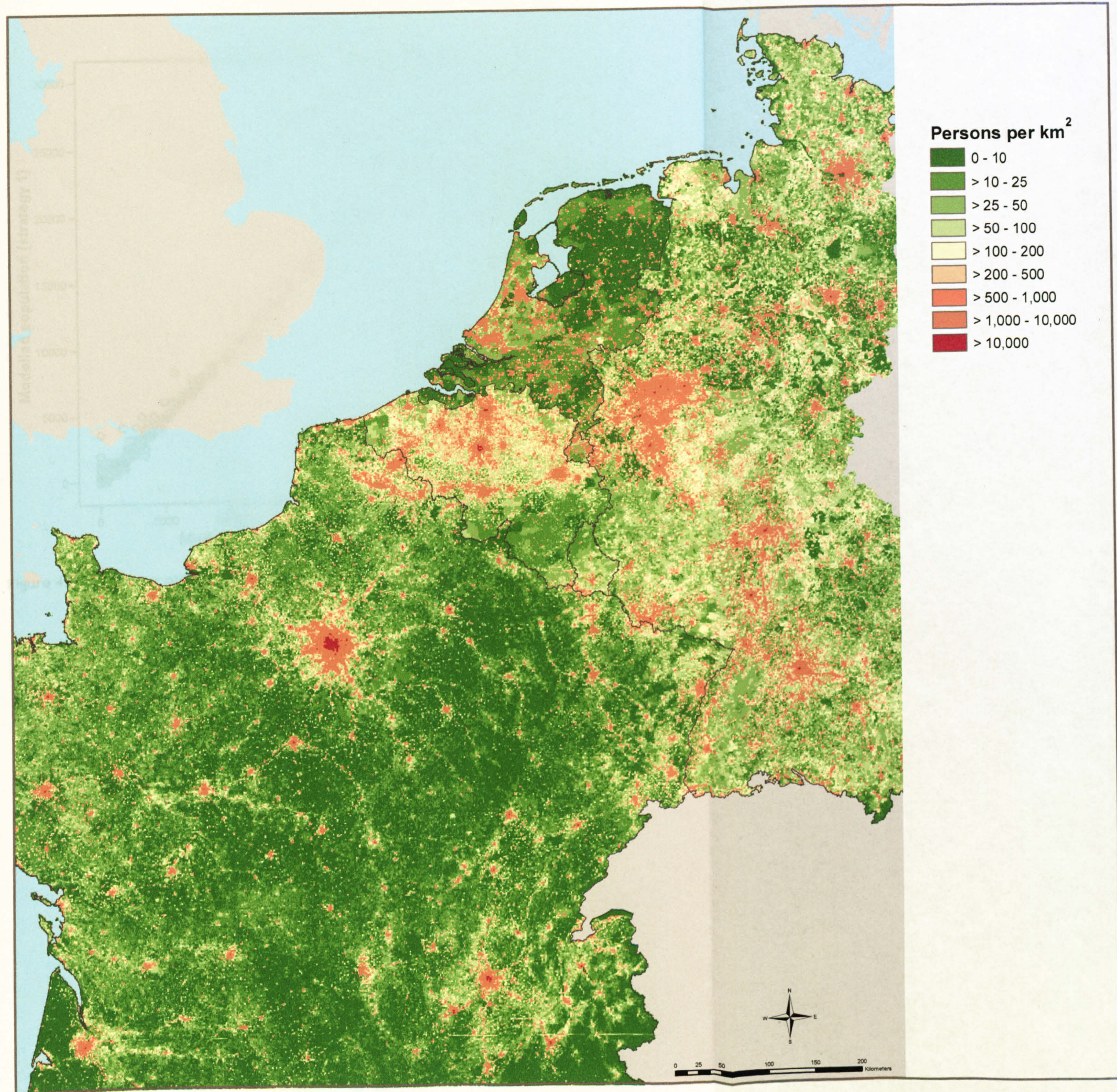


Figure 4.4
Population density: national models (strategy 2)
Showing Belgium, the Netherlands, France,
Luxembourg and western Germany

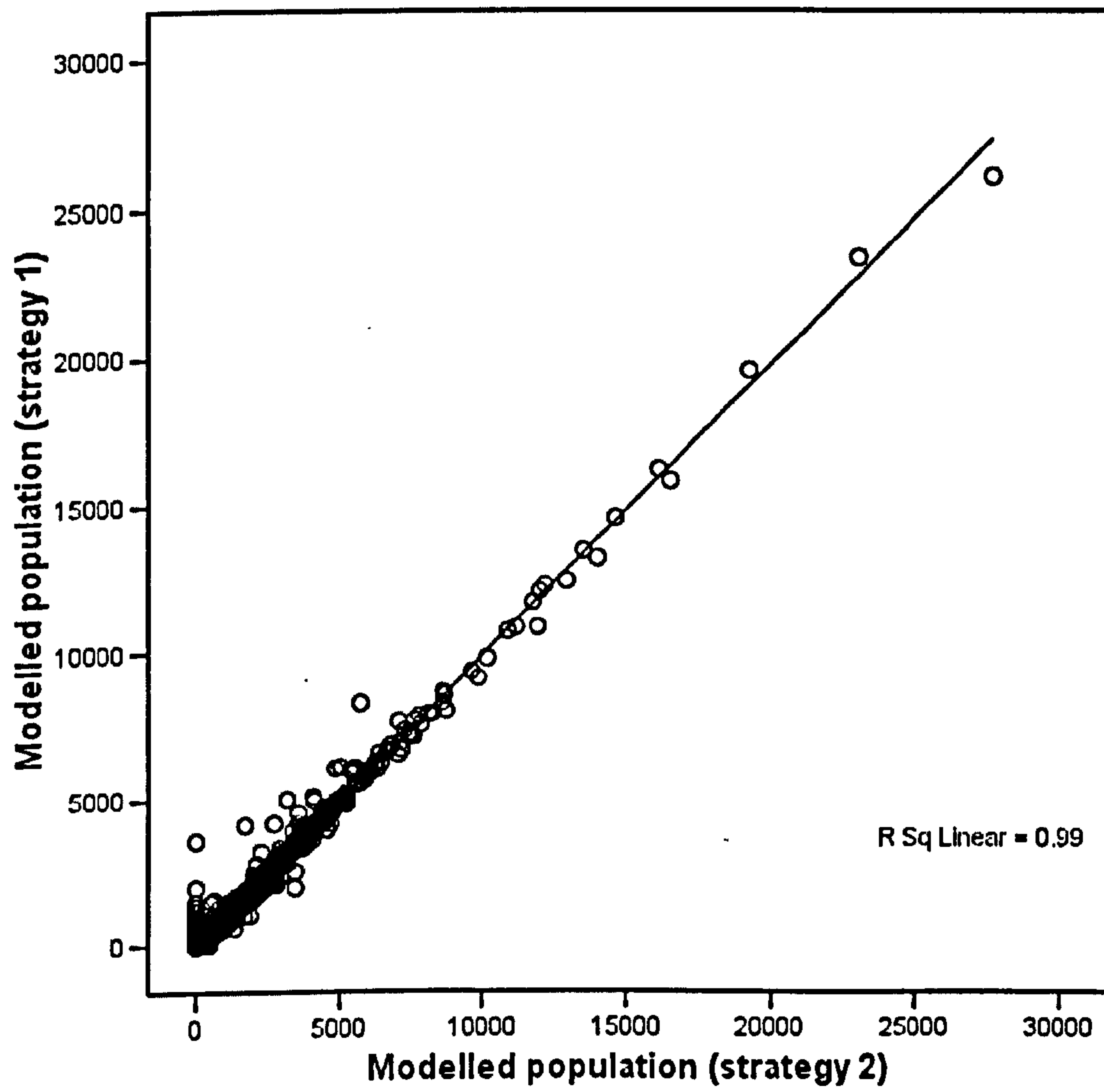


Figure 4.5 **Scatter plot EU strategy 1 vs. strategy 2**
 (Sample of 1% of grid squares)

4.4 Validation

Although the modelled population was conditioned to sum to unity at the NUTS 5 level, potential modelling errors may have occurred in the resulting distribution of population within NUTS 5 regions. Validation was, therefore, conducted to assess modelling error within these regions and determine if, and at what level, the modelled population was robust at smaller spatial scales.

Due to the lack of independent, small-area population counts, validation was only possible in Great Britain using the detailed population counts provided by the ONS. For 1991, data are available for enumeration districts (ED, n = 147,596) while for 2001 they are available both for output areas (OA, n = 218,038) and postcodes (headcount data, n = 1.38 million). EDs and OAs are both smaller than NUTS 5 regions; EDs, for example, nest into wards which are equivalent to NUTS 5 regions.

Four sets of reference population data, against which the modelled 200 metre point populations were compared, were compiled for the validation studies:

- ED1991 - enumeration district totals, derived from the 1991 census
- KM1991 - 1km totals for 1991, obtained by postcode weighting of 1991 ED totals to a 1km grid
- OA2001 - output area totals, derived from the 2001 census
- KM2001 - 1km totals for 2001, obtained by summing the 2001 headcount data to a 1km grid

The modelled 200 metre point populations for Great Britain, representing the centroid of each pixel, were used in the validation so that point-in-polygon overlay, instead of the computationally more intensive polygon overlay, could be conducted. To facilitate these overlay operations, the points were reprojected into British National Grid to match the boundaries of the reference data. Modelled population derived from both the EU and GB national models were used in the validation studies.

Ms. Daniela Fecht (SAHSU, Imperial College) is acknowledged for her substantial contribution to pre-processing of the population data and carrying out this validation.

4.4.1 ED1991 and OA2001

The ED1991 and OA2001 validation studies involved assessing how well the modelled population could predict 1991 ED and 2001 OA population totals. For this validation, the 200 metre points were intersected with the respective census boundaries and summed to the ED/OA level using the ARC STATISTICS command. The statistics file was then imported into SPSS where the correlation between actual and modelled population was determined by linear regression.

It should be noted that some urban units (i.e. approximately 15%) were smaller than the 200 metre cell size, meaning that some EDs did not contain any of the 200 metre points. All EDs with ONS population counts, however, regardless of the presence/absence of a modelled 200 metre population point, were included in the validation.

4.4.2 KM1991 and KM2001

Postcode locations were obtained from the 1991 All Fields Postcode Directory maintained by the Post Office. These data were subjected to extensive checking and revision by comparison with later postcode files prior to use (work conducted by Dr. Kees de Hoogh at SAHSU, Imperial College). Intersection between the corrected postcode locations and EDs was done in ArcGIS.

The postcodes and modelled 200 metre point populations were then intersected with the 1km British National Grid. The 1km totals for the 1991 reference populations were derived from the ED totals by postcode-weighting, using Equation 4.2. For 2001, the 1km population counts were obtained by summing the postcode headcount data to the target map unit, based on an intersection of postcode locations from Code-Point 2001. The 1km totals (KM1991 and KM2001) were then imported into SPSS where linear regression was again used to assess the correlation between the modelled and reference populations.

4.4.3 Model comparison

The results from the validation studies are presented in Table 4.6. The Pearson correlation coefficient (r) and root mean square error (RMSE) were used as measures of performance. Across all validation studies and strategies, the GB national models

performed marginally better than the EU models. The GB models may have performed better as they were built on the land cover classes specific to Great Britain, whereas the EU models were optimised on land classes found across the whole study area.

At the 1km scale, the KM1991 and KM2001 validation showed a strong correlation for all strategies, with r ranging from 0.93-0.96 and RMSE in the order of 270 people. For ED1991 the correlations were much weaker (r = 0.33-0.40), and at the finer level of output area (OA2001) the performance was further weakened with r ranging from 0.23-0.27 and RMSE between 275 and 400.

Table 4.6 Validation results for Great Britain

Validation data set		Reference population		Strategy 1		Strategy 2		Strategy 3		Strategy 4	
File	N	Mean	SD	r	RMSE	r	RMSE	r	RMSE	r	RMSE
GB models											
ED1991	147,596	370	191	.391	346	.391	345	.394	343	.397	342
KM1991	238,790	228	835	.958	239	.958	238	.959	239	.959	239
OA2001	218,034	239	134	.264	276	.264	277	.265	275	.266	274
KM2001	263,101	218	838	.951	265	.951	264	.951	263	.951	263
EU models											
ED1991	147,596	370	191	.334	412	.334	411	.343	401	.353	397
KM1991	238,790	228	835	.938	290	.938	281	.942	276	.944	240
OA2001	218,034	239	134	.234	320	.233	321	.235	316	.240	312
KM2001	263,101	218	838	.931	307	.931	306	.935	298	.937	293

The distributions of the residuals were examined in an attempt to identify a pattern in the poor correlations in the ED1991 (range = -1,477 to 9,275; mean = 1.09; SD = 411) and OA2001 (range = -2,726 to 9,278; mean = 12.57; SD = 320) validations. Figure 4.7 shows the residual population for the EU strategy 1 model mapped across Great Britain for both validation studies. As can be seen in these maps, the residuals for much of Great Britain (86% of the EDs and 88% of the OAs) are within ± 1 standard deviation unit of the mean residual number of people. Roughly 9% of the larger EDs and OAs in Scotland, however, show over predictions that are +2 standard deviations or greater above the mean. Although under predictions (e.g. -2 to <-2.5 standard deviation) are difficult to see in Figure 4.7, approximately 5% of the EDs and 3% of the OAs are under predicted in terms of population. The under predicted areas tend to occur in small administrative units which are on average 0.3 km² compared to 1.6 km² and 1.1 km² for all other EDs and OAs, respectively.

4.5 Discussion

A quick comparison between Figure 4.1, a choropleth map of the 1991 population density at the NUTS 5 level, and the 1km modelled population density map (Figure 4.6) highlights the importance of small-area population density measures, especially if the map is to be used for assessing exposure. While simple methods, such as area-weighting, may improve the disaggregation of census totals to smaller target units, it is known that populations are not evenly distributed across administrative units. More intelligent methods using data to reflect lower level variations in population, therefore, are crucial in achieving reliable measures of small-area population density.

The GIS-based modelling strategy described here involved disaggregating NUTS 5 totals to the 200 metre level on the basis of land cover data. While the methods employed here were broadly similar to those used by Yuan et al. (1997), the criteria set out in the supervised step-wise regression method (to ensure only predictors with significant, positive coefficients were accepted into the models) eliminated the need to rescale the coefficients prior to using them as weights. A further improvement on their method involved the inclusion of light intensity as an additional ancillary data set to strengthen the correlations between land cover and population.

Including light intensity in the models (Table 4.4) allowed for local differences in population density, even when the land cover classes were grouped in strategies 3 and 4. In the absence of light intensity (Table 4.3), however, the more detailed land cover classifications in strategy 1 were needed to provide discrimination of variations in population density. Generally, incorporating light intensity improved the models for all countries. These improvements were more noticeable in some of the southern countries, such as Spain/Portugal, Greece and France, indicating that the land cover classification criteria or the relationship between land cover and night-time light intensity might differ across Europe.

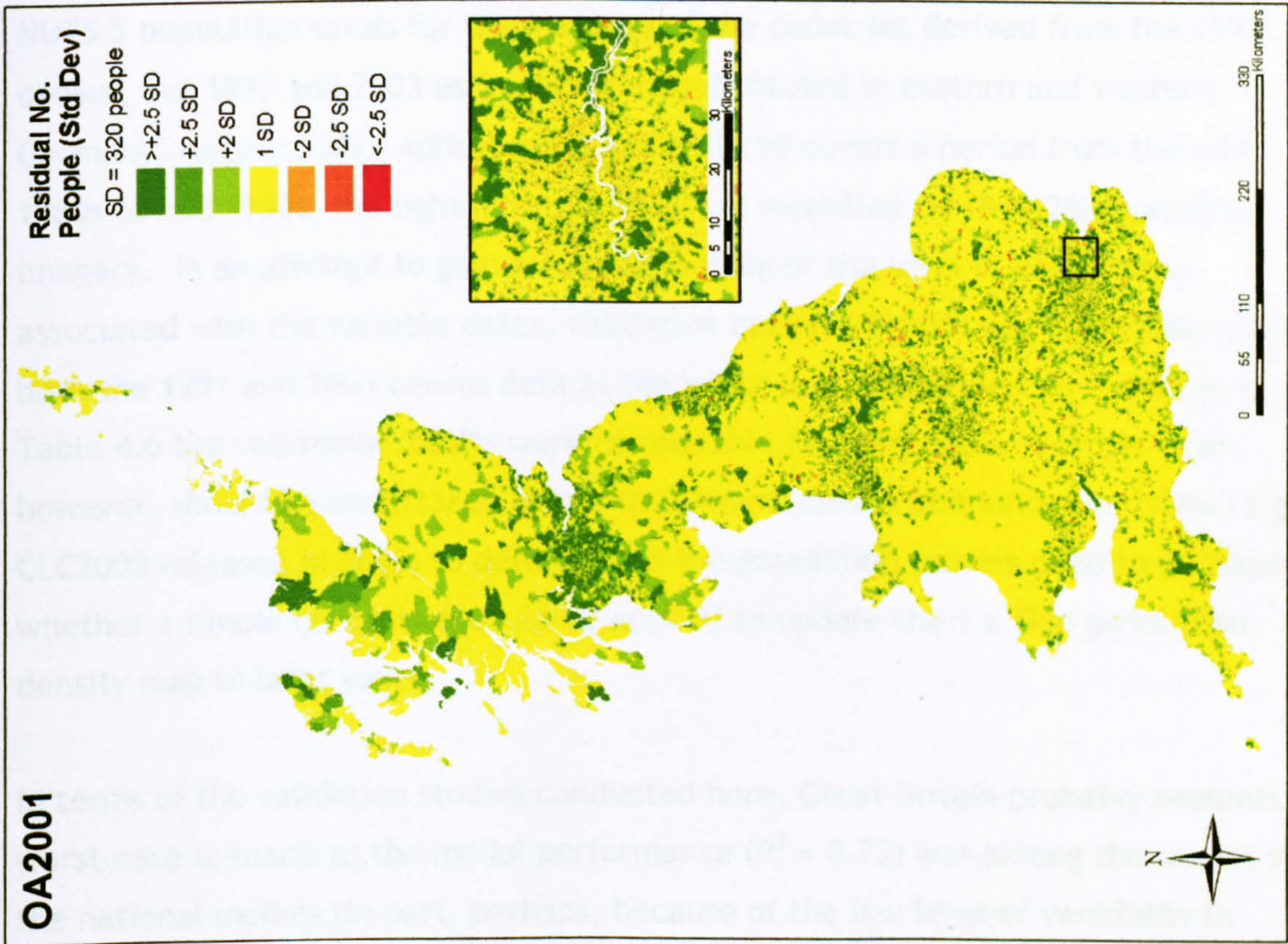
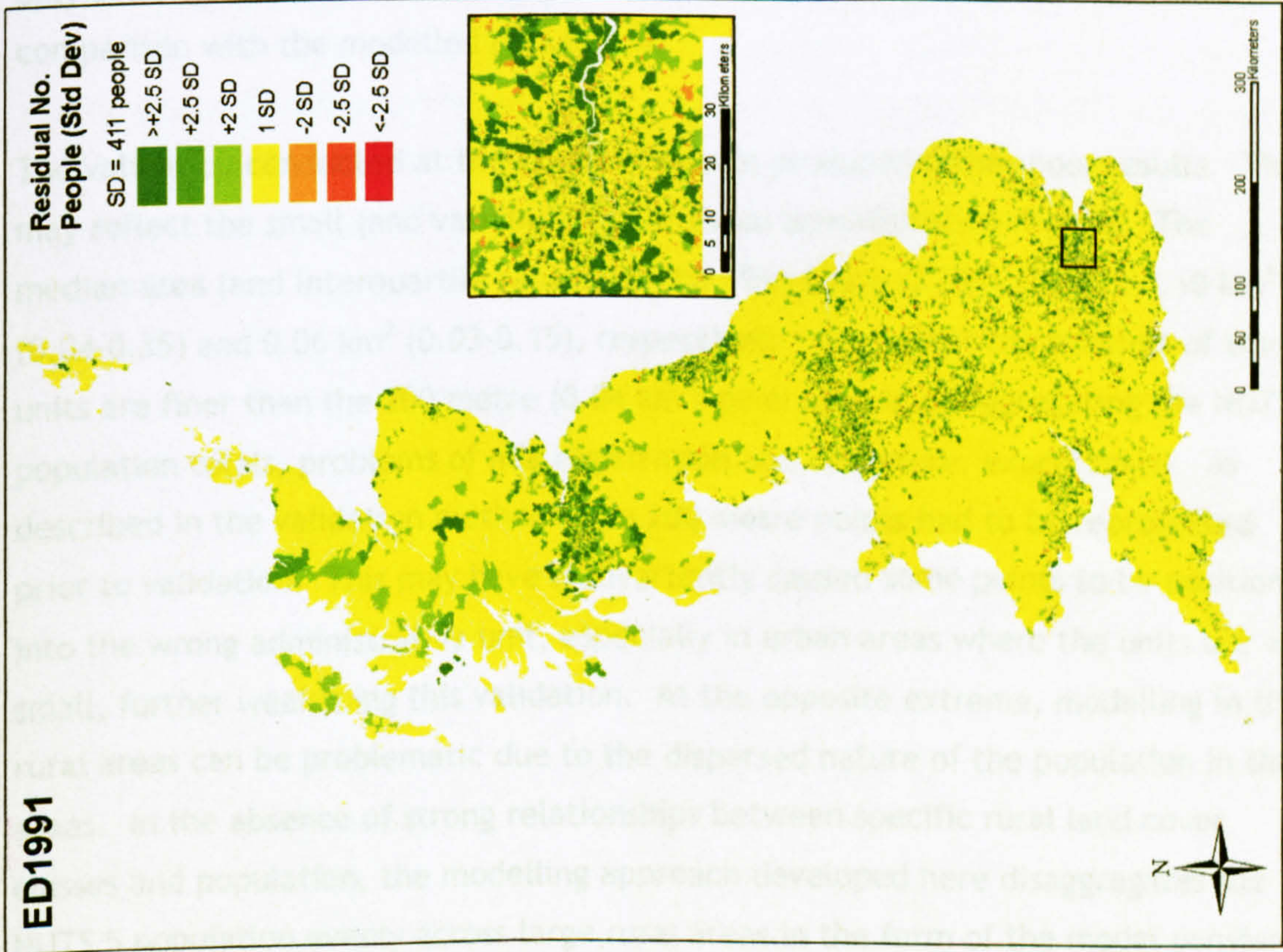


Figure 4.7 Residual population in Great Britain (EU strategy 1 model)

The data sets used as a basis for population modelling span more than ten years. The NUTS 5 population totals for the majority of the countries derived from the 1991 census, but 1997 and 2003 estimates were substituted in eastern and western Germany, respectively. Additionally, while CLC90 covers a period from the mid 1980s to mid 1990s, the light intensity map was modelled on 1999-2000 satellite imagery. In an attempt to gain an understanding of the level of uncertainty associated with the variable dates, validation in Great Britain was undertaken using both the 1991 and 2001 census data as the reference population. As indicated in Table 4.6 the validation results were comparable for both years. Further work, however, should be undertaken using more recent census and land cover data (e.g. CLC2000 released in 2005) to determine if the population models need to be rerun or whether a simple calibration could be applied to update the 1 x 1km population density map to later years.

In terms of the validation studies conducted here, Great Britain probably presents a worst-case scenario as the model performance ($R^2 = 0.73$) was among the lowest of the national models (in part, perhaps, because of the low level of variability in population numbers between NUTS 5 areas, compared to other countries). It was the only country, however, for which independent data were available to facilitate comparison with the modelled population.

The validation conducted at the ED and OA level produced rather poor results. This may reflect the small (and variable) size of these administrative regions. The median area (and interquartile range) for the 1991 EDs and 2001 OAs are 0.10 km² (0.04-0.35) and 0.06 km² (0.03-0.15), respectively. Given that the smallest of these units are finer than the 200 metre (0.04 km²) pixel used in disaggregating the NUTS 5 population totals, problems of mis-registration are, therefore, exacerbated. As described in the validation methods, the 200 metre points had to be reprojected prior to validation. This may have inadvertently caused some points to be positioned into the wrong administrative unit, especially in urban areas where the units are very small, further weakening this validation. At the opposite extreme, modelling in the rural areas can be problematic due to the dispersed nature of the population in these areas. In the absence of strong relationships between specific rural land cover classes and population, the modelling approach developed here disaggregates the NUTS 5 population evenly across large rural areas in the form of the model constant (i.e. 'background' population).

Despite the poor result at the finer scale, the validation studies indicated that the EU and GB models provided a good estimate of population density in Great Britain at the 1km target grid unit. The 1 x 1km EU-wide population density map can, therefore, be considered broadly reliable, though validation in additional countries would strengthen this conclusion.

4.6 Conclusions

A hybrid model incorporating dasymetric mapping and regression analysis was successfully developed and applied to disaggregate NUTS 5 census data, on the basis of land cover and light intensity data, to produce a small-area estimate of population across the EU. Regression models fitted at the national level outperformed those fitted at the EU level. The EU models, however, produced a more consistent high resolution population density map with less apparent step changes at national borders. While the independent validation in Great Britain showed poor correlations between the modelled and actual population at the enumeration district and output area level, validation at the 1km grid achieved very good agreement at 88%. Given that the 1 x 1km target map unit was the desired resolution for the population map needed for the EU-wide exposure assessment, this result was considered more than acceptable. Further research in this area of population modelling should include additional validation of the model in other countries as land cover and light intensity data become available. Additionally, the final population density map should be regenerated on the basis of recently available, updated versions of the census and land cover data.

5 Affinity Zone Stratification

This chapter aims to explore the utility of the air pollution monitoring network in the EU as a basis for pollution modelling. It does so using an approach called affinity zone stratification (AZS) which involves multivariate statistical techniques to define areas (i.e. affinity zones) likely to have similar air pollution levels based on their underlying physiographic features and land use characteristics.

5.1 Introduction

Most epidemiological studies derive measures of exposure to air pollution from the monitoring network - for example by taking the average concentration from the nearest available relevant sites. Such a simplistic method, however, takes no account of the spatial variability of the pollution surface. Variability in the surface is expected, if only because the sites theoretically represent different environments such as the city centre, suburbs and rural areas. This highlights the importance of recognising that data from air pollution monitoring sites can only be considered to represent a limited, and often local, area. The degree of representativeness in this area is likely to vary depending not only on proximity to the monitoring site, but also the degree of similarity in the environment (e.g. topography, emission sources, building characteristics and dispersion potential, meteorology). McGregor (1996) defined these areas as 'affinity zones' - i.e. an area within which air quality displays consistent behaviour over space, and thus in which the associated monitoring sites can be considered representative. The importance of defining affinity zones is twofold. First, they provide a basis for pollution modelling by showing to what areas data from any monitoring site (or set of sites) may legitimately be extrapolated; secondly, they can help to identify types of areas in which monitoring is currently inadequate, and might usefully be improved.

In order to explore affinity zones in relation to the AIRBASE-e monitoring network described in Chapter 3, methods of affinity zone stratification (AZS) were developed and applied. This approach is grounded in theoretical and practical studies in the area of soil (Webster & Burrough 1974) and ecological mapping (Bernert et al. 2003; Briggs & France 1983; Bunce et al. 1996a; Carter 1997; Fairbanks & Benn 2000; Harding & Winterbourn 1997). AZS techniques also form the basis of the ITE (Institute of Terrestrial Ecology) Land Classification (Bunce et al. 1996a; Bunce et al.

1996b; Bunce et al. 1996c) used in the UK Countryside Survey (Barr et al. 1993; Haines-Young et al. 2000). These studies all have one thing in common: the need to simplify complex geographical phenomena into distinct units using measurable environmental variables.

As discussed in Webster & Burrough (1974), the difficulty in soil classification is that membership in a particular class is generally based on the observations of several soil characteristics. In their research, therefore, they explored multiple discriminant analysis to distinguish soil classes. In doing so they calculated discriminant functions from data for pre-selected soil profiles to develop a classification. The discriminant function was then applied to unclassified soil profiles to allocate them to the appropriate class. Their results showed that allocation by discriminant functions was consistently better than allocation by the soil surveyor alone, and that the final groups were more compact and better separated from one another.

Similar to soil classification, the stratification of landscapes into management areas is complex because membership in a particular class is usually based on the observations of several important environmental characteristics. Multivariate statistical techniques were explored by Carter (1997), for example, in developing a soil fertility management strategy for Western Kenya. To stratify the landscape appropriately they had simultaneously to consider the population densities and growth rates in the region, large scale changes in land use, and the natural resource base and its variability, as well as migration effects on the labour force in maintaining the soil fertility. The numerous variables were first entered into principal components analysis (PCA) to reduce the data set. The 5km cells were then stratified by applying cluster analysis (CA) to the principal components. As a final step, discriminant analysis was used to verify the cluster membership and to identify the original variables most strongly associated with each cluster. In another study, Harding & Winterbourn (1997) employed cluster analysis to define ecoregions on the basis of climatic regions, rainfall, relief, vegetation, soils and geology for environmental management in the South Island of New Zealand. A similar approach involving cluster analysis followed by discriminant analysis was applied by Briggs & France (1983) to classify landscapes and habitats for regional environmental planning for South Yorkshire, UK. In this study, cluster analysis was performed on a sample of the grid squares, and then discriminant analysis was used to allocate unassigned grid

squares to one of the clusters. Discriminant analysis achieved a correct classification in 81% and 83% of the cases in the landscape and habitat analysis, respectively.

The UK Countryside Survey is the most comprehensive survey of land use and change, landscape features, habitats and species in the UK. As detailed surveys of the environmental resources over large areas are impracticable, surveys are often conducted on representative test areas to devise a classification scheme. On the basis of easily measured variables, unsampled locations can then be allocated to one of the established classes. This is the approach taken in the ITE classification (Bunce et al. 1996a; Bunce et al. 1996b; Bunce et al. 1996c) upon which the UK Countryside Survey is based (Barr et al. 1993; Haines-Young et al. 2000). The initial ITE classification was originally conducted in 1977 on a sample of 1,212-1km squares at the intersections of a 15 x 15km grid overlaid on Great Britain. Information on climate, topography, human geography and geology was recorded for each of the 1km squares. The data were then classified using indicator species analysis, a divisive polythetic classification technique, to produce a hierarchical stratification of 32 groups. The 32 land classes were used as strata to select sites for field survey of vegetation, soils and land cover across GB in order to assess the effectiveness of the classification. Correlation analyses were used to test the overall concordance between the environmental variables and various sets of survey data. The classification was subsequently extended to all 1km squares across Great Britain using a combination of logistic discrimination and discriminant functions.

Similar approaches have, as yet, rarely been used to study air pollution. The closest example is the work by McGregor (1996) who derived the concept of air pollution affinity zones. McGregor (1996) applied this concept to air pollution time-series data from Birmingham, UK to determine if zones of similar air quality existed in the urban area. The methods involved principal components analysis followed by cluster analysis. The data set, which comprised 571 daily SO₂ measurements from 17 monitoring stations, was reduced to four principal components using PCA. The relationship between the monitoring sites and the component loadings provided a measure of the extent to which a component represented the temporal behaviour of air pollution at a given location. The component loadings were then clustered in order to group the monitoring sites with similar behaviour of air pollution concentrations over time and space (i.e. affinity zones). Four SO₂ affinity zones - comprising light to heavy industry, low to medium density residential housing,

residential with open spaces, and industrial-vehicle areas associated with heavy iron and steel production - were identified.

Several additional studies have used similar stratification techniques to classify the monitoring sites on the basis of pollutant concentrations. Zelenka (1997) conducted a study to evaluate the relationship between meteorology and ambient concentrations of aerosol strong acidity (H^+) in Uniontown, Pennsylvania. Daily meteorological data were analysed for the factors which were shared by groups of days having similar meteorological conditions. This was achieved by first entering the meteorological data into PCA to minimise collinearity. A two-stage cluster analysis, involving average linkage followed by K-means CA, was then performed on the principal component scores to determine those days with similar meteorological conditions. This procedure identified seven groups of days representing the general meteorological patterns for the study period. Finally, forward stepwise regression analysis was used to identify the meteorological variables that most influenced the ambient H^+ concentrations in the overall data and for each of the clusters. Similar multivariate techniques were employed by Ho et al. (2002) to characterise volatile organic compounds, poly-aromatic hydrocarbons and carbonyl compounds at a roadside monitoring station in Hong Kong. They identified four main contributing sources to the concentration of toxic air pollutants at the monitoring station.

Cluster analysis was also used in a study by Lavecchia et al. (1996) to examine summer ozone patterns in the area surrounding Milan, Italy where air quality limits are frequently exceeded. Hourly ozone measurements from 18 monitoring sites, representing urban and rural landscapes, were hierarchically clustered to group the monitoring sites on the basis of concentration and temporal trends. The monitoring sites were clustered into four to five groups covering three main regions (south, intermediate and north). An increasing gradient in ozone concentration was found going from south to north, corresponding to geography and a decrease in the number of inhabitants, traffic density and factories. A final example is the work by Flemming et al. (2005) in which hierarchical CA was used to explore the classification of air pollution monitoring sites in Germany. The classification was based on hourly O_3 , NO_2 , SO_2 and PM_{10} /TSP concentration data from approximately 650 sites during the period 1995-2001. Clustering analysis revealed six O_3 and NO_2 and five SO_2 and PM_{10} regimes. These included the five relatively common regimes 'rural', 'urban', 'suburban', 'street' and 'urban street', plus a 'mountain' regime for O_3 and a

‘severely polluted street’ regime for NO₂. The agreement between the O₃ and NO₂ classifications was reasonably strong for most stations of the same regime. Agreement between O₃ and the other two pollutants, however, was not good given the large range of SO₂ and PM₁₀ concentrations within specific O₃ regimes.

As these examples suggest, there are two different approaches to AZS - one spatial and the other temporal. The first uses geographic variables to distinguish between and group the monitoring sites while the second uses time-series data (e.g. concentrations) to define monitoring sites with similar pollution profiles. These are referred to here as the area-based and site-based approaches, respectively. As was demonstrated above, most research has taken a site-based approach to evaluating the monitoring networks. The site-based air pollution studies, although limited, illustrate the importance of and need to better understand how well monitoring sites characterise the areas they are meant to represent. As illustrated in the research by Flemming et al. (2005), this involves determining whether or not sites classified as urban, for example, measure similar pollution concentrations. This has implication because the assumed representativeness of a monitoring site may differ depending on its classification. Observations in street canyons, for example, are assumed less spatially representative than those measured in rural background locations (Flemming et al. 2005).

The area-based approach takes a different view in that it asks ‘what is the extent or area to which the data from these monitoring sites may be meaningfully extrapolated?’ In the context of this study the specific need is to consider the geographical distribution of sites in relation to characterising long-term average concentrations: thus, an area-based approach is taken.

5.2 Methods

As was discussed in Section 2.1.3, pollution patterns vary according to factors including the distribution and activity of emission sources, chemical processes, and processes influencing dispersion. These variations can, perhaps, be characterised in terms of land cover, topography and meteorology - an assumption that, in full or in part, underlies the current air pollution models such as regression mapping. By classifying the study area in terms of these characteristics it may be possible to

identify areas that are similar in pollution. These areas can, thus, be used to define affinity zones, representing specific groups of monitoring sites.

The various statistical approaches applied in previous soils and ecological studies can be further developed as a means of analysing pollution monitoring networks, and as a basis for pollution mapping. This can be achieved through the affinity zone stratification (AZS) approach developed and presented in this chapter. An area-based approach is taken to explore the spatial patterns of air pollution and relationships with the underlying physiographic and landscape features. It involves first identifying and selecting a relevant subset of variables with which to characterise the pollution environment. The selected variables are then used to classify areas into different types or 'affinity zones'. The variation in monitored pollution concentrations within each of these affinity zones is then analysed. An overview of the approach is presented in Figure 5.1.

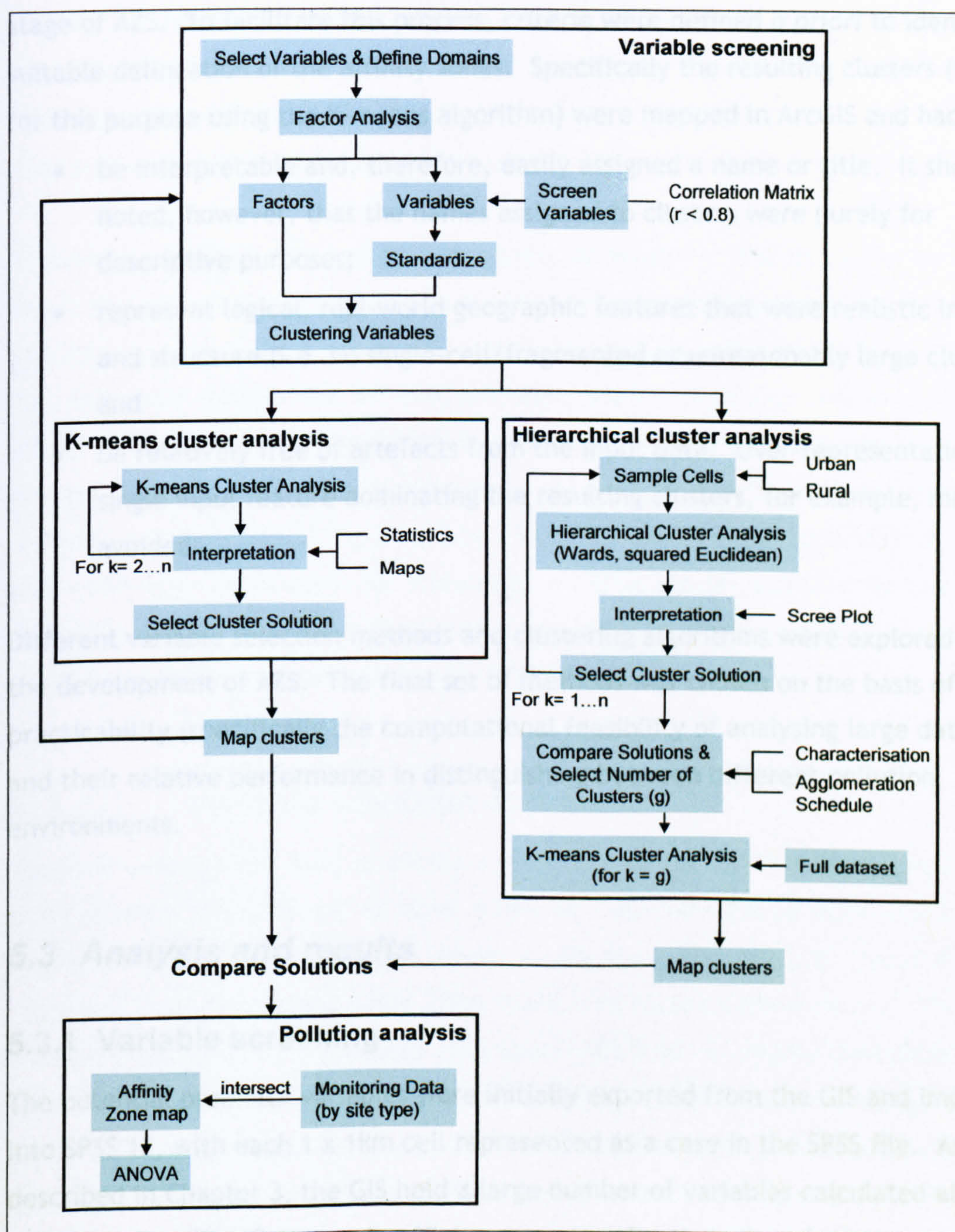


Figure 5.1 Affinity zone stratification approach

The effectiveness of the AZS process depends, amongst other things, on how well the variables selected for clustering represent the pollution environments under study. Given that a wide range of often highly correlated variables might be used for this purpose, some form of screening needs to be undertaken of the possible descriptors. The result of screening, however, cannot be evaluated until the selected variables are clustered and viewed as a map. For this reason, there exists a feedback loop (shown in Figure 5.1) between the cluster analysis stage and the variable selection

stage of AZS. To facilitate this process, criteria were defined *a priori* to identify a suitable delineation of the affinity zones. Specifically the resulting clusters (defined for this purpose using the K-means algorithm) were mapped in ArcGIS and had to:

- be interpretable and, therefore, easily assigned a name or title. It should be noted, however, that the names assigned to clusters were purely for descriptive purposes;
- represent logical, real-world geographic features that were realistic in size and structure (i.e. no single-cell/fragmented or unreasonably large clusters); and
- be relatively free of artefacts from the input data. Over-representation of a single input feature dominating the resulting clusters, for example, must be avoided.

Different variable selection methods and clustering algorithms were explored during the development of AZS. The final set of methods was chosen on the basis of both practicability (specifically the computational feasibility of analysing large data sets) and their relative performance in distinguishing between different pollution environments.

5.3 Analysis and results

5.3.1 Variable screening

The potential predictor variables were initially exported from the GIS and imported into SPSS 12, with each 1 x 1km cell represented as a case in the SPSS file. As described in Chapter 3, the GIS held a large number of variables calculated at various spatial scales. The first step in affinity zone stratification, therefore, was to identify the most appropriate set of variables from which to characterise the pollution environment.

Parsimonious selection is important both to limit the processing demands involved in cluster analysis, and to ensure that the resulting clusters are distinct and interpretable. It is also important to limit the degree of multicollinearity between the variables, in order to meet the requirements of CA. Two approaches may be used for variable selection: (1) variables can be selected on the basis of their assumed relevance to the specific application (in this case air pollution) and then

screened to remove individual variables exhibiting high levels of correlation (e.g. pair-wise Pearson correlation coefficient $r \geq 0.8$); or (2) statistical techniques (e.g. factor analysis) can be applied to identify 'natural' groupings of variables within the data set. Both approaches were explored separately and in combination here.

As mentioned above, the validity of the variable selection process cannot be confirmed until the results of the clustering process are mapped and examined. Variables and/or factors leading to interpretable, logical clusters for defining air pollution were sought. The variable selection, statistical and combined approaches for variable selection are described, in this order, in the subsequent sections. This process also involved a sensitivity analysis, conducted with the UK data, using different combinations of input variables to select the most appropriate cut-off value (i.e. correlation coefficient) for screening. A higher cut-off was used for the statistical approach as some correlation is desirable in identifying interrelated sets of variables in factor analysis (Hair et al. 1998).

5.3.1.1 Variable selection approach

Expert knowledge was used to decide *a priori* the subset of the variables and specific scale at which the different variables might best describe source distribution and dispersion processes important to characterising the pollutants (i.e. 1km and 5km represent local phenomena while 21km represents regional phenomena). The transport and land cover variables associated with built-up, highly-trafficked areas, for example, were included as important indicators of local source activity. The 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). Local agricultural activities involving the exposure of bare soil, such as tilling fields, may also contribute to PM_{10} ; thus, an additional variable to reflect such activity was included for PM_{10} only.

This analysis was undertaken on the full study area and included the variables presented in Figure 5.2. The pair-wise Pearson correlation coefficients (r) for these

35 variables were examined. Four pairs were found to be highly correlated ($r \geq 0.8$); thus, one from each pair was removed (indicated in blue in Figure 5.2). The remaining variables were standardised (mean = 0; standard deviation = 1) and entered into cluster analysis.

Input Variables			12 Clusters	
Major road5	WECS1	HDRES1	1	Not interpreted
Minor road5	AIRP5	Altitude	2	"
Major road1	AIRP1	TOPEX	3	"
Minor road1	PORT5	Sea dist	4	"
TBU21	PORT1	SM temp	5	"
AGR5	TRANS5	WM temp	6	"
AGR1	TRANS1	AM radiation	7	"
FOR5	IND5	AM wind speed	8	"
FOR1	IND1	A calm	9	"
URBGR5	LDRES5	AM vapour pressure	10	"
URBGR1	LDRES1	AS rain	11	"
WECS5	HDRES5		12	"

Figure 5.2 Example input for the variable selection method

Notes: (1) Variables shown in blue were excluded prior to CA; (2) AM = annual mean; AS = annual sum: A = annual; SM = summer mean; WM = winter mean; 1, 5 = 1 x 1km, etc. (3) see Table 3.3 for land cover abbreviations.

As can be seen in Figure 5.3, this method did not produce suitable clusters as some were apparently blocky. The blockiness can be attributed to the coarse resolution of the meteorological data which seemed to dominate the large rural clusters. Given this poor result, the clusters from this analysis were not interpreted. Instead the alternate statistical approach was explored, as described in the next section.

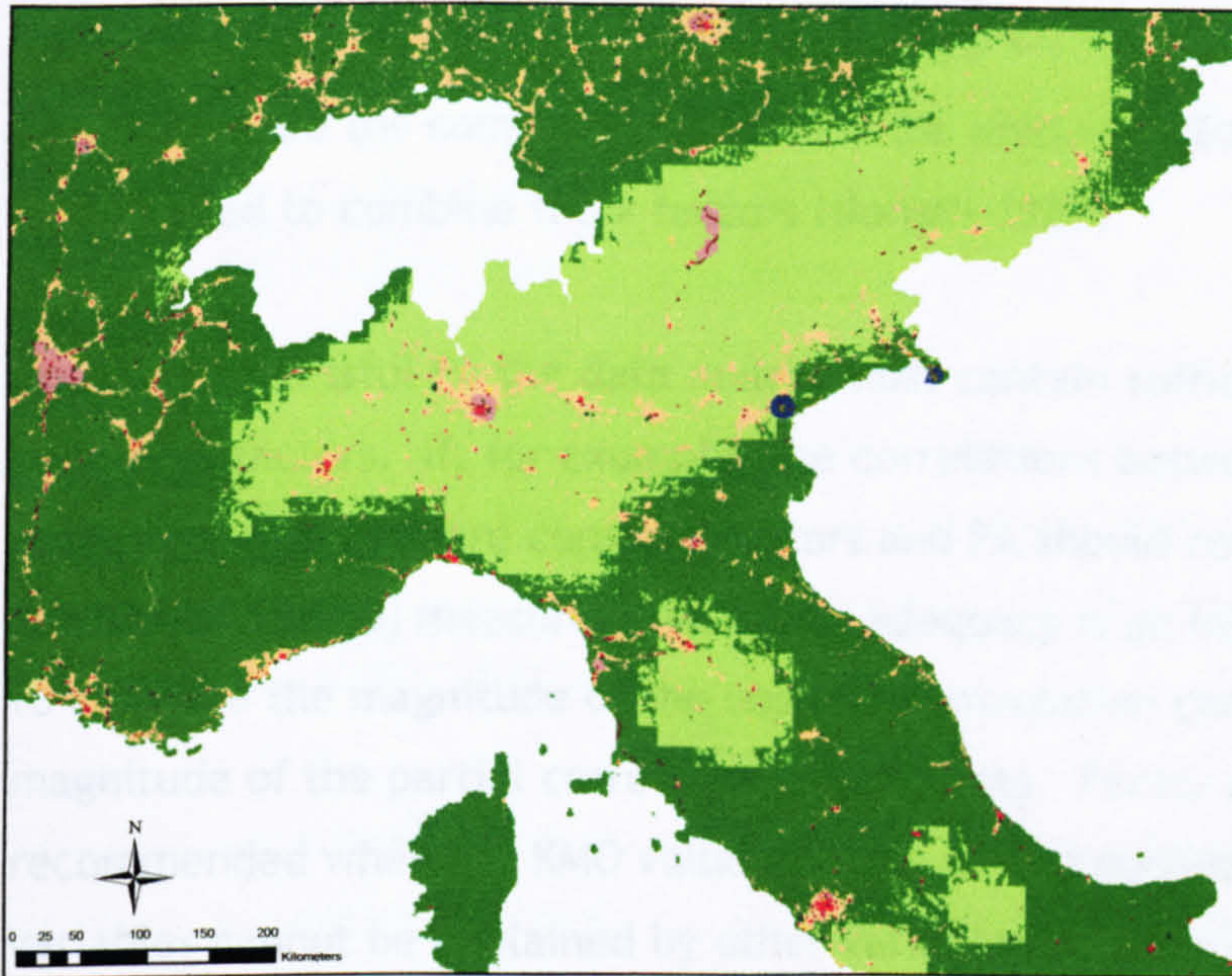


Figure 5.3 Example of blocky clusters

5.3.1.2 Factor analysis approach

Overview of factor analysis

Factor analysis (FA) is a multivariate statistical technique used to simplify complex sets of data. It attempts to condense the information in a large number of variables, by analysing the inter-correlations in order to define a set of common underlying dimensions or factors. Factors are essentially linear combinations of the original variables. The expression for the estimate of the j th factor (F_j) is given in Equation 5.1.

Equation 5.1
$$F_j = \sum_{i=1}^p W_{ji} X_i = W_{j1} X_1 + \dots + W_{jp} X_p$$

Where: W 's are the factor score coefficients; X 's represent the variables; and p is the number of variables (Norusis 1988).

Three types of variance are considered in FA. Common variance is defined as the variance shared by all variables in the analysis, while unique variance is the variance associated with only one variable. The third type, error variance, is due to uncertainty in data collection, measurement error or to a random component in the measured phenomenon (Hair et al. 1998). The model for the i th standardized variable is shown in Equation 5.2.

Equation 5.2
$$X_i = A_{i1}F_1 + A_{i2}F_2 + \dots + A_{ik}F_k + U_i$$

Where: F's are the common factors; U is the unique factor; and A's are the constants used to combine the k factors (Norusis 1988).

To run a successful FA the data matrix must contain sufficient correlations to represent factors. If, for example, the correlations between variables are small it is unlikely that they share common factors and FA should not proceed. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is an index that can be calculated to compare 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 range of values for KMO is from 0 to 1, and is interpreted according to the following guideline: 0.80 or above, meritorious; 0.70 or above, middling; 0.60 or above, mediocre; 0.50 or above, miserable; and below 0.50, unacceptable (Norusis 1988).

The eigenvalue, or latent root, represents the amount of variance accounted for by each factor. Each factor is essentially the best linear summary of the variance left in the data after the previous factors have been extracted; thus, the first few factors may explain most of the variance in the data set (Bailey & Gatrell 1995). Several criteria can be used to determine the optimal number of factors to extract:

- The latent root criterion says that any individual factor should only be retained if it accounts for the variance of at least a single variable (e.g. eigenvalue ≥ 1). All factors with eigenvalues less than one are considered insignificant and disregarded. This test is most reliable when the number of variables is between 20 and 50 (Hair et al. 1998).
- The percentage of variance criterion is based on obtaining a specific cumulative percentage of the total variance extracted by consecutive factors. The criterion was established *a priori* at 90%.
- The scree test criterion is used to identify the optimum number of factors that can be extracted before the amount of unique variance dominates the common variance structure. A scree plot is essentially a plot of the eigenvalue against factor number. The optimal number of factors is located where the plot shows a distinct break between the steep slope and the gradual tailing off of the other factors (Norusis 1988).

In AZS, the factors were extracted by the principal components method. This extraction method is used when the objective is to summarise most of the original information in a minimum number of components for prediction purposes and, because the resulting factors are based on the total variance, when prior knowledge indicates that specific and error variances represent a small proportion of the total variance (Hair et al. 1998). The principal components method provides uncorrelated linear combinations of the observed variables, where the first component has maximum variance, and successive components explain progressively smaller portions of the variance (SPSS 2003).

The final stage of FA involves interpretation of the extracted factors. A varimax rotation, in which the factors are rotated while the individual components remain orthogonal to each other, was used in AZS to simplify the spatial interpretation. This rotation increases the segregation between component loadings so that distinct clusters of inter-correlated data are more clearly defined (Yu & Chang 2001). The factor loadings, representing the correlation between the original variables and the factors, are used to interpret the dimensions that the factors represent in the original data. Variables with high loadings (e.g. an absolute value near 1) are said to define the particular components.

Factor analysis results

The use of FA to reduce the full data set in preparation for CA was initially piloted on Great Britain. Sixty-five of the available GIS variables, at various spatial scales, were used as input. These included the topography, annual mean meteorology, land cover, population, night-time light intensity, and the transport variables shown in Figure 5.4. As the objective of FA is to identify interrelated sets of variables, some correlation is desirable. The cut-off pair-wise correlation coefficient was, therefore, increased to $r \geq 0.95$. One of each pair of highly correlated variables was pruned prior to FA (shown in blue in Figure 5.4).

The latent root criterion was used to identify the appropriate number of principal components to extract. As depicted in Figure 5.4, 19 rotated principal components were extracted from the 56 variables entered into FA. Using a comprehensive set of input variables, however, resulted in a set of principal components that were extremely difficult to interpret, which further complicated the interpretation of the

clusters. For this reason, FA on the ‘full’ data set was rejected as a viable method for variable screening and it was not tested at the EU level.

Input Variables			19 Principal Components	
Altitude	AGR1	TRANS5	1	Low altitude agriculture, warm and dry
TOPEX	AGR3	TRANS11	2	LD residential, other built-up, B-roads and high population
NW slope	AGR5	IND1	3	HD residential
NE slope	AGR11	IND3	4	Central GB, low wind speed
SW slope	FOR1	IND5	5	Forest
SE slope	FOR3	IND11	6	Road/rail land cover types
Rail1	FOR5	LDRES1	7	Industry
Rail3	FOR11	LDRES3	8	Ports
Motorway1	WECS/URBGR1	LDRES5	9	Other built-up
Motorway3	WECS/URBGR3	LDRES11	10	Airports
A-road1	WECS/URBGR5	HDRES1	11	NE, rainy
A-road3	WECS/URBGR11	HDRES3	12	Northern facing, steep slopes
B-road1	AIRP1	HDRES5	13	Motorways
B-road3	AIRP3	HDRES11	14	SE facing, steep slopes
Minor road1	AIRP5	AM vapour pressure	15	A-roads
Minor road3	AIRP11	AS rain	16	Rail networks
Light	PORT1	AM temp	17	Western GB
Population	PORT3	AM radiation	18	Minor roads, no B-roads
NE sea dist	PORT5	AM temp range	19	SE facing slopes
NW sea dist	PORT11	AM wind speed		
SE sea dist	TRANS1	A calm		
SW sea dist	TRANS3			

Figure 5.4 Principal components derived from full data set

Notes: (1) Variables shown in blue were excluded prior to CA; (2) NE, SE, NW, SW = northeast, etc.; AM = annual mean; AS = annual sum: A = annual; 1, 3, 5 and 11 = 1 x 1km, etc. (3) see Table 3.3 for land cover abbreviations

5.3.1.3 Combined approach

As demonstrated above, neither the FA nor expert knowledge approach was entirely appropriate for screening the variables from which to derive the clusters. To overcome the downfalls of these methods (i.e. blocky clusters when using variables and difficult interpretation when clustering principal components), a combined approach was explored next.

The rationale behind the combined approach is that the pollution environments should be characterised by several types of variables - termed domains. These domains represent transport, land cover, topography and climate and the variables within each of these domains are measured at different spatial scales (i.e. 1km, 5km and 21km). Furthermore, some of the variables are grouped to represent landscape features not directly measured in the data, including the local effect of major roads (e.g. motorways plus A-roads constitutes 'major roads'), and the local and regional effects of built-up areas. Grouping select variables was also done to reduce the likelihood of obtaining tiny clusters representing a single land cover class; for example airports, as was apparent in the previous attempts (Figure 5.4).

While the predictors within some of these domains may suitably be represented by principal components, others may be better left as standardised variables. Initially, individual factor analyses were, therefore, run using the unstandardised variables in each domain (by spatial scale) listed in Table 5.1. Factor analysis was only done if:

- there were sufficient variables to make FA worthwhile;
- variables had sufficient levels of correlation; and
- variables showed approximately normal distributions.

The resulting KMO (for domains for which FA was run) was examined to determine whether factoring was in fact appropriate. For those domains where it was deemed inappropriate (i.e. $KMO < 0.7$), the factors were discarded and substituted with the standardised variables. This was done to ensure that the environmental conditions were described to the fullest possible extent. The merged rows in Table 5.1 indicate variables that, if not included as principal components, were summed to create a 'grouped' standardised variable. As per the previous approaches, the pair-wise Pearson correlations were examined prior to clustering. One of each pair of highly correlated variables/principal components ($r \geq 0.8$) was pruned prior to entering the final selection of inputs into CA.

Table 5.1 Variables (by scale) used in AZS

Variable	21km	5km	1km
Transport Domain			
Motorways		✓	✓
A-roads			✓
B-roads			✓
Minor roads			✓
Land Cover Domain*			
HD residential (1)	✓	✓	✓
LD residential (2)		✓	✓
Industry (3)		✓	✓
Transport (4)			✓
Airports (6)			
Seaports (5)			
Construction (7a)			
Urban greenery (7b)			✓
Forestry (8)			✓
Tillage (12)		PM ₁₀ only	
Topography Domain**			
Altitude (transformed)#	✓		
Topex (transformed)#	✓		
Distance to sea (transformed)#	✓		
Climate Domain**			
Summer and Winter Temperature	✓		
Annual Radiation	✓		
Annual Wind speed	✓		
Annual Calm Days	✓		
Annual Vapour pressure	✓		
Annual Rainfall	✓		

Notes: * Numbers in brackets for land cover variables refer to Table 3.3, redefined groups

** Variables are derived from coarse data

Altitude is calculated as $\sqrt[nalt]{nalt / \max(nalt)}$, where $nalt = altitude - \min(altitude)$

Topex is calculated as $\sqrt[ntopex]{ntopex / \max(ntopex)}$, where $ntopex = topex - \min(topex)$

Distance to sea is calculated as $\sqrt[\text{minimum distance}]{\text{minimum distance} / \max(\text{minimum distance})}$

Domain-based FA results

Figure 5.5 diagrams the correlations ($r \geq 0.5$ only) for the variables listed in Table 5.1. Positive correlations are displayed as solid lines while negative correlations are represented by dashed lines. High correlations ($r \geq 0.8$) are shown in cyan. As some degree of correlation is necessary to generate factors, the highly correlated variables within domains were not removed prior to running the domain-based FA, the results of which are described next.

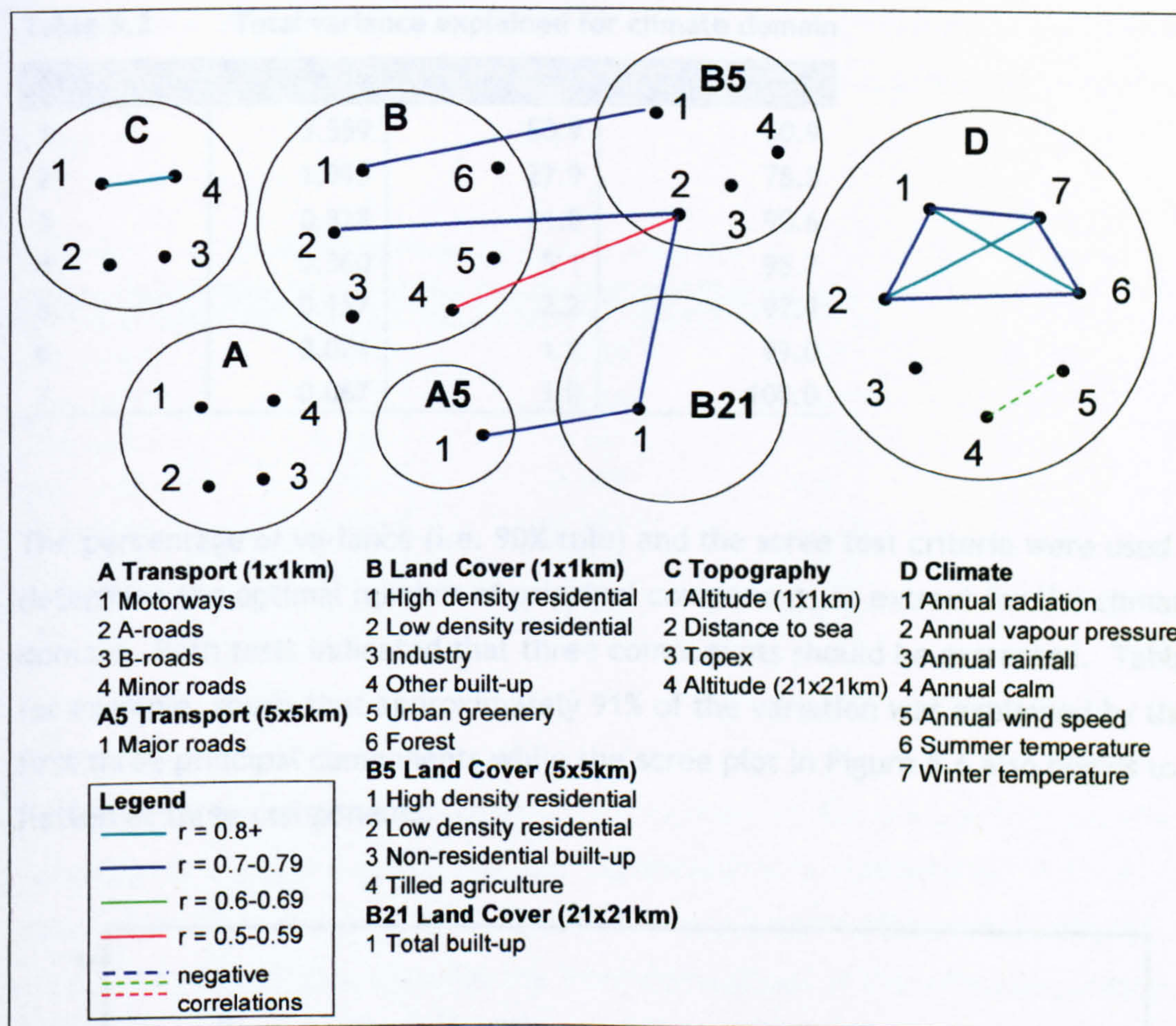


Figure 5.5 Correlation net for potential variables used in AZS

Each domain was examined in turn to evaluate the suitability of the constituent variables for FA. On the basis of the three criterion listed above, FA was not carried out for land cover (non-normality of the variables), or topography (an insufficient number of variables). Factor analysis, using the KMO values *post-hoc* to assess the appropriateness of the analysis, was performed on the other two domains. The KMO value for the 1km transport domain was below the acceptable limit (KMO = 0.54), so this, too, was considered inappropriate for FA. The FA on the climate domain, however, achieved a KMO of 0.70 which was considered adequate in this application. Table 5.2 shows the eigenvalues and total variance explained by each of the possible seven components (i.e. one component for each variable) in the climate domain.

Table 5.2 Total variance explained for climate domain

Component	Eigenvalue	% Variance	Cumulative %
1	3.559	50.9	50.9
2	1.995	27.9	78.8
3	0.828	11.8	90.6
4	0.360	5.1	95.7
5	0.157	2.2	97.9
6	0.074	1.1	99.0
7	0.067	1.0	100.0

The percentage of variance (i.e. 90% rule) and the scree test criteria were used to determine the optimal number of principal components to extract for the climate domain. Both tests indicated that three components should be extracted. Table 5.2, for example, shows that approximately 91% of the variation was explained by the first three principal components while the scree plot in Figure 5.6 also begins to flatten at three components.

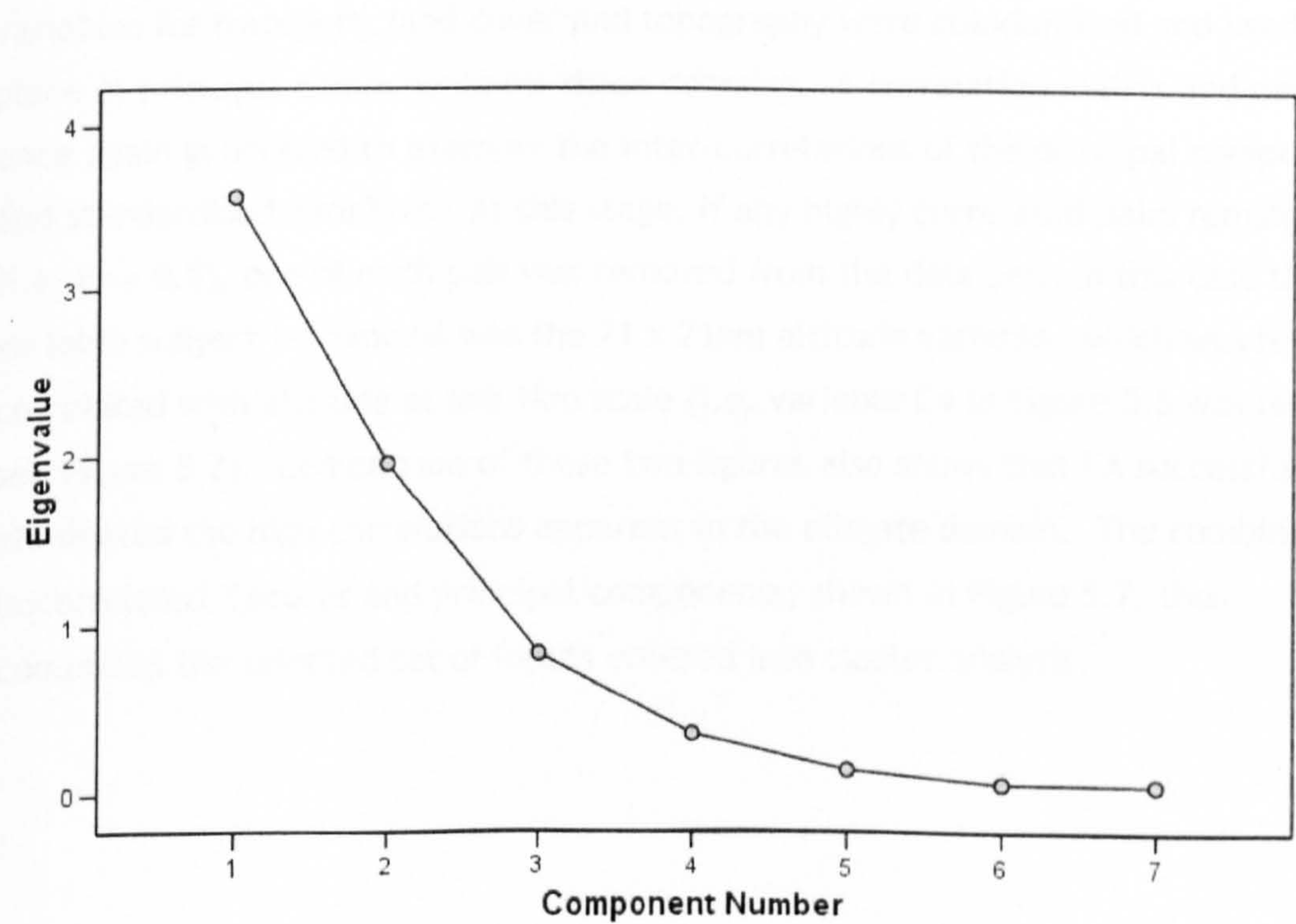


Figure 5.6 Scree plot for climate domain

The factor loadings for the climate domain are presented in Table 5.3 with the high loadings, used to interpret each of the resulting principal components, shown in blue. The results indicate that component 1 was defined in terms of year-round

temperature, radiation and vapour pressure; component 2 related to annual mean wind speed and number of calm days; and component 3 was essentially a rainfall factor.

Table 5.3 Rotated principal component loadings

Variable	Component 1	Component 2	Component 3
sm_tmp01	0.895	-0.136	-0.351
wm_tmp01	0.939	0.256	0.045
am_rad01	0.882	-0.181	-0.276
am_ws01	0.032	0.922	0.099
a_calm01	-0.071	-0.896	-0.029
am_vp01	0.922	0.238	-0.002
as_r01	-0.176	0.085	0.974

Final selection

As suitable principal components were only derived for the climate domain, the raw variables for transport, land cover and topography were standardised and used in place of principal components for those domains. A correlation matrix and net were once again generated to examine the inter-correlations of the principal components and standardised variables. At this stage, if any highly correlated pairs remained (i.e. $r \geq 0.8$), one of each pair was removed from the data set. In this case the only variable subject to removal was the 21 x 21km altitude variable, which was highly correlated with altitude at the 1km scale (i.e. variable C4 in Figure 5.5 was removed, see Figure 5.7). Comparison of these two figures also shows that FA successfully eliminated the high correlations apparent in the climate domain. The combination of uncorrelated Z scores and principal components, shown in Figure 5.7, thus, comprised the selected set of inputs entered into cluster analysis.

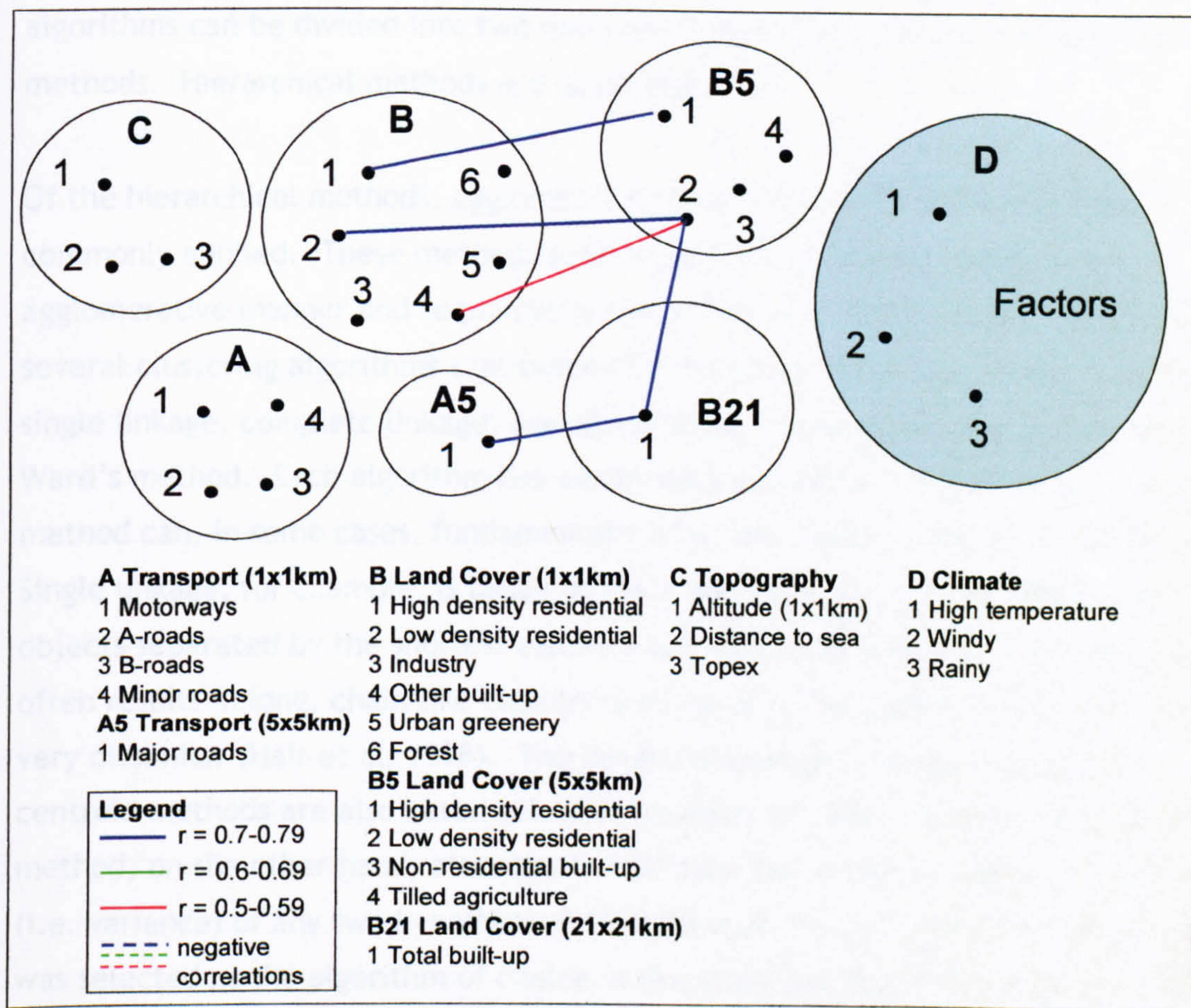


Figure 5.7 Correlation net for variables and factors used in AZS

5.3.2 Cluster analysis

Cluster analysis (CA) aims to group objects, so that those in a particular cluster are more similar to one another than they are to objects in any other cluster. Distance measures are most often used to measure similarity. Based on Euclidean distance, for example, objects having a smaller distance between them exhibit greater similarity (Hair et al. 1998). The Euclidean metric, where distance between points i and j is denoted by d_{ij} , is shown in Equation 5.3.

$$\text{Equation 5.3} \quad d_{ij} = \left\{ \sum_{k=1}^p (X_{ik} - X_{jk})^2 \right\}^{\frac{1}{2}}$$

Where: X_{ik} is the value of the k th variable for the i th case (Everitt 2003).

Association measures are used to create a similarity matrix ($N \times N$) that describes the strength of all pair-wise relationships. The most commonly used clustering

algorithms can be divided into two main categories: hierarchical and partitioning methods. Hierarchical methods are described first.

Of the hierarchical methods, agglomerative clustering methods are the most commonly applied. These methods search the $N \times N$ similarity matrix in an agglomerative manner and sequentially merge the most similar cases. There are several clustering algorithms that belong to this group of CA procedures, including: single linkage, complete linkage, average linkage, the centroid method and the Ward's method. Each algorithm has advantages and disadvantages and the choice of method can, in some cases, fundamentally alter the final clusters that are identified. Single linkage, for example, is based on minimum distance. It, therefore, finds two objects separated by the shortest distance and assigns them to the same cluster; this often results in long, chain-like clusters with cases at the opposite ends often being very dissimilar (Hair et al. 1998). The complete linkage, average linkage and centroid methods are also based solely on distance between clusters. The Ward's method, on the other hand, attempts to minimize the within-cluster sum of squares (i.e. variance) of any two hypothetical clusters that can be formed at each stage. It was selected as the algorithm of choice in this study because it tends to make stable, compact clusters. In the Ward's method, each case is in its own cluster and the within or error sum of squares (ESS) is nil at the first step of the CA. The groups or cases that result in the minimum increase in the error sum of squares then are joined at each successive stage (Everitt 2003).

Since hierarchical methods proceed in a stepwise fashion, the entire range of cluster solutions can be presented as a dendrogram, a tree-like diagram illustrating the fusions made at each level. The dendrogram or agglomeration schedule can then be used to determine the optimal number of clusters by plotting the cluster number by the fusion coefficient (e.g. squared Euclidean distance) at which various cases merge. The number of optimal clusters occurs where the plot flattens as, at this point, no new information is portrayed by additional mergers of clusters (Aldenderfer & Blashfield 1984).

In hierarchical methods, the $N \times N$ matrix is held in memory during computation, meaning that the computations can become highly demanding in terms of system performance. Hierarchical cluster analysis thus poses difficulties when analysing data sets that are large - as here. One way of avoiding these problems is to run the

analysis on repeated samples of data, and to compare the results in order to define the full cluster set. This, of course, raises the question of how to make these comparisons, and how to identify common clusters from different analyses. It can also be a time-consuming approach.

Another way of avoiding the computational limitations of hierarchical cluster analysis is to use partitioning methods. These are referred to as K-means CA, and use an algorithm designed to handle a large number of cases by dividing the objects into a predefined number of clusters. K-means CA can be thought of as reverse ANOVA, because the objects are moved between groups until a solution is found that minimises the within-cluster variability and maximises the between-cluster variability. K-means CA in SPSS begins by simultaneously selecting several cluster seeds and assigning objects within the threshold distance to the nearest seeds. It is based on nearest centroid sorting, where an object is assigned to the cluster for which the smallest Euclidean distance to the cluster centre is achieved (Norusis 1988). Partitioning methods differ from hierarchical methods in that they permit objects to move in and out of clusters, allowing for the correction of poor initial partitions at a later stage (Hair et al. 1998).

Both K-means CA (on the full data set) and hierarchical CA (on repeated samples of the full data set) were investigated in this study (Figure 5.1). To interpret the final cluster solutions derived from both analyses, the cluster centroids were calculated from the raw variables in SPSS using ANALYZE / REPORTS / CASE SUMMARIES. Similar to factor analysis, clusters with high loadings for any variables are said to be defined by those variables. The final cluster solutions were also mapped to aid interpretation and naming of the clusters. Prior to selecting the final solution, representing the air pollution affinity zones for the study area, a concordance analysis was undertaken to evaluate the sensitivity of the AZS procedure to input variables and clustering method.

5.3.2.1 K-means cluster analysis

Procedure

The K-means CA procedure in SPSS 12 is run by ANALYZE / CLASSIFY / K-MEANS CLUSTER. The challenge with K-means CA is that the practitioner must have some notion of the number of clusters that exist in the data. To address this issue one can obtain a range of cluster solutions by running the procedure several times. K-means CA was, therefore, run repeatedly to produce a range of cluster solutions from six to twenty-six, by steps of two.

When the objective of CA is to optimise the analysis using a particular clustering criterion, it is generally suggested that a plot of the criterion value against cluster number will indicate the correct number of clusters where a sharp increase/decrease (depending on the criterion) occurs in the plot (Everitt 2003). Such clustering criteria, however, tend to be based on statistical principles (e.g. within-cluster variability and between-cluster variability) considered too stringent for AZS. For example, while it might be important to minimise the within-cluster variability, it is also important to end up with reasonably distributed, sized and shaped clusters that represent real-world geographic features. The following indices, based on descriptive statistics and geographic principles, were thus established to provide a more flexible guide to selecting the optimal number of clusters:

1. Statistical index: The total (TSS) and mean sum of squared (MSS) distance to cluster centre was portrayed as a scree plot.
2. Map granularity index: A map granularity index was defined as the percentage of adjacent grid cells within a 3 x 3km and 5 x 5km rectangle window within the same cluster as the target cell. A low value indicates that there are many small or fragmented clusters, while a high value means that the clusters are very large. The *cluster_index.aml* (presented in Appendix C1) was written to calculate the map granularity index.
3. Map cluster index: The resulting clusters were mapped and a visual assessment of the cluster patterns was performed. Where possible, the clusters were interpreted and named.

Overall, solutions were sought that minimised the MSS and TSS while providing moderate granularity indices and a logical cluster map with interpretable clusters.

Results

After repeatedly running K-means CA on the Z scores/principal components (from Figure 5.7) to obtain a range of solutions, the optimal number of clusters was determined on the basis of the predefined statistical and map-based indices. Scree plots of the TSS and MSS plotted against cluster number are presented in Figure 5.8, while the map granularity indices are shown in Figure 5.9.

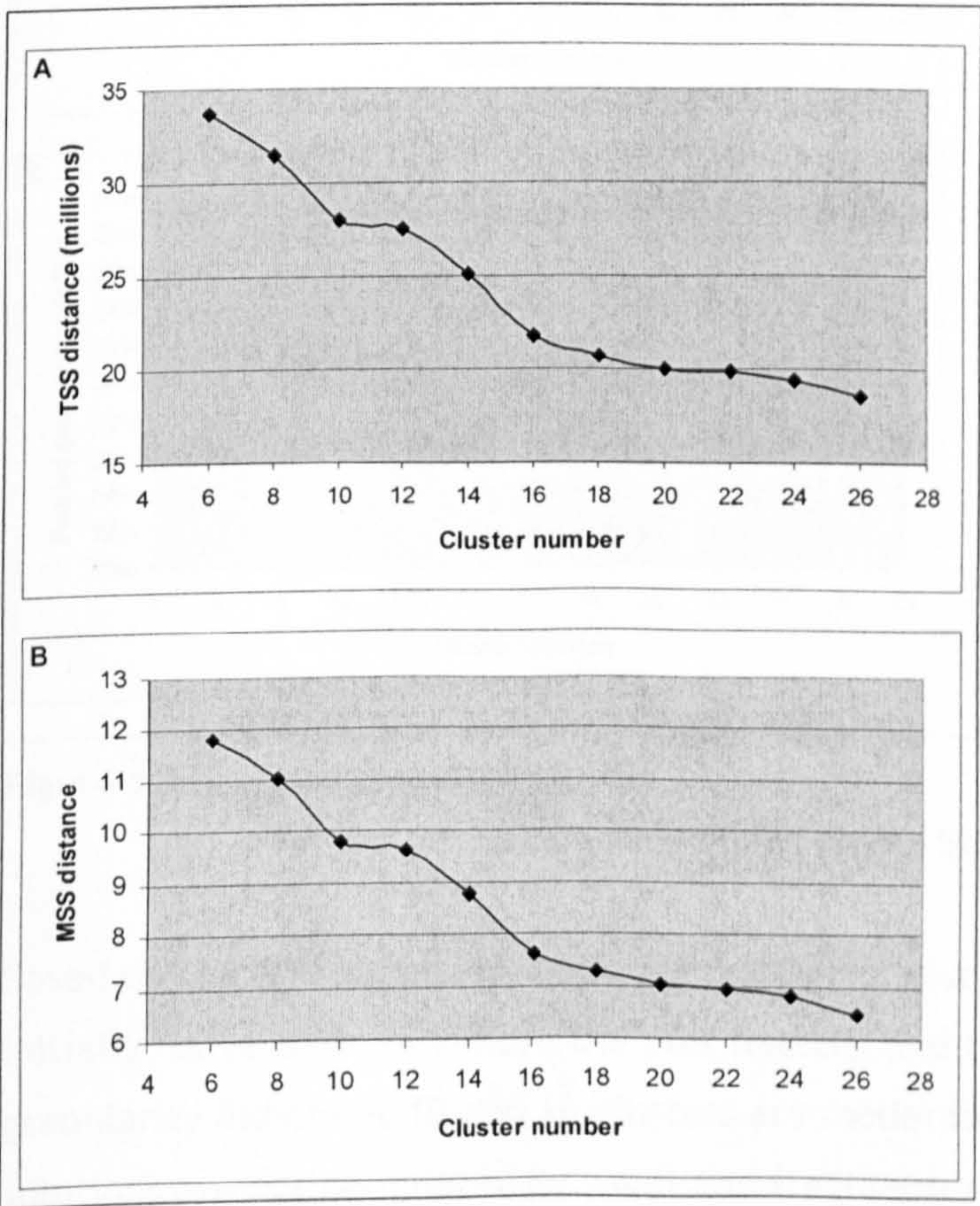


Figure 5.8 Statistical index
Scree plot of cluster number versus: (A) total sum of squared (TSS); and (B) mean sum of squared (MSS) distance to cluster centre

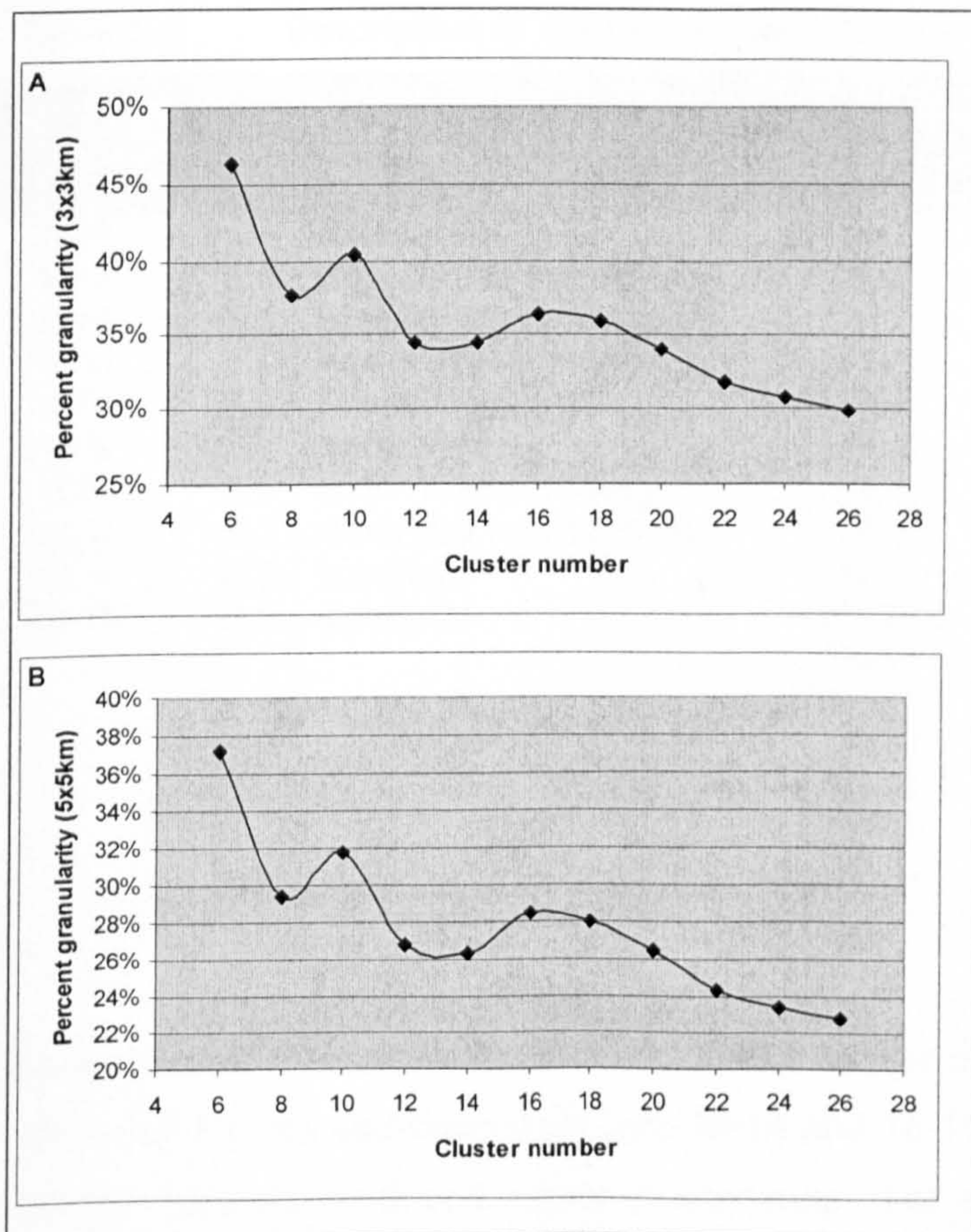


Figure 5.9 Map granularity index
For the (A) 3 x 3km window and (B) 5 x 5km window

Based on the scree plots, the optimal number of clusters could be 10 (where the plot initially flattens) or 16 (where the plot flattens and tails off). Additionally, the granularity indices at 10 and 16 clusters are moderate which indicates that these solutions are not dominated by small and fragmented or very large clusters. The wave-like structure of the granularity indices and step-like structure of the scree plots seems to suggest that there are some clusters that, when initially broken up, form loose clusters that must be further subdivided several times before settling into a stable partition of the data.

Given that a clear decision could not be taken solely on the scree and granularity plots, the 10 and 16 cluster solutions were interpreted and mapped so that the distribution of the clusters across the study area could be examined. A description of each cluster and its percent area for the two solutions under consideration is given in Table 5.4. The corresponding maps are shown in Figures 5.10 and 5.11.

Table 5.4 Description of clusters (10 and 16 cluster solutions - K-mean CA)

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 only
			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

Although the 16 cluster solution provided a better division of the large mixed rural class (10-1 splits approximately into 16-14 and 16-15), the fragmentation of the smaller clusters produced a poor cluster map. For example, minor roads were incorporated into the mixed rural (10-1) and forested upland (10-5) clusters for the 10 cluster solution, while they were represented by their own class (16-7) in the 16 cluster solution. While this may not appear to be a problem in Table 5.4, when the clusters were mapped one can see that the German landscape was dominated by minor roads in the 16 cluster solution (Figure 5.11). Additionally, a unique cluster arose in the 16 cluster solution representing built-up areas in the Netherlands only (16-10). The Netherlands may in fact have a unique urban fingerprint compared to the rest of the EU, but the 10 cluster solution provided a smoother map without obvious step changes between countries. In general, the 10 cluster solution better met the statistical and geographical criteria defined for creating EU-wide clusters, and was thus considered the final solution from the K-means CA procedure.

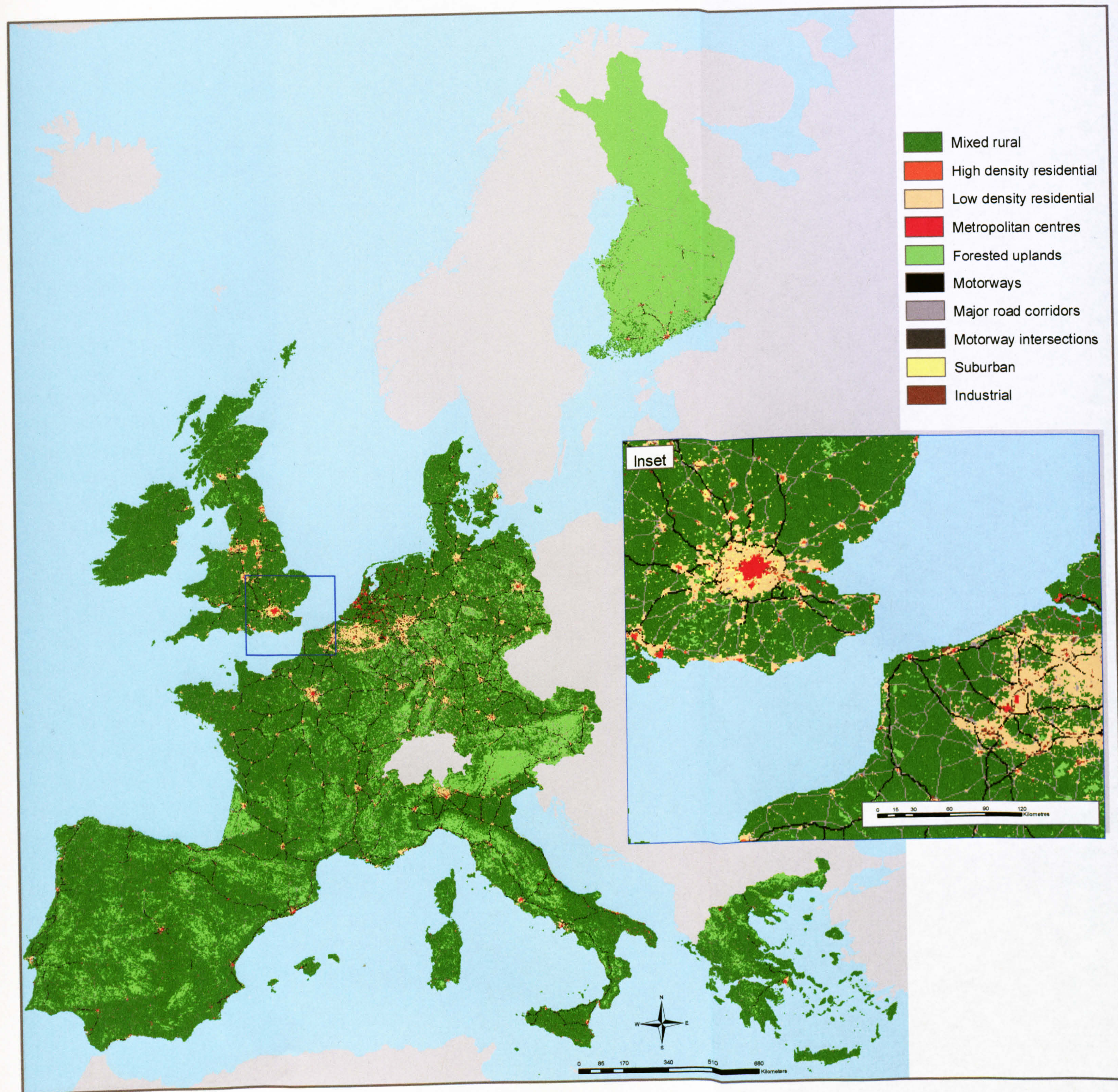


Figure 5.10
10 cluster solution (K-means CA)

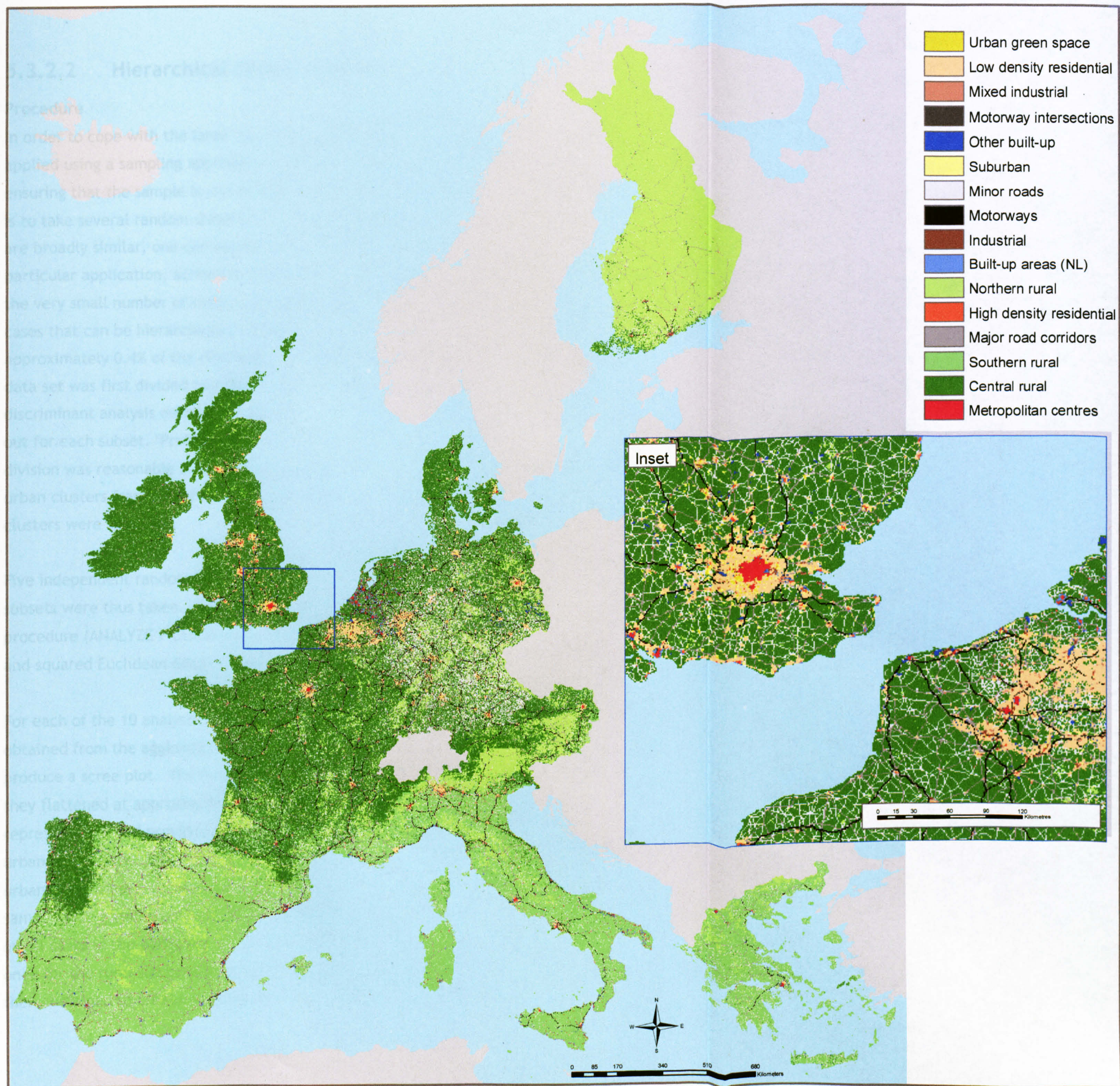


Figure 5.11
16 cluster solution (K-means CA)

5.3.2.2 Hierarchical cluster analysis

Procedure

In order to cope with the large data set (i.e. 2.8 million records), hierarchical CA was applied using a sampling approach. In general, the difficulty with sampling is in ensuring that the sample is representative of the overall data set. The usual solution is to take several random samples. If, when compared, the results from each sample are broadly similar, one can assume the samples were representative. In this particular application, achieving a representative sample was further complicated by the very small number of cases comprising each sample. The maximum number of cases that can be hierarchically clustered in SPSS, for example, is only 7,000 (i.e. approximately 0.4% of the database). To help overcome this particular problem the data set was first divided into urban and rural cells according to the results of the discriminant analysis detailed in Section 3.3.3, and separate analyses were carried out for each subset. Preliminary results from K-means CA confirmed that this division was reasonable in that the large rural clusters tended to persist while the urban clusters fragmented into smaller, dispersed areas as an increased number of clusters were specified.

Five independent random samples (of 7,000 cases each) for both the rural and urban subsets were thus taken. Each sample was entered into the hierarchical CA procedure (ANALYZE / CLASSIFY / HIERARCHICAL CLUSTER) using the Ward's method and squared Euclidean distance options. Ten analyses in total were run.

For each of the 10 analyses, the percent change in clustering coefficient was obtained from the agglomeration schedule and plotted against cluster stage to produce a scree plot. The five rural scree plots were examined to assess whether they flattened at approximately the same stage, indicating that the samples were representative and provided a stable partition. The same was done using the five urban scree plots. The five agglomeration schedules were also examined (by urban/rural subset) to evaluate if the clusters at each stage were essentially the same. Once a stable partition for each subset was confirmed, K-means CA was run on all cases in the particular subset to assign the cases to a cluster. This was done once for the rural subset and once for the urban subset of cells. The solutions were then combined to create a composite cluster map for the study area.

Results

The five scree plots for each of the rural and urban subsets (example shown in Figure 5.12) were examined to determine where each of the plots flattened. The five rural scree plots flattened at between four and seven clusters, while the urban scree plots flattened at between six and eight clusters. Examination of the agglomeration schedules by urban/rural subset confirmed that the clusters derived in these solutions were essentially the same, indicating that the samples were representative of the larger data set. On this basis, the mean solution from each subset was selected - 5 rural clusters and 7 urban clusters. K-means cluster analysis was then run individually on the rural and urban subset of cells to group the appropriate 1km cells into 5 and 7 clusters, respectively. A composite map containing 12 clusters, shown in Figure 5.13, was then created.

A description of each cluster and its percent area is given in Table 5.5. Although the proportions are slightly different, essentially the same clusters as those identified in K-means CA (Table 5.4) emerged. A more detailed comparison of the two clustering methods was conducted by a sensitivity analysis described in Section 5.3.3.

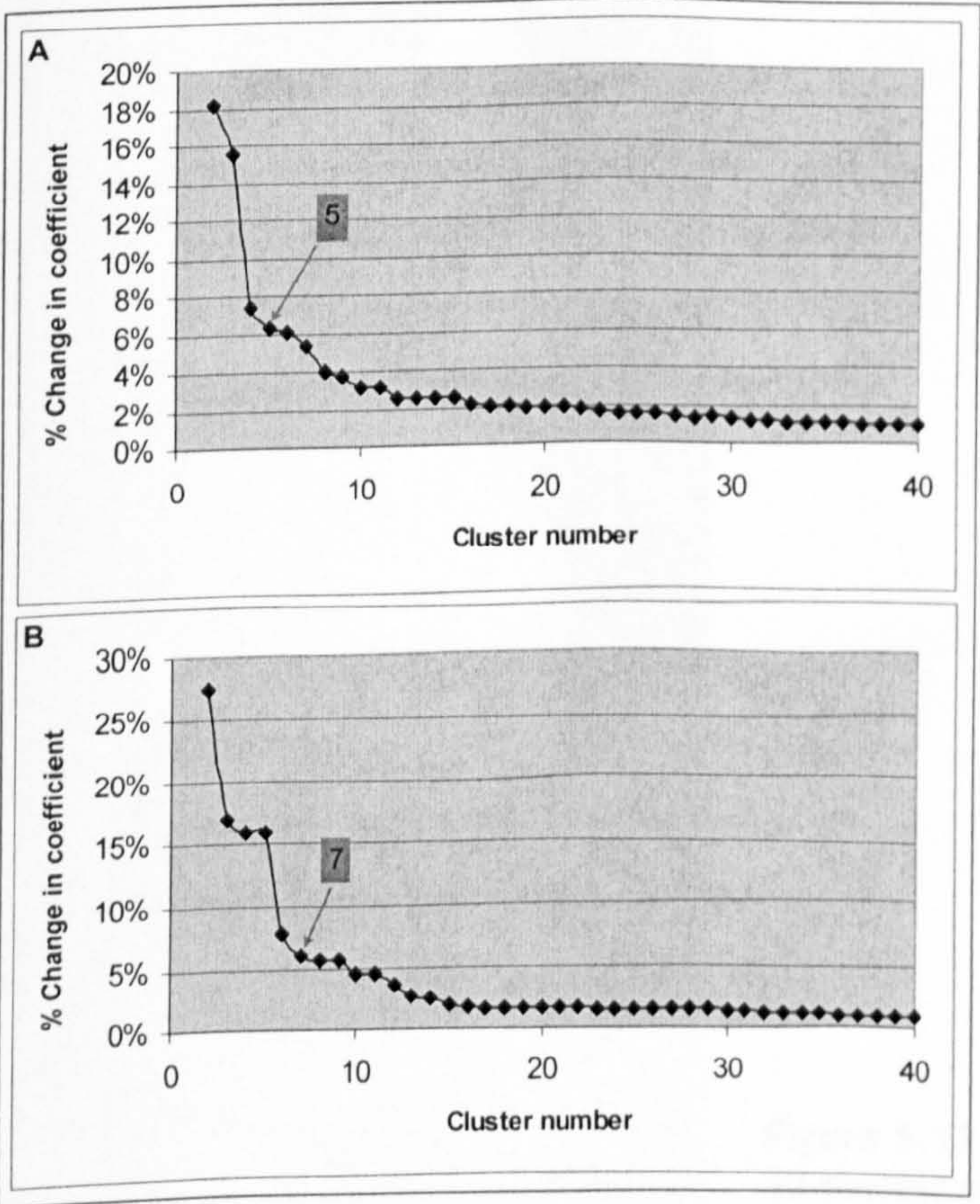


Figure 5.12 Example scree plots for hierarchical CA
(A) rural sample and (B) urban sample

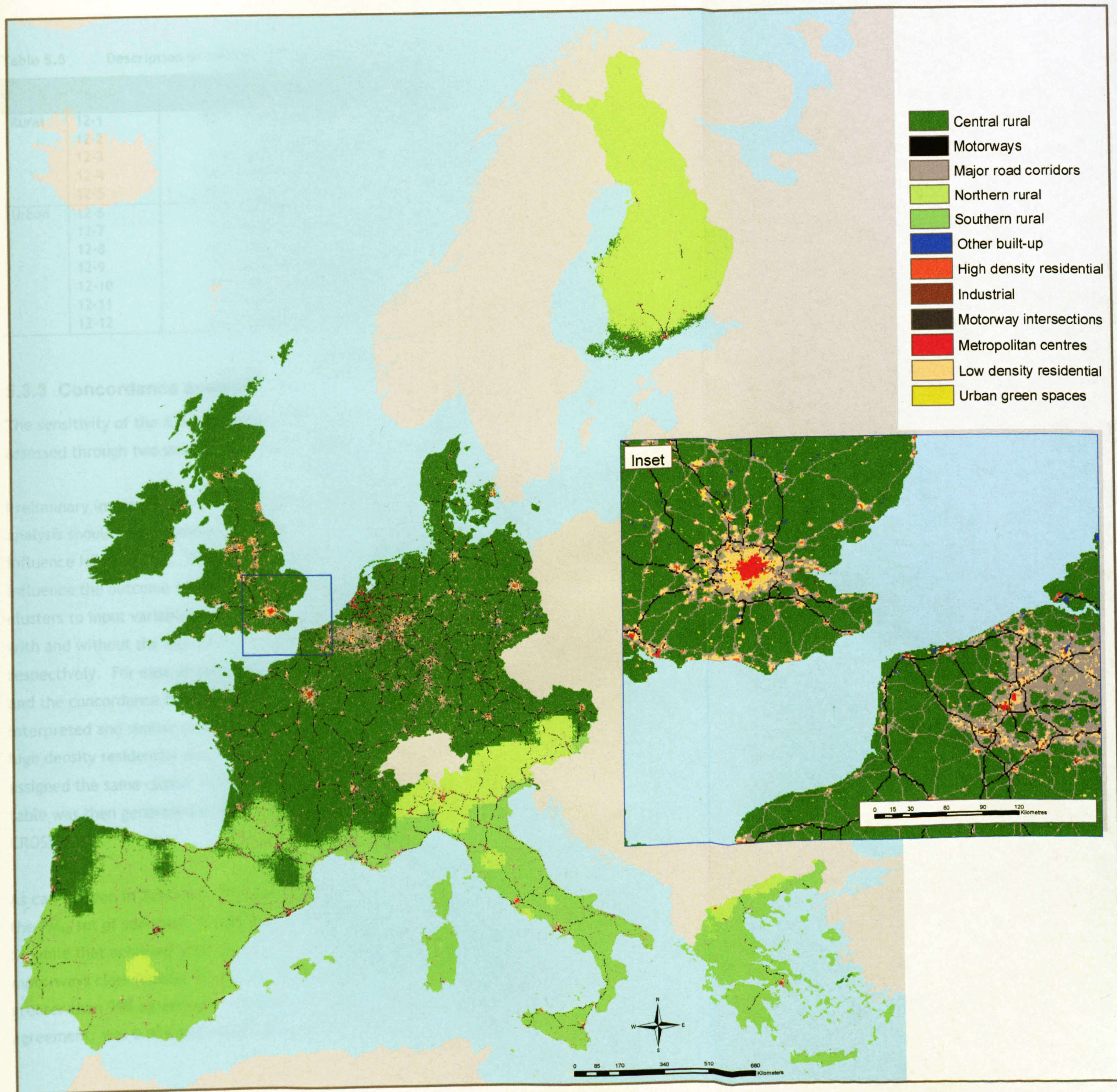


Figure 5.13
12 cluster solution (HCA)

Table 5.5 Description of clusters (12 cluster solution - HCA)

Subset of cells	Cluster No.	% Area	Description
Rural	12-1	40.7	Central rural
	12-2	1.3	Motorways
	12-3	6.8	Major road corridors
	12-4	16.2	Northern rural
	12-5	32.8	Southern rural
Urban	12-6	0.2	Other built-up
	12-7	0.2	High density residential
	12-8	0.3	Industrial
	12-9	0.1	Motorway intersections
	12-10	0.1	Metropolitan centres
	12-11	1.4	Low density residential
	12-12	0.2	Urban green spaces

5.3.3 Concordance analysis

The sensitivity of the AZS procedure to input variables and clustering method were assessed through two simple concordance analyses, described below.

Preliminary investigations in AZS involved an assessment of whether or not cluster analysis should be customised by pollutant using the variables specifically thought to influence NO₂ and PM₁₀. For example, would including a tillage variable for PM₁₀ influence the outcome of the CA? Sensitivity of the identification of landscape clusters to input variable selection was, therefore, assessed by running K-means CA with and without the inclusion of the tillage variable to represent PM₁₀ and NO₂, respectively. For ease of comparison, a 14 cluster solution for each CA was selected and the concordance between the two results was examined. The clusters were first interpreted and similar clusters assigned the same number in both solutions. The high density residential class identified in the NO₂-based solution, for example, was assigned the same cluster number as that of the PM₁₀-based solution. A concordance table was then generated in SPSS using ANALYSE / DESCRIPTIVE STATISTICS / CROSSTABS.

As can be seen in Table 5.6, thirteen of the clusters were similar. The solution with the PM₁₀ set of variables, however, contained a unique mixed industrial class (cluster X) while that achieved with the NO₂ set of variables produced a unique urban motorways class (cluster Y). Eleven of the 14 clusters were highly similar with greater than 99% agreement. The kappa value, which represents the overall agreement, was 0.984 and significant for the 13 matched clusters (excludes cluster X

and Y). Only two clusters, numbers 10 and 12, exhibited poor agreement; however, they were both motorway classes which were difficult to differentiate from each other during the interpretation.

Table 5.6 Percent agreement between NO₂ and PM₁₀ results

NO ₂	PM ₁₀														Total
	1	2	3	4	6	7	8	9	10	11	12	13	14	X	
1	99.5	0.1		0.0				0.1	0.3	0.2		0.0		56.6	0.5
2	0.0	99.9		0.0	0.0		0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.3	0.3
3	0.0		100.0						0.0	0.2					0.1
4				99.6	0.0		0.0	0.0	0.1	0.0	0.1	0.1	0.0	10.8	2.5
Y	0.3	0.0				0.1			26.2						0.1
6					99.0		0.3					0.6	0.6	10.7	41.5
7	0.2					99.9		0.1	0.0						0.1
8				0.0	0.8		99.0					0.1	0.3	7.6	32.6
9	0.0			0.0	0.0		0.0	99.5	0.0		0.0	0.0		0.3	0.4
10		0.0		0.2				0.1	73.3	0.1	10.6	0.0		0.4	0.3
11				0.0	0.0		0.0		0.1	99.5		0.0		0.3	0.3
12		0.0		0.0	0.0		0.0	0.0		0.0	89.2	0.1	0.0	1.1	1.1
13				0.1	0.0		0.0	0.0		0.0	0.0	99.0	0.1	8.1	4.2
14				0.1	0.1		0.7	0.0			0.1	0.1	99.0	4.0	16.1
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Notes: (1) Industrial; (2) Other built-up; (3) Urban green spaces; (4) Low density residential; (6) Central rural; (7) Metropolitan centres; (8) Southern rural; (9) High density residential; (10) Motorway intersections; (11) Suburban; (12) Motorways; (13) Major road corridors; (14) Northern rural; (X) Mixed industrial; (Y) Urban motorways.

The second concordance analysis involved the comparison of a K-means CA and hierarchical CA solution derived using the exact same input variables and resulting in the same number of clusters. As per the analysis described above, the clusters were first interpreted and similar clusters assigned the same number in both solutions. Once again a concordance table was then generated using the CROSSTABS command.

The concordance between the results obtained by K-means CA and hierarchical CA is shown in Table 5.7. The overall agreement in terms of the kappa value, which was significant, was 0.936. The percent agreement ranged from approximately 52% for two clusters, to >93% for more than half of the clusters. Clusters having the lowest percent agreement were all transportation classes (clusters 2, 3 and 9), which, as stated above, were difficult to differentiate during interpretation.

Table 5.7 Percent agreement between HCA and K-means CA results

K-means	Hierarchical CA												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
1	98.8		20.1	0.0	0.1						0.2		41.6
2	0.1	77.8	0.8	0.0	0.1	0.1	0.0	0.1	28.6	0.1	0.2		1.1
3	0.1	0.2	57.2	0.5	0.8		0.0	0.2	0.2		1.9		4.2
4	0.0		0.2	99.1	0.0						0.3		16.1
5	0.4		1.3	0.1	98.5						0.3		32.7
6	0.2	0.4	0.7	0.0	0.1	99.9	0.2	0.6	1.3		0.5	0.1	0.3
7	0.1	0.5	1.1	0.0	0.3		80.1	0.2	0.2		1.4	0.0	0.4
8	0.1	0.8	1.6	0.1	0.1		0.1	98.7	6.1		1.5	0.2	0.5
9		19.2					0.1	0.0	52.3		0.1		0.3
10		0.0	0.0	0.0			19.5	0.2	0.5	99.8	0.0	0.4	0.1
11	0.2	1.1	16.9	0.2	0.1		0.0		10.9		93.5	0.0	2.5
12	0.0	0.0	0.1	0.0	0.0					0.1		99.2	0.1
Total	100	100	100	100	100	100	100	100	100	100	100	100	100

Notes: (1) Central rural; (2) Motorways; (3) Major road corridors; (4) Northern rural; (5) Southern rural; (6) Other built-up; (7) High density residential; (8) Industry; (9) Motorway intersections; (10) Metropolitan centres; (11) Low density residential; (12) Urban green spaces.

Overall, these concordance analyses showed that slight changes in the input variables or clustering method did not significantly alter the outcome of the analysis. As a result, the use of a common set of input variables (Table 5.1) to define clusters for both NO₂ and PM₁₀ was considered justified. In terms of the preferred CA method, K-means CA was selected as it was more straightforward than hierarchical CA. Despite the approximately 3-fold longer computer processing time, interpretation of the K-means CA results was easier and quicker than for the HCA method. On this basis, the 10 cluster solution derived from K-means CA was selected as the definitive set of affinity zones upon which the pollution analysis and mapping was based. Figure 5.10, thus, represents the final affinity zone map.

5.3.4 Pollution characteristics of affinity zones

The monitoring sites in each affinity zone provide a basis for specifying pollution characteristics of the zone. If the delineation of affinity zones is successful in this respect, zones might be expected to show small within-zone variations in monitored concentrations, but large between-zone variations. The extent to which this was true was analysed using ANOVA techniques.

ANOVA is a statistical technique used to determine if samples from two or more groups originate from populations with equal means. It uses a single dependent variable (in this case monitored pollutant concentration) and two or more independent categorical variables (in this case affinity zone number and site types) to assess group differences.

In this analysis, ANOVA was limited to affinity zones with at least twenty monitoring sites in order to ensure an adequate number of cases to provide reliable representation of the pollution conditions across zones. The one-way ANOVA procedure in SPSS was run for each pollutant by site type (background and traffic) using ANALYSE / COMPARE MEANS / ONE-WAY ANOVA. The Bonferroni *post hoc* test was used to identify differences between each pair of means at 0.05 confidence level.

The ANOVA results presented in Table 5.8 indicate that the differences in concentrations (also depicted in the box plots, Figure 5.14) were significant for all analyses except background PM₁₀. For NO₂ background sites, highest concentrations 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. Additionally, as shown in the concordance analysis (Table 5.7, cluster number 3), the discrimination of 1km cells in this zone was inconsistent between K-means and hierarchical CA.

Post hoc tests confirmed that many of the NO₂ background sites were significantly different. None of the rural or traffic-related zones had a sufficient number of monitoring sites to be included in the analysis of traffic sites; thus, only the four built-up zones were considered. This fact alone indicates that some of the affinity zones are not adequately monitored. The highest concentrations at traffic sites were found in zone 4 (metropolitan centres) for both pollutants. Significant difference in concentrations were found between zone 4 and all other affinity zones for NO₂ and between zone 4 and zone 3 and 10 for PM₁₀.

Table 5.8 One-way ANOVA results

		Sum of Squares	df	Mean Square	F	Sig.
NO ₂	Background					
	Between Groups	49107.2	7	7015.3	92.0	0.00
	Within Groups	68044.8	892	76.3		
	Total	117152.0	899			
	Traffic					
	Between Groups	14624.8	3	4874.9	21.5	0.00
PM ₁₀	Background					
	Between Groups	175.6	4	43.9	0.9	0.46
	Within Groups	17829.6	366	48.7		
	Total	18005.2	370			
	Traffic					
	Between Groups	2596.0	3	865.3	8.3	0.00
	Within Groups	22337.2	215	103.9		
	Total	24933.1	218			

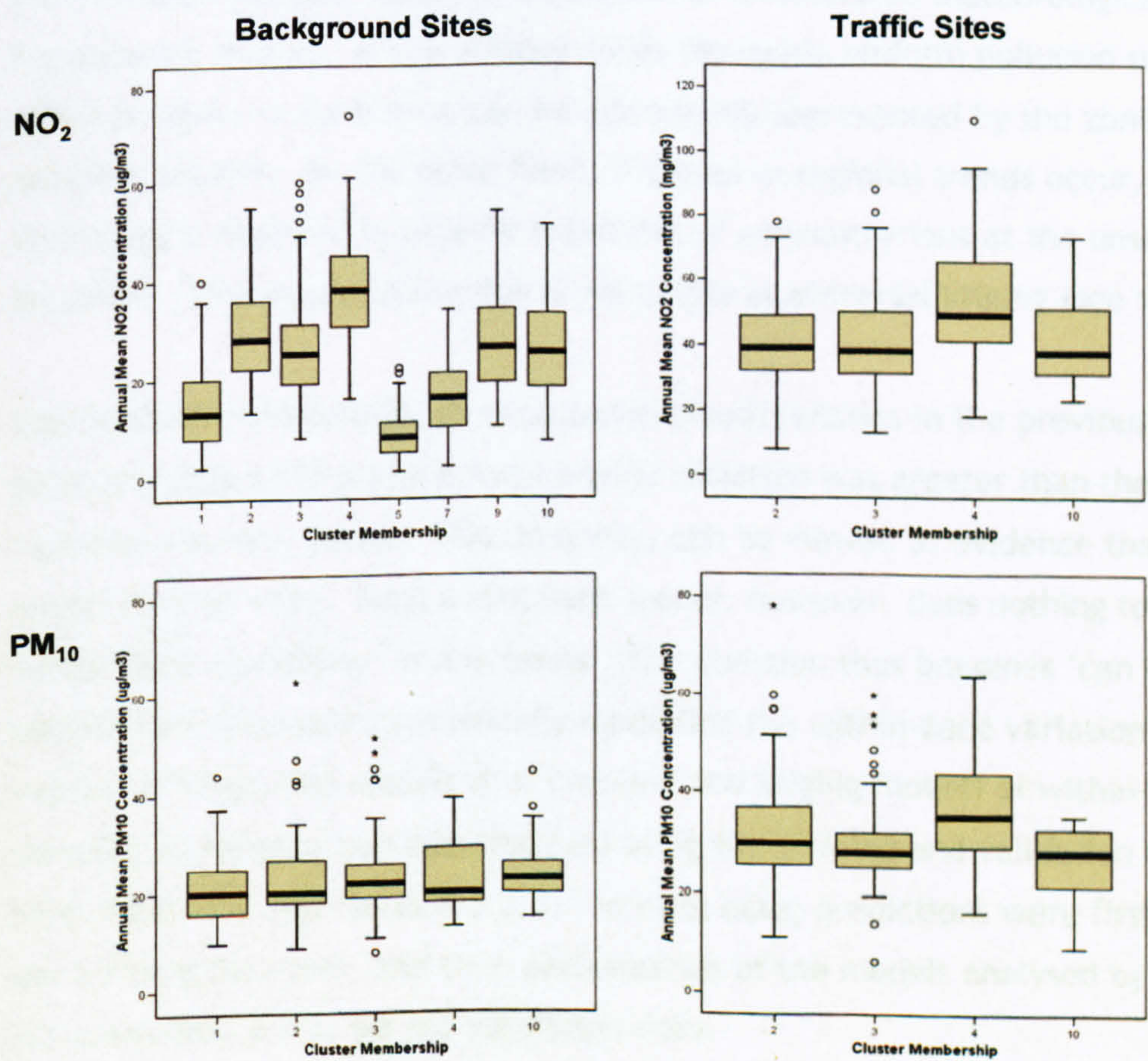


Figure 5.14 Box plots of air pollutant concentrations by cluster

5.4 Pollution mapping by affinity zone

In principle, the affinity zones define areas that are represented by 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. At the most basic, each zone might be expected to represent an area of uniform pollutant concentration (as was tested through the ANOVA investigation in the previous section). In this case, a disjointed surface would be anticipated, with little variation within each zone, but marked step changes at the zone boundaries. Alternatively, some degree of spatial trend might be expected within affinity zones; for example, from north to south reflecting broad-scale patterns of long-range air pollution. These trends might be constant across all zones (i.e. global), or may vary by affinity zone or region depending on more local factors such as characteristics of the emission sources, urban configuration or unmeasured meteorological and topographic factors. If the affinity zones represent uniform pollution surfaces, then concentrations in each zone can be adequately represented by the zone mean (i.e. baseline model). On the other hand, if global or regional trends occur, then further modelling is required to provide estimates of concentrations at the unsampled locations. One way of doing this is via simple or universal kriging (see section 2.3.1).

Exploration of the affinity zone pollution characteristics in the previous section generally showed that the between-zone variation was greater than the within-zone variation for most zones. This, in a way, can be viewed as evidence that the baseline model may be valid. Such a simplistic model, however, does nothing to quantify the within-zone variability - if any exists. The question thus becomes 'can the pollution surface be improved by specifically modelling the within-zone variation?' The validity of these two models (i.e. baseline and kriging model) of within-zone variation in air pollution was analysed using the training and validation data sets from AIRBASE-e (see Section 3.3.5). In each case, predictions were first made using the training data sets, and then performance of the models analysed by computing the prediction errors for the validation sites.

5.4.1 Procedure

The basic steps involved in mapping air pollution on the basis of affinity zones were as follows:

1. The training sites (designated for model building) were intersected with the affinity zone map (Figure 5.10).
2. The model of interest was applied in turn to each affinity zone:
 - a. Baseline model - the mean concentration from the training sites was calculated and applied to all cells in the zone.
 - b. Kriging model - using ArcGIS Geostatistical Analyst, simple kriging with a Gaussian model was applied to the training sites to construct a continuous surface. Simple kriging was selected over universal kriging because inconsistent trend was apparent in the data. To control for non-normality, a natural logarithm transformation was applied to the concentration data within the kriging procedure. In this application kriging was only applied to zones having at least 40 training sites, in order to ensure that sufficient pairs of sites were available in each lag (inter-site distance bin) to give adequate estimates of the semivariance. This limited the analysis to affinity zones 1-4 and 10 for background NO₂.
3. For each model, the predictions for each zone were combined in a composite map.
4. The validation sites were intersected with each composite map and measures of model performance calculated (see below).

5.4.2 Model validation

To assess the predictive capability of the models, and compare them with other modelling methods, it is necessary to validate each concentration surface with independent data. Several goodness-of-fit statistics were, thus, calculated using the reserved validation sites: R^2 , standard error (SEE), root mean square error (RMSE) (Equation 5.4), fractional bias (FB) (Equation 5.5) and factor-of-two (F2). These measures were also used to validate the air pollution models developed in Chapter 6, and have been widely used for testing and validating dispersion models such as ADMS (Carruthers et al. 2003; McHugh et al. 1999). They are also, for example, included in the American Society for Testing and Materials "D 6589-05 Standard Guide for

Statistical Evaluation of Atmospheric Dispersion Model Performance” (ASTM International 2006).

$$\text{Equation 5.4} \quad RMSE = \sqrt{\sum_{i=1}^n (Co_i - Cp_i)^2}$$

Where: Co_i is the observed concentration at monitoring station i ; Cp_i is the predicted concentration at station i ; and n is the total number of stations.

Factor-of-two represents the percentage of cases (i.e. sites) with predicted concentrations falling between one-half and twice the observed values. Fractional bias is a measure of agreement of the mean values. A negative FB value represents an under prediction while a positive value indicates over prediction. A value of 0.1, for example, represents an over prediction of 10%. Generally, the best model would have $FB = RMSE = 0$ and a F2 value of 100%.

$$\text{Equation 5.5} \quad FB = \frac{\overline{Cp} - \overline{Co}}{0.5(\overline{Co} + \overline{Cp})}$$

Where: Co is the average observed concentration; and Cp is the average predicted concentration.

Note for this particular analysis only:

One disadvantage of kriging in ArcGIS is that predictions cannot be made outside the extent of the input data (i.e. the training sites) so that the prediction map often does not provide full coverage of the study area. Recall from Table 3.10 that there were 714 training and 228 validation sites for background NO_2 . Validation was only possible, however, at 217 sites as the remaining 11 sites were outside the extent of predictions. A similar restriction was, therefore, applied to the baseline NO_2 model to ensure comparability between the validation results.

5.4.3 Results

The average and standard deviation of NO_2 and PM_{10} concentrations by cluster and site type are shown in Table 5.9 for the training sites. For the baseline model, the mean for each affinity zone was simply applied to all cells within that particular zone.

For the more advanced kriging representation, a Gaussian model was fitted to the variogram for each applicable affinity zone (the parameters of which are presented in Table 5.10). The variogram range is the distance beyond which there is little or no autocorrelation between the concentrations measured at the monitoring sites. The parameters in Table 5.10 show that the range varies substantially between affinity zones, from approximately 200km for the low density residential zone to 1,600km for the mixed rural zone. The mixed rural zone also has the largest nugget effect indicating that measurement error or the distance associated with microscale variations (or both) are greater for this zone compared to the others. The notable difference in range across models indicates that kriging within zones is likely to be more appropriate than applying a global kriging model, as the different zones have different distances across which the monitors may be considered representative. To confirm this assumption a global kriging model (also using simple kriging with a Gaussian model) was applied to all 714 of the NO₂ training sites. The parameters for this model are also given in Table 5.10.

Table 5.9 Pollutant concentration by cluster - training sites

Affinity Zone		Background			Traffic		
		N	Mean (µg/m ³)	SD (µg/m ³)	N	Mean (µg/m ³)	SD (µg/m ³)
NO ₂							
1	Mixed rural	125	14.16	8.11	12	32.92	19.10
2	High density residential	50	30.30	9.91	92	39.11	13.25
3	Low density residential	297	26.31	9.08	112	40.06	13.70
4	Metropolitan centres	54	38.06	11.83	99	51.56	17.44
5	Forested uplands	40	8.73	4.80	1	56.00	-
6	Motorways	13	25.62	10.97	8	47.50	8.02
7	Major road corridors	36	15.64	7.21	14	28.71	14.49
8	Motorway intersections	5	36.00	6.52	13	46.92	16.58
9	Suburban	25	26.72	9.23	7	36.14	22.70
10	Industrial	69	26.70	9.29	13	38.92	9.68
PM ₁₀							
1	Mixed rural	47	21.04	7.26	0	-	-
2	High density residential	35	23.23	8.63	50	32.72	9.93
3	Low density residential	141	23.04	6.70	54	30.65	9.67
4	Metropolitan centres	26	22.50	6.51	50	37.56	10.09
5	Forested uplands	11	13.73	4.13	1	12.00	-
6	Motorways	2	22.00	1.41	3	30.00	1.00
7	Major road corridors	15	21.40	7.12	7	27.43	8.79
8	Motorway intersections	1	25.00	-	6	25.50	8.38
9	Suburban	13	23.38	5.35	4	24.25	4.03
10	Industrial	35	23.74	6.45	18	26.11	8.32

Table 5.10 Simple kriging (Gaussian model) parameters - background NO₂

Affinity Zone		Range (km)	Partial sill (ln µg/m ³)	Nugget (ln µg/m ³)
1	Mixed rural	1,636.7	0.397	0.164
2	High density residential	630.1	0.115	0.066
3	Low density residential	217.4	0.118	0.065
4	Metropolitan centres	1,207.7	0.182	0.021
10	Industrial	567.8	0.135	0.068
All zones (global kriging)		1,970.0	0.383	0.223

The validation results for the two models are presented in Table 5.11. These results show that baseline model was inadequate in characterising PM₁₀ as indicated by the low R² (0.07 for background and 0.01 for traffic). The baseline model was also a poor predictor of traffic-related NO₂ across the affinity zones (R² = 0.17). A comparison with the kriging model could not be made for these components due to the lack of sufficient monitors per affinity zone.

The baseline model produced a relatively reasonable map of background NO₂ concentrations, with a 42% percent agreement between the measured and modelled concentrations at validation sites. This result was improved to 51% by applying kriging to the five affinity zones with sufficient numbers of NO₂ monitors. Reduction in the SEE, RMSE and fractional bias and an increase in the factor-of-two for the kriging model also indicate improved performance. In comparison, a poor result was obtained by the global kriging model (R² = 0.27; SEE 9.029), generated independent of the affinity zones. This further indicates that variation in pollution concentration does occur at different scales (i.e. ranges) depending on the pollution environment, and confirms the assumption that the sites in different affinity zones are representative over different distances. Better models, therefore, are derived if these differences are taken into consideration.

The validation statistics by each zone for which kriging was performed were also calculated. The R² and SEE could not be calculated for the baseline model in SPSS because the predicted values by zone were constant. The RMSE, factor-of-two and fractional bias, however, are reported in Table 5.12. The validation results generally show that kriging within zones produced better models of the within-zone variability compared to the baseline model. Exceptions include an elevation in RMSE for the high density residential zone and a slight increase in bias associated with the mixed rural zone kriging models.

The increase in RMSE may, in part, be caused by an insufficient number of monitoring sites in the high density residential zone to characterise the pollution surface (Table 5.12). The best kriging models were obtained for the low density residential and metropolitan centre zones. For the low density residential areas, this may be a result of the large number of monitoring sites present in the zone. The metropolitan centre zone, however, has significantly fewer monitoring sites. This is a reminder that the coverage or number of sites itself is a poor indicator of the representativeness of the monitoring network. The variability in the pollution environment, which has been shown to differ both between as well as within zones, should be an important driver in determining the optimal number (and location) of the monitoring sites.

Table 5.11 Model validation results across all affinity zones

Type	Model	N	Validation Sites						
			R ²	Adj R ²	SEE	RMSE	F2	FB	Method
NO ₂ models									
background	baseline	942	0.42	0.42	8.035	8.126	88.9%	0.053	LS (n=217)
	kriging	942	0.51	0.51	7.368	7.415	91.7%	0.024	LS (n=217)
	Global kriging	942	0.27	0.27	9.029	9.037	84.8%	0.006	LS (n=217)
traffic	baseline	500	0.17	0.16	15.671	3.554	96.1%	-0.061	LS (n=129)
PM ₁₀ models									
background	baseline	430	0.07	0.06	6.766	6.702	98.1%	-0.004	LS (n=104)
traffic	baseline	251	0.01	-0.01	11.981	12.22	89.3%	0.032	LS (n=56)

Table 5.12 Model validation results by affinity zone

Affinity Zone		Model	Total N	Validation Sites			
				N	F2	FB	RMSE
1	Mixed rural	baseline	163	38	65.8%	-0.036	9.053
		kriging			76.3%	-0.058	8.344
2	High density residential	baseline	69	19	94.7%	0.047	7.332
		kriging			94.7%	-0.015	8.205
3	Low density residential	baseline	384	87	96.6%	0.089	7.836
		kriging			98.9%	0.047	6.450
4	Metropolitan centres	baseline	70	16	100.0%	-0.002	6.981
		kriging			100.0%	-0.006	6.284
10	Industrial	baseline	93	24	91.7%	0.071	9.938
		kriging			91.7%	0.058	9.255

The area around Madrid is shown in Figure 5.15 to demonstrate the difference between the baseline and kriging models. As expected, the surface derived from the baseline model is notably flat (i.e. showing little variation in concentrations) in comparison to the kriging model because all cells within each affinity zone were assigned the same concentration.

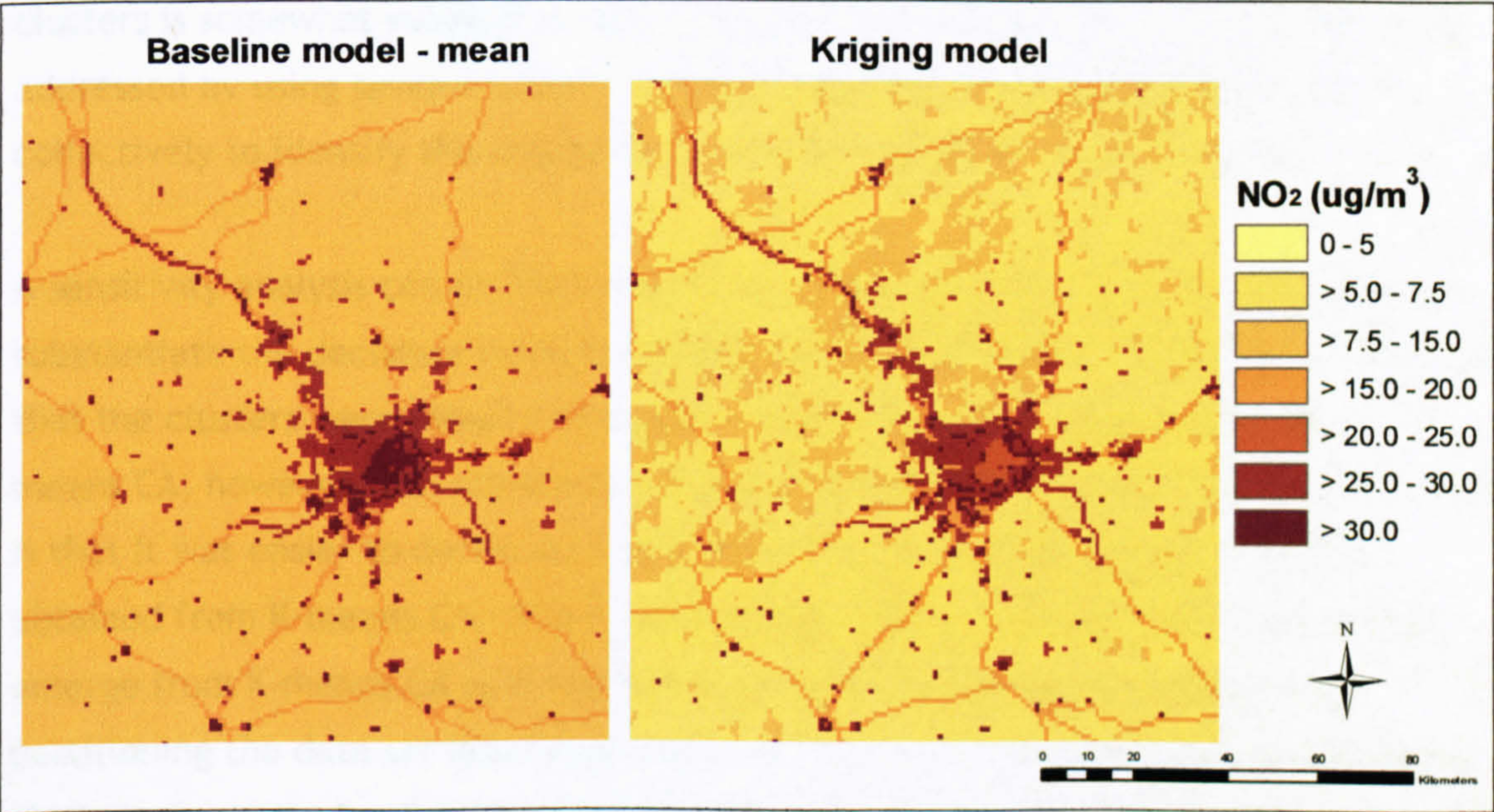


Figure 5.15 **Comparison of baseline vs. kriging model**
Background NO₂ concentrations (µg/m³) in the area surrounding Madrid

5.5 Discussion

5.5.1 Creating air pollution affinity zones

The objective of AZS was to develop and test a methodology for defining affinity zones, using multivariate statistical techniques, as a way of characterising the pollution environments represented by the monitoring network. Ten interpretable and coherent zones were identified, comprising four related to densely populated areas (metropolitan centres, high density residential, low density residential and suburban), four representing industry and transport infrastructure, and two rural zones (Figure 5.10).

The main difficulty with the AZS procedure was in choosing the most appropriate CA method and correct number of clusters to extract. While hierarchical CA can help determine the optimal number of clusters in the data set by way of the

agglomeration schedule, the number of records involved in this study greatly exceeded the maximum for the SPSS procedure. The only way in which hierarchical CA could be used was to run repeated analysis on very small samples (0.4%) of the data set. On the other hand, while K-means CA can provide the best partition when the number of clusters is known, it is often difficult to determine the 'optimal number' *a priori*. As discussed in Carter (1997), selecting the optimal number of clusters is somewhat subjective, and trial and error is necessary. This problem was addressed by using several criteria, based on statistical and geographical indices, collectively to identify the appropriate number of clusters for K-means CA.

A sensitivity analysis comparing K-means CA with hierarchical CA was undertaken to substantiate the decisions taken in implementing each method. This analysis showed that the clusters were robust to changes in the clustering method (Table 5.7). K-means CA, however, was chosen over hierarchical CA for two main reasons. The first is that it was easier to derive an appropriate solution and interpret the clusters obtained from K-means CA. The second reason is that more natural clusters could emerge from K-means CA as it was not necessary to constrain the analysis by partitioning the data set into urban and rural subsets. One would expect that some clusters, in particular those related to transport, would occur in both urban and rural areas.

A second sensitivity analysis (Table 5.6) also showed that the clusters were robust to minor changes in the input variables, indicating that the same set of clusters were suitable for evaluating both the NO₂ and PM₁₀ monitoring sites. If, however, one wanted to define affinity zones for another pollutant with very different sources or dispersion properties, such as ozone, appropriately customising the input variables may result in a different set of affinity zones.

As was discussed in Briggs & France (1983), the success of these classification methods are, to a great extent, dependent on the initial choice of input variables. Bunce et al. (1996a) made a similar comment, stating that if the input variables are mostly climatic, for example, then the classification will be predominantly based on climatic factors. This was clearly demonstrated in the variable screening stage of AZS where attempts to define the set of input variables produced blocky maps dominated by the meteorological data (Figure 5.3). To arrive at the best stratification of the landscape and physiographic features to define air pollution

affinity zones, the approach, in the words of Carter (1997), had to be exploratory, iterative and time consuming! It needs to be emphasised, in this context, that there will usually be no single, ideal set of clusters that can characterise a monitoring network. On the other hand, the process of constructing and comparing different stratifications 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 outlined below.

5.5.2 Characterising pollution within affinity zones

Once a set of ten affinity zones had been identified, the next stage of AZS involved assessing the within-zone variability of measured concentrations to determine if they truly represented air pollution affinity zones. ANOVA techniques were used for this purpose.

Except for NO₂ background sites, which generally provided good coverage of the affinity zones, at least half of the zone had to be excluded from the analysis as they contained too few monitoring sites (i.e. <20 sites). This in itself may be taken as evidence that the monitoring network is too sparse in some areas. This was a particular problem with the traffic sites, where sufficient site numbers were available only for zones 2, 3, 4 and 10 (high and low density residential, metropolitan centres and industrial areas, respectively). Notably, none of the traffic-related affinity zones had sufficient NO₂ or PM₁₀ coverage by traffic sites to facilitate the analysis. This presents a problem not only for AZS, but also for air pollution modelling by other means on the basis of this network. The lack of a sufficient number of monitoring sites in some area, especially those known to have variable pollution surfaces or high populations, may be a major limitation to air pollution modelling. For this reason, modelling efforts in Chapter 6 were limited to characterising annual mean background concentrations across the study area; the traffic component was not modelled.

ANOVA identified differences between some of the affinity zones in the background NO₂ concentrations. The rural zones 1 and 5 (mixed rural and forested upland) had significantly different concentrations, and were both significantly different from the four urban zones 2, 3, 4 and 9 (high and low density residential, metropolitan centres and suburban, respectively). Differences in the performance of the AZS were also

found between NO_2 and PM_{10} . While zones with distinct background pollutant concentrations could be defined for NO_2 , those for PM_{10} were 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 - a result, perhaps, of the greater influence of long-range particulate transport. The difficulties of modelling PM_{10} are further discussed in the Chapter 6 in relation to the focalsum method and results.

5.5.3 Mapping concentrations with affinity zones

Section 5.4 explored the advantage of mapping air pollution on the basis of the affinity zones. First, maps were derived by applying the mean concentration from the monitoring sites in a particular affinity zone to all cells within that zone. As demonstrated by validation with independent monitoring stations, this simplistic method generally provided only a moderate characterisation of the pollution surface. This was not a surprising result as pollution is expected to vary not only between affinity zones, but also within zones. To explore the within-zone variability, kriging by affinity zone was explored where there were sufficient numbers of sites. This was only possible in five of the affinity zones for NO_2 sites measuring background concentrations. In these cases, however, kriging reduced the errors (SEE, RMSE) in predictions of background NO_2 concentrations by almost 10% and gave marked improvement in the correlation coefficient.

This analysis demonstrated the potential for modelling on the basis of affinity zones. The success of kriging, however, relies on having enough monitoring sites in each affinity zone. Additionally, the predictions from simple kriging are based only on the location of the monitors. A more sophisticated method such as universal kriging to take account of global trend may be better suited to air pollution which often exhibits broad-scale variation. Universal kriging, however, was not applied as the monitoring data in the affinity zones showed no clear, unidirectional trend. Co-kriging or regression mapping, which include covariates as proxies of emissions and the dispersion environment, are also often better suited to the task of characterising the more local variations in air pollution. Applied within affinity zones, these methods may greatly improve the maps of long-term air pollution across the study area.

5.5.4 Assessing representativeness

Many of the air pollution modelling methods reviewed in Chapter 2 rely on a monitoring network to enable predictions. A monitoring network with adequate spatial coverage is, thus, vital to the success of pollution modelling. This is especially important if an existing network is to be used in epidemiological applications to link exposures to air pollution to effects on human health. A major consideration in this context is that ground-based routine monitoring networks are normally designed for assessing whether air quality standards have been exceeded - not as a basis for exposure assessment. The value of data from routine monitoring networks may also be limited given that the areas covered by monitoring and where people actually reside are not always the same (WHO 1999).

One must also recognise that monitoring stations provide information for point locations and, thus, present only a partial picture of the air pollution situation in the vicinity of the monitor. Additionally, the spatial and temporal variability of ground level air pollution, which was discussed in Section 2.1.3, is difficult and expensive to resolve with routine monitoring networks. A very dense network of monitoring stations, for example, would be required to describe these variations to a high degree of accuracy. Costs prohibit this level of monitoring on a routine basis and, even when purpose-designed monitoring is carried out as part of an epidemiological study, the number of monitoring sites is rarely adequate.

Against this background, there are obvious benefits in using modelling to improve the spatial and temporal resolution of data obtained from the monitoring networks. Prior to developing spatial models on the basis of a monitored data, however, some form of evaluation of the network should be undertaken to gain insight into its suitability for making predictions at unmonitored locations. In other words, we need to confirm the notion that the monitoring sites are representative of the surrounding area before using them as a basis for modelling. In addition to mapping concentrations, the affinity zones potentially offer a means by which the representativeness of the network can be assessed. This topic is further explored in Chapter 7.

5.6 Conclusions

Guided by established classification methods in soil and ecological mapping, multivariate statistical techniques were explored to define air pollution affinity zones across the EU. The affinity zones were then used as a basis for mapping long-term pollution concentrations, with some success for background NO₂. Further research, however, is needed to fully evaluate the concept of mapping pollution on the basis of affinity zones. In particular, nesting more complex models (i.e. those that incorporate covariate information) into the affinity zones might improve predictions by better characterising the within-zone variability of the pollution surface. With good covariate information regression mapping might be particularly advantageous because, unlike kriging, it does not rely on spatial structure of the monitoring sites. This leads to a final point - that of the spatial representativeness of the monitoring network for exposure assessment. The affinity zones, in association with detailed information on population density, potentially offer a forum in which this topic may be addressed.

6 Modelling Air Pollution Using Moving Window Methods

6.1 Introduction

Mapping exposures to air pollution relies on assessing concentrations at locations where people live and spend their time. It is obviously impossible to measure pollution everywhere, so we frequently depend on data from a network of monitoring stations at point locations as a basis for exposure assessment. As the previous chapter has indicated, however, to translate these point data into measures of exposure, the data need to be extrapolated in some way to all the unmonitored locations where exposures actually occur.

The challenge facing practitioners in this context is to devise a method that can suitably represent the variable pollution surface using the often limited number of monitoring stations. Even a very dense network, which may only be available at a city or regional level through purpose-designed monitoring, is rarely adequate without further modelling to resolve the spatial and temporal variations we wish to represent (Brauer et al. 2003; Briggs et al. 1997). Various methods have been developed for this purpose. Researchers have, for example, explored methods such as location-allocation techniques to site monitors optimally in areas of the city exhibiting the greatest variability or highest population (Kanaroglou et al. 2005). The EU has also considered this issue in providing guidelines to the Member States for selecting sites for the EUROAIRNET monitoring network (Larssen & Lazaridis 1998). Affinity zone stratification, applied to the AIRBASE-e sites (Chapter 5), has nevertheless suggested that the network coverage remains less than ideal in many areas of Europe and the monitoring data alone cannot provide a complete picture of exposures for the population of the EU. Clearly the monitoring data need to be enhanced through the use of models.

Possibly the most reliable way of estimating concentrations (and, by implication, exposures) at unsampled locations is through the use of dispersion models (Section 2.3.3). Advances in modelling methods and in the understanding of atmospheric processes has led to the development of sophisticated dispersion models which can give reliable estimates of pollution at unmonitored locations when calibrated and validated with monitoring data (Carruthers et al. 2003; Colville & Briggs 2000; McHugh

et al. 1999). For two main reasons, however, dispersion modelling is not yet appropriate for generating high resolution EU-wide pollution estimates. The first is the need for powerful computing facilities, especially when the models are to be applied across large areas and at high spatial or temporal resolution. The second is the severe data requirements of the sophisticated dispersion models. While suitably detailed information on source distribution, activity and emissions rates, meteorological conditions and surface terrain is often available at a local level, these data are rarely available across the whole EU at an adequate spatial resolution. Further complication is added in modelling pollutants such as particles, where it is necessary to combine both local and long-range dispersion models if the aim is to map total concentrations.

Geographic information systems offer an alternative to dispersion modelling. As already mentioned, GIS include a range of powerful techniques for modelling spatial data and offer an environment in which models and maps of air pollution can be generated. GIS also provide the means to intersect and combine different data sets, making it easy to derive new variables for mapping and modelling (Chapter 3). With this set of tools, it is relatively straightforward to enhance the information from the monitoring network using spatial models that incorporate covariate information as proxy measures of emissions and dispersion. Section 2.3 on the 'state of the art' in air pollution modelling discussed a range of methods presently used for modelling exposures in epidemiological studies. Included among these techniques are geostatistical (Finkelstein et al. 2003; Jerrett et al. 2005; Liu & Rossini 1996) and land use regression methods (Brauer et al. 2003; Briggs et al. 1997; Briggs et al. 2000; Kanaroglou et al. 2005) which take advantage of the GIS toolset.

GIS-based moving window analysis is presented in this chapter as an alternative way in which to model exposures to air pollution. These methods are routinely used to compute digital elevation models for hydrological applications. They are also regularly used to filter spatial data in applications such as remote sensing and image analysis to remove extreme values, in effect smoothing the data to enhance interpretation (Burrough & McDonnell 1998b). Using the GIS functionality of moving window analysis it is also possible to smooth the spread of highly resolved emissions on the basis of meteorological or other data to generate a map of concentrations. One of the most powerful features of this approach is provided by the so-called

FOCALSUM command, available in the ArcInfo GRID environment, which allow for rapid computation of large data sets.

The focal function works by applying calculations within a predefined window, cell-by-cell across an input grid. Focal functions can be used to calculate new values for each cell in a grid using some function of the input value at that location and the values of the surrounding cells specified by the window. As outlined in Section 3.2.1, the window configuration is determined by a text file called a kernel. The kernel specifies the cells, and associated weightings, around the target cell that are to be included in the calculation of each output value.

The focalsum method is built on the principle that proximity to the source is one of the main determinants of pollutant concentrations. To generate a smoothed surface, the window is moved across an input grid to collect (i.e. sum) the source contributions and assign them to the central or target cell. With a weighted kernel one can assign greater contributions from nearby sources. The shape and direction of the kernel can also be defined to mimic the influence of regional wind direction or wind speed. Cells directly downwind of an industrial stack, for example, are expected to receive a greater source contribution than cells adjacent to but upwind of the cell in which the source is located.

Moving window analysis as a technique for mapping air pollution was first explored by Collins (1998a) in a comparison of three GIS-based approaches for modelling traffic-related pollution in Huddersfield, UK. The moving window approach involved the use of a distance-weighted kernel to model the near-source traffic-related component which was then added to a map of background concentrations derived from kriging. Validated against independent monitoring data ($n = 8$), kriging performed only moderately ($R^2 = 0.44$). The moving window approach performed considerably better ($R^2 = 0.67$), comparably to dispersion modelling ($R^2 = 0.63$), though not as well as the regression-based approach ($R^2 = 0.82$).

A similar grid-based approach was explored by Loibl & Orthofer (2001) to model NO_2 concentrations in Austria. They used weighted focalsum equations to apply existing dispersion profiles to point, line and area source emissions to generate a 250m grid of concentrations. The dispersion profiles were used to account for the presence/absence of street canyons, stack height, wind direction and wind speed. In

general the model under predicted the measurement data, especially at monitoring stations with high NO_x contribution from traffic. Validation of the model showed that 54% (72 sites) of the measurements at monitoring sites were within $\pm 15 \mu\text{g}/\text{m}^3$ of the model predictions and 76% (101 sites) were within $\pm 25 \mu\text{g}/\text{m}^3$. The authors suggest that the validation results were reasonable given that measurements at a particular site, which quantifies short-term concentrations at a given location, are not necessarily representative of long-term average spatial patterns in air quality in the surrounding area.

Aside from the work by Collins (1998a) and Loibl & Orthofer (2001), the use of moving window approaches to model air pollution does not appear in the scientific literature. This research attempts to demonstrate and assess the value of moving window techniques as a basis for air pollution modelling, both for emissions and concentrations. Focalsum methods were applied in the APMoSPHERE project to model emissions data, and the work presented in this chapter involves the extension of this approach to model pollution concentrations on basis of these emissions. This chapter aims to demonstrate that the moving window approach offers an efficient method for modelling concentrations at high resolutions over large areas by incorporating the fundamental elements of dispersion.

6.2 Methods

6.2.1 Data requirements

Modelling of concentrations using moving window techniques requires three main sets of data: (1) monitored air pollution at a sample of sites to calibrate and validate the models; (2) emissions (or a proxy for them) to provide measures of source intensity within the moving window; and (3) a distance function by which to weight cells in the source grid according to their relative importance. The last of these might be directional and take account of factors such as wind direction, wind speed or topography. This section describes the way in which these data were sourced, selected and processed prior to analysis.

Emissions data

Emissions can be defined as the quantity of pollutants released into the air from industry, transport and other human activity. The use of emissions maps in the focalsum approach, therefore, provides a measure of this source activity and intensity.

Emissions are usually estimated on the basis of quantities of materials consumed or produced. Emissions inventories and the corresponding emissions maps are typically constructed using a bottom-up or top-down approach. Bottom-up approaches involve collecting and combining detailed data on emissions from individual sources with modelled data based on source activity and predefined emissions factors. Top-down approaches, on the other hand, usually start from national data on consumption and activity (e.g. fuel usage) from which emissions totals are estimated using relevant emissions factors. These are then disaggregated to a finer spatial resolution on the basis of more localised measures of source activity.

A wide range of local and national emissions inventories have been developed in Europe, using one or the other of these approaches (sometimes in combination). Examples include the national 1km emissions inventory in Great Britain (King et al. 2006) and the 5km emissions inventory for the Netherlands (CBS 2005). In London, a 1km emissions inventory has been developed (GLA 2005), and similar inventories are found in many other cities. Although these inventories are often based on best available data in the areas to which they relate, and can thus be considered broadly reliable, for EU-wide mapping they are problematic because of differences in methodology, resolution and emissions categories.

At the international level, one of the most important emissions inventories is that produced by EMEP. Under the 1979 Geneva Convention on Long-Range Transboundary Air Pollution, the European Union is required to report emissions annually for SO₂, NO_x, NMVOC, CH₄, CO, NH₃ and various heavy metals and persistent organic pollutants to EMEP (EEA 2005a). National emissions are reported using eleven main source categories, known as SNAP (Selected Nomenclature for Sources of Air Pollution) categories, which include stationary and mobile combustion sources, industrial processes, fossil fuel extraction, solvent use, waste and agriculture. National estimates for each category are made in a variety of ways, though primarily on the basis of nationally aggregated consumption or production data (e.g. total fuel

consumption by road traffic, energy production, and waste disposal), using internationally agreed emissions factors.

There is also a requirement for the EMEP emissions estimates to be mapped at a 50km grid, which has prompted several European countries to develop GIS protocols for mapping the spatial distribution of emissions. The methods used by Member States to create these maps, however, are not based on an EU-wide protocol which may lead to unrealistic differences, or step changes, across country borders. Given this limitation and the need for higher resolution emissions data, work undertaken in the APMoSPHERE project involved the development of methods (applied to consistent input data) to improve on the EMEP emissions data. This was achieved in part by the use of moving window techniques to disaggregate the national totals to a local (1km) scale.

The APMoSPHERE project emissions modelling protocol (Figure 6.1) involved disaggregation of the national emissions totals (by SNAP category) using a combination of input data on source distribution and activity and other proxies, mainly population density, to distribute the national totals to progressively finer scales. Using this methodology, 1km grids representing 2001 emissions for NO_x, PM₁₀, SO_x and CO were created for each SNAP category outlined in Table 6.1. The SNAP category grids were also reorganised into six classes (C1 to C6) to consolidate the point, line and area sources to provide coherent inputs for air pollution modelling (Table 6.2). It is these six grids listed in Table 6.2 that were used in the focalsum modelling; the limitations which are presented in Table 6.3.

Further description of the emissions modelling protocol used in the APMoSPHERE project is provided on the project website (<http://www.apmosphere.org/>). This work was undertaken by Dr. Chris Dore, Mr. Steve Pye and Ms. Claire Thomson from AEA Technology and Dr. Kees de Hoogh of SAHSU, Imperial College.

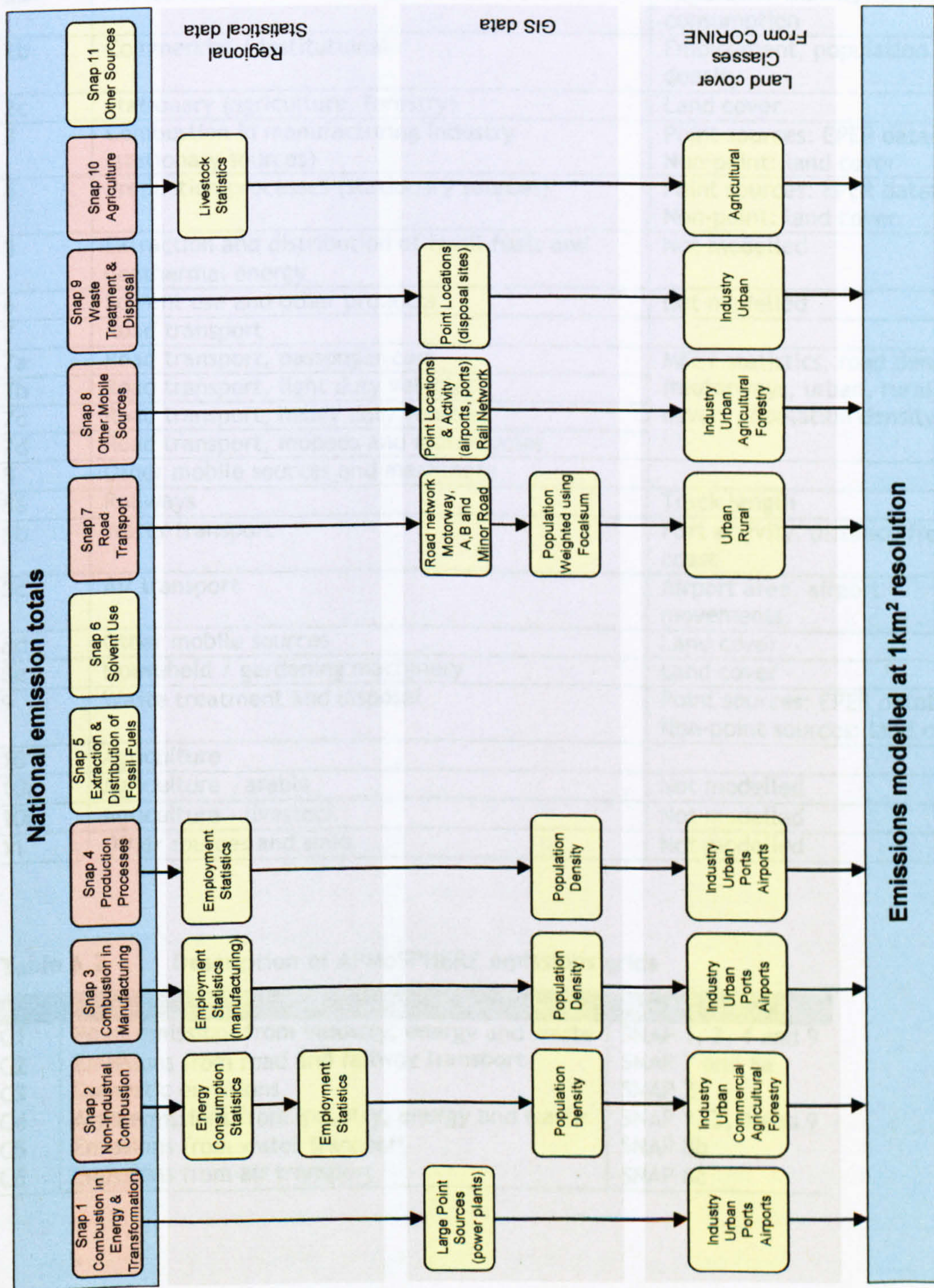


Figure 6.1 Emissions modelling protocol

Table 6.1 Emissions modelling predictor variables by SNAP category

Sector	Sector Description	Predictor variables
1	Combustion in energy and transformation industries (stationary sources)	Point sources: EPER database Land cover
2	Non-industrial combustion plants (stationary sources)	
2a	Residential	Population density, energy consumption
2b	Commercial / Institutional	Employment, population density
2c	Stationary (agriculture, forestry)	Land cover
3	Combustion in manufacturing industry (stationary sources)	Point sources: EPER database Non-point: land cover
4	Production processes (stationary sources)	Point sources: EPER database Non-point: land cover
5	Extraction and distribution of fossil fuels and geothermal energy	Not modelled
6	Solvent use and other products	Not modelled
7	Road transport	
7a	Road transport, passenger cars	MEET statistics, road density (motorways, urban, rural), land cover, population density
7b	Road transport, light duty vehicles	
7c	Road transport, heavy duty vehicles	
7d	Road transport, mopeds and motorcycles	
8	Other mobile sources and machinery	
8a	Railways	Track length
8b	Water transport	Port activity, distance from coast
8c	Air transport	Airport area, airport movements
8d	Other mobile sources	Land cover
8e	Household / gardening machinery	Land cover
9	Waste treatment and disposal	Point sources: EPER database Non-point sources: land cover
10	Agriculture	
10a	Agriculture - arable	Not modelled
10b	Agriculture - livestock	Not modelled
11	Other sources and sinks	Not modelled

Table 6.2 Description of APMoSPHERE emissions grids

Grid	Description	Derived from
C1	Point emissions from industry, energy and waste	SNAP 1, 3, 4 and 9
C2	Emissions from road and railway transport	SNAP 7 and 8a
C3	Domestic emissions	SNAP 2
C4	Area emissions from industry, energy and waste	SNAP 1, 3, 4 and 9
C5	Emissions from water transport	SNAP 8b
C6	Emissions from air transport	SNAP 8c

Table 6.3 Emissions grids, by pollutant, used in focalsum modelling

Modelled Pollutant	Input Grids	Rationale/Limitations
NO ₂	NO _x : C2, C3, C4, C5, C6	The main contribution to background NO ₂ concentration is from traffic-related sources. Industrial point sources (C1) were excluded.
PM ₁₀	PM ₁₀ : C2, C3, C4, C5, C6	The main contribution to background PM ₁₀ concentration is from traffic-related and area sources. Industrial point sources (C1) were excluded. Missing data for point source emissions totals for some SNAP categories meant that all emissions were distributed as area sources for some countries. The PM ₁₀ emissions grids (C1-C6) modelled on the basis of complete data include: Austria, Belgium, Finland, Great Britain and the Netherlands.
SO ₂	SO _x : C1, C2, C3, C4, C5, C6	All emissions categories were used, including industrial point sources which are the main contributor to concentrations.

Wind direction data

Meteorological data, especially wind speed and wind direction, can influence the dispersion of emissions. The higher the wind speed, for example, the further the emissions may be dispersed from a particular source. Wind direction is also important as it can be used to identify areas that are predominantly downwind of an important emission source. Both these elements were, thus, used in focalsum modelling to ‘disperse’ source emissions in a more appropriate form from source to receptor cells.

The wind speed and direction data originate from the 50 x 50km resolution ECMWF database. These data were supplied as monthly mean U (easterly) and V (northerly) wind vector components from which the wind speed and direction were computed using Pythagorean Theorem. Section 3.2.7 describes the methods used to perform these calculations in ArcInfo to derive the seasonal wind speed and direction for every 1 x 1km cell in the study area.

For use in focalsum modelling, the raw data were also used to calculate a 1km grid of wind direction for each month. These monthly grids were then reclassified into the eight directional classes shown in Figure 6.2 to provide a measure of frequency of wind direction (i.e. number of months per year that the wind blows from a particular direction). This was facilitated by the *calcwind_month.aml* in Appendix D1. The eight resulting grids (i.e. one for each direction) were then used to weight wind direction so that the relative importance of upwind sources could be attributed to the target cell during focalsum modelling.

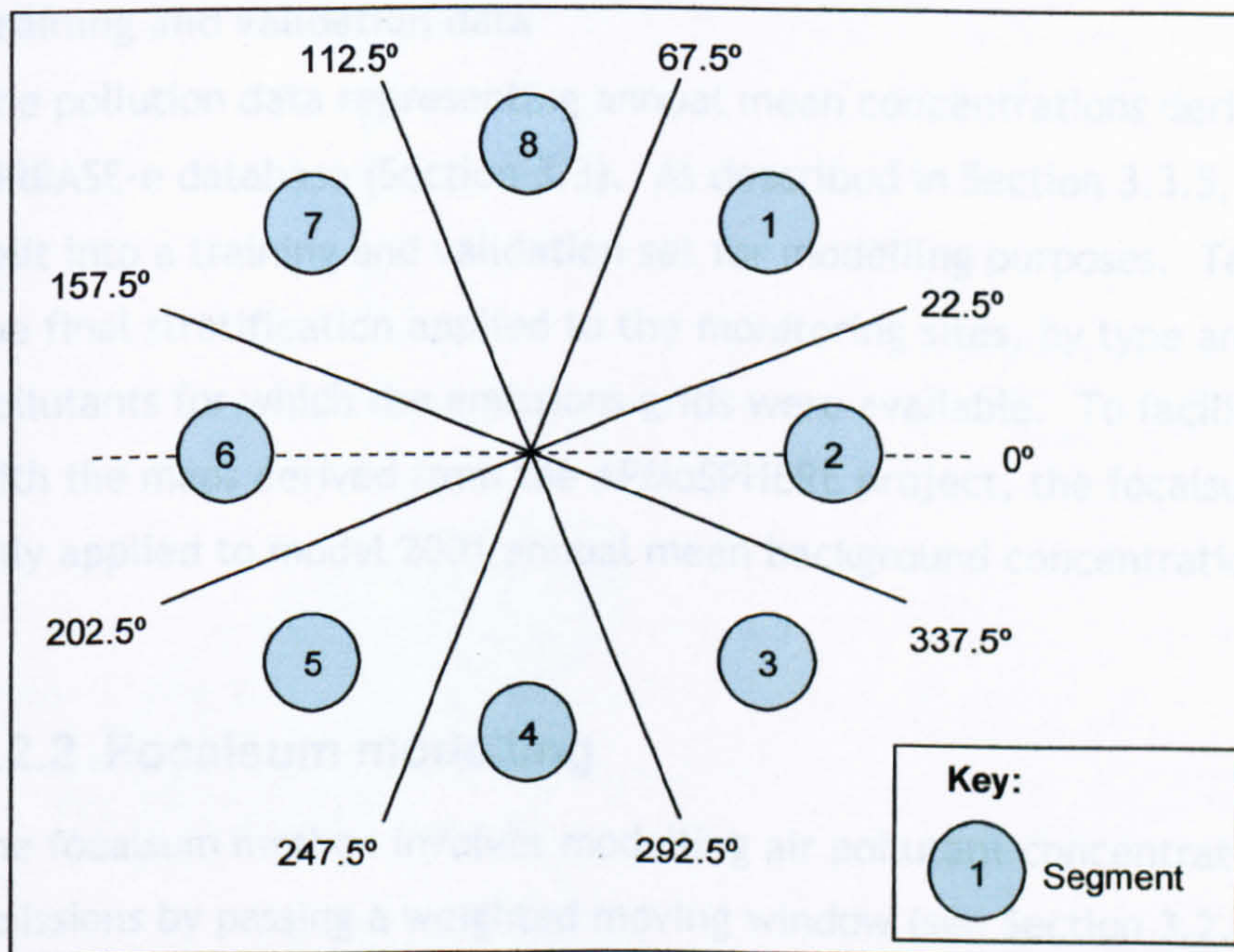


Figure 6.2 Eight segments for wind direction

In incorporating wind direction into the modelling, one must remember that wind direction indicates the direction from which the wind is blowing. In the focalsum method, therefore, the window is positioned over the opposite segment to collect the contributing sources which will be assigned to the target cell positioned at the origin of Figure 6.2. If the wind is in segment 8 (i.e. blowing from the south), for example, the window is positioned over segment 4 to collect the contributing sources.

As already mentioned, there is a relationship between wind speed and distance in that increasing wind speed typically causes emissions to be dispersed further from the source. Wind speed might also alter the shape of the distance decay function. To explore these relationships the wind speed grid was reclassified into three grids representing light air (< 1 m/s), light breeze (≥ 1 m/s and ≤ 2.5 m/s) and breeze (> 2.5 m/s) according to the Beaufort Scale (MET Office 2006). Modifications to the kernel shape (e.g. circular for low wind speeds and elongated for higher wind speeds) and distance decay functions (e.g. slower decay with high wind speeds) could then be examined in relation to the predominant wind speed.

Training and validation data

The pollution data representing annual mean concentrations derived from the AIRBASE-e database (Section 3.3). As described in Section 3.3.5, these data were split into a training and validation set for modelling purposes. Table 3.10 outlines the final stratification applied to the monitoring sites, by type and location, for the pollutants for which the emissions grids were available. To facilitate comparison with the maps derived from the APMoSPHERE project, the focalsum methods were only applied to model 2001 annual mean background concentrations.

6.2.2 Focalsum modelling

The focalsum method involves modelling air pollutant concentrations on the basis of emissions by passing a weighted moving window (see Section 3.2.1) cell-by-cell across an emissions grid to collect source contributions. Distance decay functions are applied to the window so that the contribution of emissions closer to the target cell are more heavily weighted than those that are further away. The emissions value at each cell inside the window is multiplied by its respective weight, and the cell values inside the window are summed and assigned to the target cell. This process produces an index grid which can then be calibrated with the concentration data from the training sites. Using the calibration equation, the index grid (based on emissions) is converted to a map of concentrations which is then validated with concentration data from the reserved sites. A flow diagram for the basic focalsum procedure is given in Figure 6.3, and modifications to the basic procedure are described below.

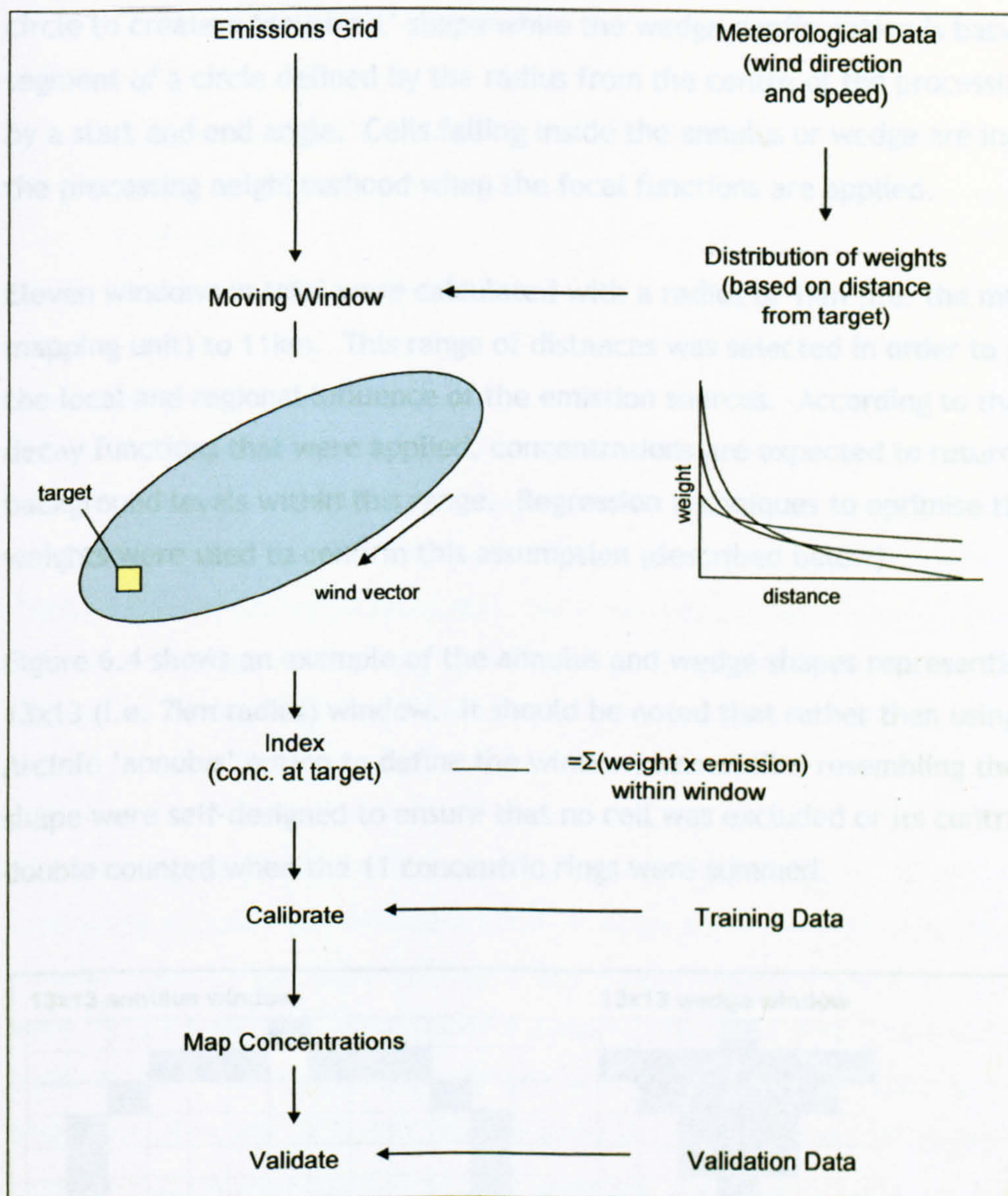


Figure 6.3 Basic focalsum procedure

The focalsum methodology was developed at the European level using the 2001 emissions grids and NO₂ measurements from the AIRBASE-e monitoring network. Models incorporating wind data were also explored. The models were calibrated using the monitoring sites designated for training and validated with the reserved sites. After refining the methodology for NO₂, the focalsum method was extended, where possible, to other pollutants and geographies for which reliable emissions grids were available.

Window configurations

Two main window configurations were used in focalsum modelling: annulus and wedge (Figure 6.4). The annulus shape comprises a smaller circle within a larger

circle to create a 'doughnut' shape while the wedge configuration is basically a segment of a circle defined by the radius from the centre of the processing cell and by a start and end angle. Cells falling inside the annulus or wedge are included in the processing neighbourhood when the focal functions are applied.

Eleven windows in total were calculated with a radius of 1km (i.e. the minimum mapping unit) to 11km. This range of distances was selected in order to pick up both the local and regional influence of the emission sources. According to the distance decay functions that were applied, concentrations are expected to return to background levels within this range. Regression techniques to optimise the distance weights were used to confirm this assumption (described below).

Figure 6.4 shows an example of the annulus and wedge shapes representing the 13x13 (i.e. 7km radius) window. It should be noted that rather than using the ArcInfo 'annulus' option to define the window, kernel files resembling the annulus shape were self-designed to ensure that no cell was excluded or its contribution double counted when the 11 concentric rings were summed.

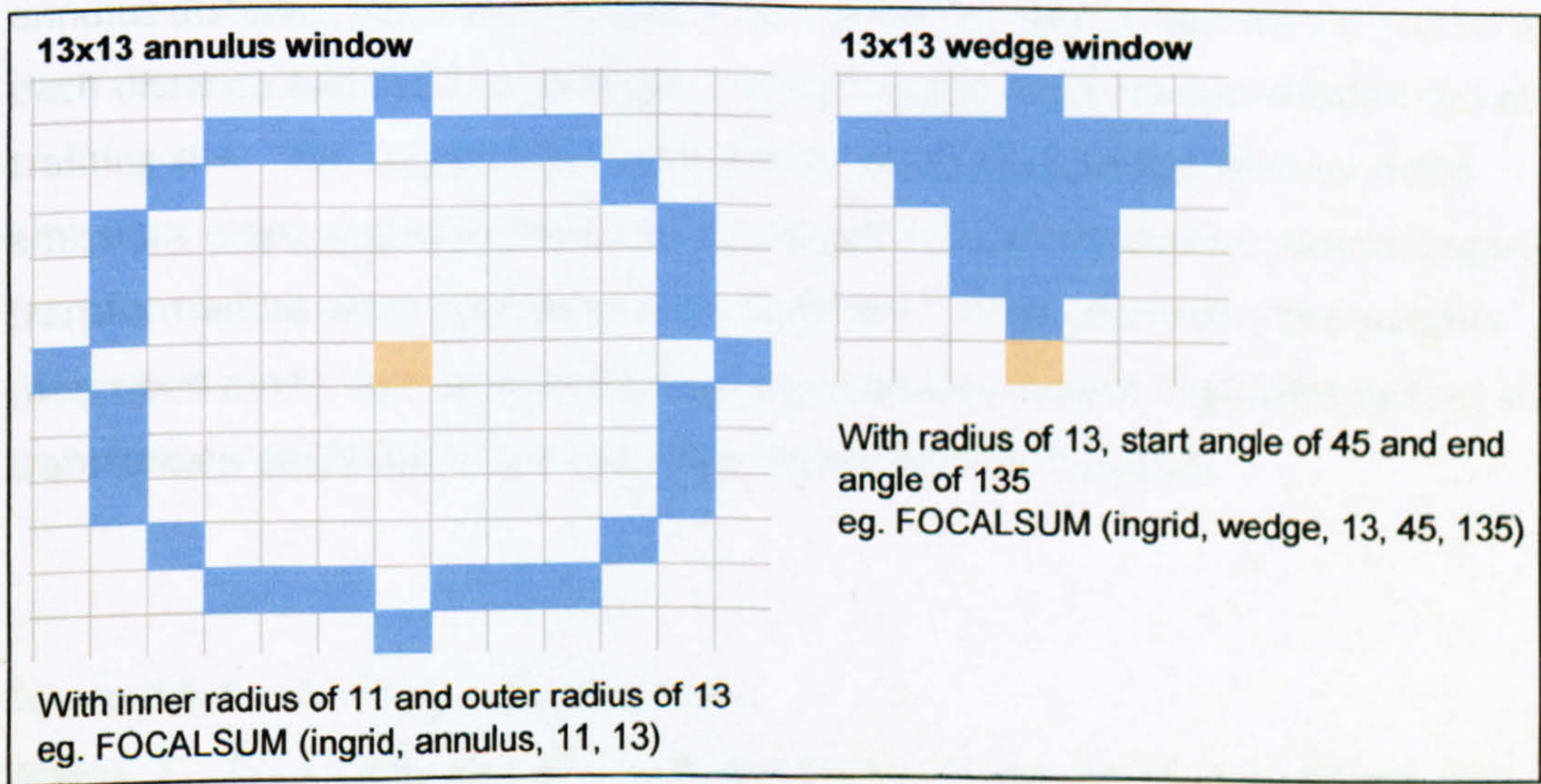


Figure 6.4 Annulus and wedge windows for 13x13 neighbourhood

Progressively complex focalsum models were run, as follows:

- **Analysis 1:** Annulus windows as 11 concentric rings weighted by $1/d^2$, where d is distance.

- **Analysis 2:** Annulus windows as 11 concentric rings, with optimised distance weights.
- **Analysis 3:** Eight wedge windows weighted by monthly frequency of wind direction (as per Figure 6.2), to a radius of 11km weighted by some function of distance.

Each of these analyses is described below.

Analysis 1 and 2: Annulus windows

Focalsum modelling with the annulus windows was conducted in several stages. Using the *focalsum_annulus.aml* (Appendix D2) a series of eleven annulus kernels of increasing radius (i.e. 1-11km) were passed over the relevant emissions grid (see Table 6.3 for specifics by pollutant). As each window was moved across an emissions grid, the contribution (i.e. the sum of emissions within each annulus ring) was attributed to the target cell in an output grid. Eleven resulting grids, one from each annulus kernel, were thus generated.

The training sites were then intersected with the eleven resulting grids. A file containing annual mean pollutant concentration and sum of emissions for each annulus distance was imported into SPSS. A set of weights was then imputed for each distance and used to compute a distance-weighted emissions index (I_c) at each training site. The relationship between annual mean concentration and the emissions index was examined and, to control for non-normality, natural logarithm transformations were applied to both variables. After optimising the weights (described next), the log-transformed concentrations were regressed against the log-transformed emissions index to derive the calibration function.

$$\text{Equation 6.1} \quad I_c = aE_1 + bE_2 + \dots + nE_{11}$$

Where: $E_1 \dots E_{11}$ are the sum of emissions for the 1st through to the 11th annulus window; and $a \dots n$ are the imputed weights.

For Analysis 1, the weight for the inner window (E_1) was set to 1 and all other weights were set to $1/d^2$ (i.e. 0.25, 0.111... 0.0083).

For Analysis 2, the weights were adjusted manually (starting from $1/d^2$), in a stepwise manner. Weights were adjusted according to two constraints: the

maximum weight was set to 1 for E_1 ; and each outer ring had to have a weight less than, or equal to, the adjacent inner ring (i.e. to portray the distance decay relationship of the dispersion environment, in that concentrations decrease with distance from the source). Weights were adjusted in a sequential, stepwise fashion from E_2 to E_{11} then from E_{11} back to E_2 , etc. until the optimal solution was found. Initially the values from the $1/d^2$ distribution were used to adjust the weight for a given distance up or down by one unit (e.g. E_2 was first adjusted to 1, then to 0.25). In later stages of the process deviations from the $1/d^2$ distribution were explored, using smaller increments. The index of emissions was re-calculated after each adjustment and linear regression analysis was used (in SPSS) to calibrate the log-transformed emissions index to a measure of log-transformed concentrations at the training sites. The optimal set of weights was selected based on maximisation of R^2 .

Analysis 3: Wedge window incorporating weighted wind direction

The *focalsum_wwind.aml* (Appendix D2) was used to calculate a single emissions index grid which summed two sets of weights for each target cell. First, a circular window was constructed from 8 wedges (Figure 6.2), each weighted by the monthly frequency of wind direction. Next, a function was applied to each concentric ring of the circular window (from a 1km to 11km radius) to distance-weight the rings. Initially the function $1/d^2$ was computed. A follow-up sensitivity analysis was also performed to evaluate the effect of other distance weightings including $1/d$, $1/\ln(d)$ and $1/d^3$.

As per Analysis 1 and 2, the training sites were then intersected with the index grid and imported into SPSS. Natural logarithm transformations were applied to the emissions index and annual mean pollutant concentrations, and the log-transformed concentrations regressed against the log-transformed emissions index to derive the calibration function.

Modelling other pollutants and spatial scales

After refining the methodology on NO_2 , the focalsum method was¹ applied to PM_{10} . Modelling PM_{10} , however, was limited by the quality of the emissions grids and increased sparseness of the monitoring sites. As outlined in Table 6.3, many of the PM_{10} SNAP category grids were compromised by missing data. Models limited to the countries with emissions grids developed on the basis of complete data (i.e. Austria,

Belgium, Finland, Great Britain and the Netherlands) were also investigated. EU-wide focalsum modelling was also explored for SO₂.

In follow-up analyses, focalsum modelling was also applied to nationally available data for Great Britain to assess whether a denser network of monitoring sites and/or the nationally reported emissions data could improve model performance. National emissions grids for NO_x, PM₁₀ and SO_x were available from the National Atmospheric Emissions Inventory (NAEI) maintained by the National Environmental Technology Centre (Netcen). Additional NO₂ (n= 587 background sites) and SO₂ (n = 135) concentration data were obtained from the UK National Air Quality Information Archive non-automatic network. Similar to AIRBASE-e (Section 3.3.5 and 6.2.1), the monitoring sites were randomly stratified into two groups: 75% for training and 25% for validation.

As described in Section 3.3.1, the monitoring sites included in EUROAIRNET (i.e. reported in AIRBASE) derive from existing networks around the EU. Data quality objectives, such as the requirement to use the reference method (or equivalent) for monitoring are defined to ensure that the measurements are comparable. The UK network, however, differs not only because the sites are non-automatic but also in how the sites record measurements.

The NO₂ monitors in the UK network comprise diffusion tubes which measure concentrations by passive sampling. Passive samplers rely on the gas diffusing to an absorber without pumping. Because these monitors are inexpensive and do not require electrical power they are useful for large-area surveys to characterise spatial variations in pollution. The use of diffusion tubes in some applications, however, is limited given that they are labour and laboratory intensive. Additionally, the measurements are not directly comparable with standards as each tube is exposed for days to weeks, providing a measure of average concentrations (Colls 2002). The diffusion tubes in the UK network, for example, are exposed for a period of one month (Netcen 2006).

In the UK network SO₂ is absorbed in hydrogen peroxide and the concentration is calculated by acid titration. Comparison of the measurements recorded by the UK network with the AIRBASE-e sites is less problematic as the measurements are taken daily (Netcen 2006).

6.2.3 Mapping concentrations

To facilitate mapping, the 1km index grid was created in ArcInfo by applying Equation 6.1 (appropriately weighted) to the 11 annulus grids. The natural logarithm of the index grid was taken, and the appropriate calibration function applied. The resulting grid was then exponentiated, thus, producing the 1km grid of pollutant concentrations. This process was done using the *create_map* routine in the *focalsum_annulus.aml* (for Analysis 1 and 2) and *focalsum_wwind.aml* (for Analysis 3). These AMLs are presented in Appendix D2.

6.2.4 Model validation

As described in Section 5.4.2, measures of performance were calculated in order to assess the predictive capability of individual models and to compare them with other modelling methods. The goodness-of-fit measures calculated on the basis of the reserved validation sites included: R^2 , standard error (SEE), root mean square error (RMSE), fractional bias (FB) and factor-of-two (F2).

Jackknife, or leave-one-out (LOO), is a technique commonly used to assess the robustness of a model when there is insufficient data for validation with an independent subset. This technique was used to assess model performance when fewer than 100 total monitoring sites (i.e. training plus validation sites) were available. It was only necessary to use jackknife for models built at the GB level on the basis of the GB sites from the European monitoring network (i.e. AIRBASE-e sites). Jackknife involves excluding one monitoring site at a time and predicting its concentration from the remaining sites (in this case, using linear regression). The process is repeated until all sites have been left out, and thus predicted once. The model is then evaluated using linear regression to examine the relationship between the measured concentration and the jackknife prediction. The RMSE, F2 and FB can also be calculated.

6.2.5 Comparison with reference method

The focalsum results were compared to the best models derived from the APMoSPHERE project. For NO_2 , PM_{10} and SO_2 a method which combined stochastic and geostatistical modelling, undertaken by Utrecht University participants (Dr. Gerard Hoek, Dr. Edzer Pebesma and Rob Beelen), outperformed the other models

explored in the APMoSPHERE project. This was, thus, defined as the reference method against which the focalsum models developed in this chapter were compared.

The overall modelling strategy undertaken in the APMoSPHERE project was developed on the principle that air pollution patterns represent the combined effect of global trend, reflecting mainly broad-scale meteorological and topographic factors, and a local component reflecting nearby emissions sources and dispersion characteristics. Separate analyses, therefore, were carried out for the global background, rural and urban scale (with urban and rural cells defined by land cover - described in Section 3.3.3).

Prior to modelling, the pollutant concentrations were transformed to the natural logarithmic scale. Models were developed using the training data. Regression models using a supervised, forward stepwise procedure were developed in SAS using the list of potential predictor variables in Table 6.4. The regression models were constrained to variables conforming to the expected direction of association (indicated by the + and - signs in the table) that increased the adjusted R^2 by more than one percent.

The global background models were constructed first using the rural background training sites and global background predictors. The rural predictors were then added to the preferred global background model to produce the rural scale model. The urban scale was modelled separately using the urban background training sites by first computing the residuals from the preferred global model at the urban background sites. The residuals were then entered into a separate regression analysis as the dependent variable with the urban scale predictors as the independent variables. After each stage of modelling (global background, rural and urban) the 'best' model was selected based on an evaluation of the model performance. This was determined by comparing predictions with the monitored concentrations at the relevant validation sites according to the measures outlined in Section 6.2.4 (i.e. R^2 , SEE, RMSE, F2 and FB).

The gstat package within the R statistical software environment was used for ordinary kriging or ordinary co-kriging, as appropriate. If, for example, covariates were identified during the regression analysis, they were included in ordinary co-

kriging; otherwise ordinary kriging was used. First, the variograms were constructed and different kriging models (e.g. spherical, Gaussian, exponential and linear) were compared with different cuts-off and boundaries. The preferred rural and urban models were then used to predict concentrations on a global neighbourhood using all available measurements within relevant (rural/urban) grid cells. A composite of the two maps was then created to produce an EU-wide concentration map for each pollutant. The final models for the reference method are presented in Tables 6.5 to 6.7.

Table 6.4 Variables used in APMoSPHERE pollution modelling

Variable	Global Background (1km)			Rural scale (5km, 21km)			Urban scale (1km)		
	NO ₂	PM ₁₀	SO ₂	NO ₂	PM ₁₀	SO ₂	NO ₂	PM ₁₀	SO ₂
Motorways				+	+	+	+	+	+
A-roads									
B-roads									
Minor roads									
Railways									
High density residential				+	+	+	+	+	+
Low density residential									
Industry									
Transport									
Airports				+	+	+	+	+	+
Seaports									
Construction									
Urban greenery									
Forestry							-	-	-
Agriculture					+				
Altitude [#]	-	-	-				-	-	+
Topex [#]				-	-	-	-	-	-
Distance to sea [#]	+	-	+						
Combined into 5 meteorological factors									
Seasonal temperature	+	+	+						
Annual radiation	+	+	+						
Seasonal wind speed	+	+	+						
Seasonal calm [*]	+	+	+						
Annual vapour pressure	+	+	+						
Annual rainfall	+	+	+						

Notes: + = positive direction; - = negative direction

Altitude is calculated as $\sqrt{\text{nalt}/\text{max}(\text{nalt})}$, where $\text{nalt} = \text{altitude} - \text{min}(\text{altitude})$

Topex is calculated as $\sqrt{\text{ntopex}/\text{max}(\text{ntopex})}$, where $\text{ntopex} = \text{topex} - \text{min}(\text{topex})$

Distance to sea is calculated as $\sqrt{\text{minimum distance}/\text{max}(\text{minimum distance})}$

Table 6.5 Reference method - NO₂ background model

Variable	Parameter estimate	Standard error
Global Model = Ordinary Co-Kriging (R ² = 33.5%, RMSE = 8.006)		
Intercept	2.590	0.107
Altitude	-2.305	0.262
Distance to sea	0.946	0.178
Climate factor 5: cold calm winters, warm windy summers	0.679	0.114
Rural Model = Ordinary Co-Kriging (R ² = 47.0%, RMSE = 7.235)		
Intercept	3.987	0.711
Altitude	-1.592	0.244
Distance to sea	0.705	0.179
Climate factor 5	0.518	0.102
Minor roads	0.010	0.003
Major roads	0.021	0.008
Non-residential urban	0.053	0.016
Topex	-2.753	1.062
High density residential	0.041	0.017
Urban Model = Ordinary Co-Kriging (R ² = 50.7%, RMSE = 6.389)		
Intercept	0.231	0.034
High density residential	0.004	0.001
Major road	0.073	0.015
Minor road	0.031	0.009

Table 6.6 Reference method - PM₁₀ background model

Variable	Parameter estimate	Standard error
Global Model = Ordinary Co-Kriging (R ² = 24.6%, RMSE = 8.607)		
Intercept	3.204	0.098
Climate factor 1: hot year round, windy	0.051	0.039
Climate factor 2: hot summers, cooler winter, calm	0.096	0.027
Climate factor 5: cold calm winters, warm windy summers	0.327	0.110
Altitude	-0.751	0.266
Rural Model = Ordinary Co-Kriging (R ² = 36.5%, RMSE = 7.747)		
Intercept	4.148	0.725
Climate factor 1: hot year round, windy	0.054	0.038
Climate factor 2: hot summers, cooler winter, calm	0.069	0.027
Climate factor 5: cold calm winters, warm windy summers	0.294	0.106
Altitude	-0.313	0.270
Agriculture	0.004	0.001
Low density residential	0.009	0.003
Topex	-2.013	1.075
High density residential	0.018	0.010
Urban Model = Ordinary Kriging (R ² = 65.7%, RMSE = 3.323)		

Model building

Table 6.8 presents the calibration function and associated regression statistics for each NO₂ model. The optimised weights for models built using Analysis 2 are also given. For each group of models, the best calibration function was achieved from Analysis 2 in which the distance weights were optimised. This is indicated by the increase in R^2 and decrease in SEE for Analysis 2 in each group. Nevertheless, compared to the corresponding Analysis 1 run (based on $1/d^2$), optimisation of the weights resulted in only a small increase in the percent agreement between the log-transformed concentrations and log-transformed emissions index at the training sites. The optimised weights derived for run 2, for example, achieved only a 0.8% increase. The largest increases in R^2 were noted for runs 8 and 11 which are also reflected in the larger increase in percent agreement at the validation sites shown in Table 6.9.

The calibration functions derived for runs 3 to 6, representing EU models incorporating monthly weighted wind direction in combination with a distance decay function, all had a reduced R^2 compared to Analysis 2 (run 2). Of these models, however, the best calibration was achieved using the $1/d^2$ function applied in run 3. The annulus model (run 2) shows that the optimal decay function is much more gradual than $1/d^2$ for the inner few annulus rings (to a distance of about 3km). This indicates that there is a function different to $1/d^2$ which could be applied to the monthly weighted wind direction model to produce an even better calibration function.

The GB models, which were based on different combinations of national and EU-level concentration and emissions data, all had very different optimal distance decay weights. The weights for run 11 resembled the $1/d^2$ function while run 8 showed no decay until a distance beyond 6km. The weights for run 14 were approximately midway between those of runs 8 and 11. Also, the weights for run 8 did not decay to approximately zero (i.e. background level), thus, indicating that the annulus rings should have been expanded beyond 11km for this analysis only.

Table 6.7 Reference method - SO₂ background model

Variable	Parameter estimate	Standard error
Global Model = Ordinary Kriging (R ² = 5.9%, RMSE = 5.645)		
Rural Model = Ordinary Co-Kriging (R ² = 12.3%, RMSE = 5.458)		
Intercept	3.937	1.220
Non-residential urban	0.075	0.019
Topex	-3.970	1.817
Urban Model = Ordinary Kriging (R ² = 31.9%, RMSE = 3.088)		

6.3 Results

Focalsum models for annual mean concentrations were built and validated for NO₂, PM₁₀ and SO₂. The success in modelling at the scale of the EU was doubtful for some pollutants, largely due to the reduced number of monitoring sites (Table 3.10) and lesser quality emissions grids in the case of PM₁₀ (Table 6.3). For this reason, the GB models were also developed to assess the performance of the method where national data could provide a denser network of monitoring sites and high quality emissions grids. This section describes the model building and validation results for the focalsum approach.

6.3.1 NO₂ models

Results from the model building for NO₂ are presented in Table 6.8 while the validation results are provided in Table 6.9. Each analysis is assigned a unique ‘run’ number, and the analyses are also grouped for comparison. The groups comprise Analysis 1 to 3 applied to the same set of input data. The EU models were built using the AIRBASE-e monitoring network and the EU-wide emissions grids. Three combinations of input data were available for the GB analysis:

- A. AIRBASE-e sites and the EU-wide emissions grids (runs 7-9)
- B. AIRBASE-e sites and the national NAEI emissions grids (runs 10-12)
- C. National network sites and the national NAEI emissions grids (runs 13-15)

Table 6.8 NO₂ model building
Including distance weights applied to annulus rings

Run/ Analysis	Method	Calibration				Distance Weights for Each Annulus Ring										
		N	Function	R ²	SEE	1	2	3	4	5	6	7	8	9	10	11
EU-wide FOCALSUM models																
1/1	FS annulus (1/d ²)	714	exp ^{1.912 + 0.254 lnlc}	0.63	0.372	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
2/2	FS annulus ow	714	exp ^{1.801 + 0.241 lnlc}	0.64	0.368	1.0	1.0	0.5	0.028	0.028	0.01	0.01	0.01	0.01	0.01	0.01
3/3	FS wind (1/d ²)	714	exp ^{2.232 + 0.234 lnlc}	0.63	0.371											
4/3	FS wind (1/d)	714	exp ^{1.970 + 0.250 lnlc}	0.60	0.386											
5/3	FS wind (1/ln(d))	714	exp ^{1.773 + 0.243 lnlc}	0.56	0.406											
6/3	FS wind (1/d ³)	714	exp ^{2.397 + 0.210 lnlc}	0.62	0.375											
GB-wide FOCALSUM models																
7/1	FS annulus (1/d ²)	62	exp ^{2.399 + 0.178 lnlc}	0.54	0.228	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
8/2	FS annulus ow	62	exp ^{2.432 + 0.197 lnlc}	0.58	0.216	1.0	1.0	1.0	1.0	1.0	1.0	0.557	0.557	0.439	0.439	0.439
9/3	FS wind (1/d ²)	62	exp ^{2.787 + 0.134 lnlc}	0.45	0.248											
10/1	FS annulus (1/d ²)	62	exp ^{2.112 + 0.244 lnlc}	0.54	0.227	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
11/2	FS annulus ow	62	exp ^{2.036 + 0.265 lnlc}	0.58	0.217	1.0	0.167	0.069	0.069	0.017	0.017	0.017	0.017	0.017	0.017	0.017
12/3	FS wind (1/d ²)	62	exp ^{2.544 + 0.199 lnlc}	0.52	0.232											
13/1	FS annulus (1/d ²)	441	exp ^{1.645 + 0.284 lnlc}	0.33	0.281	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
14/3	FS annulus ow	441	exp ^{1.736 + 0.259 lnlc}	0.35	0.277	1.0	1.0	0.444	0.444	0.16	0.16	0.16	0.16	0.062	0.062	0.062
15/3	FS wind (1/d ²)	441	exp ^{2.011 + 0.255 lnlc}	0.25	0.298											

Notes: ow = optimised weighted; calibration based on log-transformed emissions index and log-transformed concentrations

Table 6.9 Background NO₂ modelling results

Run/ Analysis	Method	Network	Emissions	Total N	Validation Results						
					R ²	Adj R ²	SEE	RMSE	F2	FB	Method
Reference model from APMoSPHERE (EU-wide)											
	Reference	AIRBASE-e		942	0.61	0.61	6.580	6.725	94.7%	-0.039	LS (N=228)
EU-wide FOCALSUM models											
1/1	FS annulus (1/d ²)	AIRBASE-e	2,3,4,5,6	942	0.60	0.60	6.719	6.803	90.8%	0.005	LS (N=228)
2/2	FS annulus ow	AIRBASE-e	2,3,4,5,6	942	0.61	0.61	6.641	6.673	93.0%	-0.028	LS (N=228)
3/3	FS wind (1/d ²)	AIRBASE-e	2,3,4,5,6	942	0.61	0.61	6.588	6.617	91.7%	-0.009	LS (N=228)
4/3	FS wind (1/d)	AIRBASE-e	2,3,4,5,6	942	0.58	0.58	6.882	7.008	90.4%	-0.012	LS (N=228)
5/3	FS wind (1/ln(d))	AIRBASE-e	2,3,4,5,6	942	0.54	0.54	7.200	7.373	89.9%	-0.018	LS (N=228)
6/3	FS wind (1/d ³)	AIRBASE-e	2,3,4,5,6	942	0.60	0.60	6.657	6.652	92.5%	-0.010	LS (N=228)
GB-wide FOCALSUM models											
7/1	FS annulus (1/d ²)	AIRBASE-e	2,3,4,5,6	62	0.52	0.51	6.805	6.706	98.4%	-0.002	LOO (N=62)
8/2	FS annulus ow	AIRBASE-e	2,3,4,5,6	62	0.56	0.55	6.543	6.440	98.4%	-0.001	LOO (N=62)
9/3	FS wind (1/d ²)	AIRBASE-e	2,3,4,5,6	62	0.42	0.41	7.504	7.413	96.8%	-0.003	LOO (N=62)
10/1	FS annulus (1/d ²)	AIRBASE-e	NAEI at00	62	0.53	0.52	6.753	6.670	98.0%	-0.002	LOO (N=62)
11/2	FS annulus ow	AIRBASE-e	NAEI at00	62	0.57	0.56	6.464	6.375	98.0%	-0.002	LOO (N=62)
12/3	FS wind (1/d ²)	AIRBASE-e	NAEI at00	62	0.46	0.45	7.220	7.128	98.0%	-0.002	LOO (N=62)
13/1	FS annulus (1/d ²)	UK	NAEI at00	587	0.43	0.42	5.731	5.844	97.3%	-0.054	LS (N=146)
14/2	FS annulus ow	UK	NAEI at00	587	0.44	0.44	5.653	5.771	97.3%	-0.055	LS (N=146)
15/3	FS wind (1/d ²)	UK	NAEI at00	587	0.30	0.30	6.328	6.474	96.6%	-0.069	LS (N=146)

Notes: ow = optimised weighted; LS = least squares regression; LOO = leave-one-out (jackknife)

Notes: ow = optimised weighted; LS = least squares regression; LOO = leave-one-out (jackknife)

EU models

As the model complexity was increased from Analysis 1 to 3 (runs 1 to 3 in Table 6.9), the performance also increased from an R^2 of 0.58 to 0.61. While Table 6.8 showed that Analysis 2 produced the best calibration function, the best validated model was actually achieved in Analysis 3 (run 3) where monthly weighted wind direction in combination with a $1/d^2$ distance decay function were applied ($R^2 = 0.61$; $SEE = 6.588 \mu\text{g}/\text{m}^3$). The performance of run 3 was equivalent to that of the reference method ($R^2 = 0.61$; $SEE = 6.580 \mu\text{g}/\text{m}^3$). The other measures of performance were also very similar. Only slightly more of the predictions from the reference method were within a factor-of-two (95% compared to 92% for run 3) while the focalsum model had a slightly lower measures of fractional bias. The negative fractional bias values indicate that both models tended to under-predict concentrations by less than 0.4%.

Runs 4 to 6 in Table 6.9 represent results of the sensitivity analysis undertaken to evaluate the effect of changing the distance decay functions for Analysis 3. None of the functions examined here ($1/d$, $1/\ln(d)$ and $1/d^3$) performed as well as run 3, which suggests that $1/d^2$ was the best tested distance decay function when monthly weighted wind direction was incorporated into the model. As indicated by the calibration results, however, it is likely that a more suitable function would be one exhibiting a more gradual decay closer to the source (i.e. similar to the optimised weights from run 2).

The agreement between the predicted and measured concentrations (for run 3) at the validation sites was also examined by site type to determine if the reference method and best focalsum model performed equally in urban versus rural areas. As shown in the scatterplots in Figure 6.5, both of the methods validated better at the urban sites ($n = 151$) than at the rural sites ($n = 77$). At the urban sites, the focalsum method ($R^2 = 0.50$) marginally outperformed the reference method ($R^2 = 0.49$), while the opposite was true for the rural sites (i.e. R^2 of 0.44 for the reference method compared to 0.41 for focalsum). The small differences in performance further suggest that NO_2 model derived from focalsum is comparable to that of the reference method.

Maps of the reference method and best focalsum model (run 3) are shown in Figures 6.6 and 6.7, respectively. While both methods reflected expected trends in annual NO_2 concentrations (i.e. higher concentrations in urban centres, southeast England,

Belgium, the Netherlands, the Ruhr area in Germany and the Po Valley), the reference model produced a somewhat smoothed map compared to focalsum. The emissions from road transport, for example, are more evident in the focalsum map where the pattern of major roads is clearly visible across the study area. Additionally, the area of elevated NO₂ levels shown in the map inset exhibits less variation in the reference map - especially noteworthy is the absence of lower level concentrations which are apparent in the focalsum map.

A further comparison of the maps was undertaken by examining statistics calculated on the basis of the 1km predictions across the whole study area. The concentration data upon which the models were built had a range of 1.0-74.0 µg/m³ (SD = 11.64 µg/m³). The range of predicted values in the focalsum map were comparable (minimum = 0.02 µg/m³; maximum = 62.9 µg/m³; mean = 9.5 µg/m³; SD = 5.1 µg/m³), but the reference method produced unreasonably high predictions in a small proportion of cells (minimum = 1.0 µg/m³; maximum = 424.5 µg/m³; mean = 10.1 µg/m³; SD = 5.9 µg/m³). Approximately 77 grid cells (located in the Netherlands, Paris, Barcelona, Seville, Turin, Rome, Naples and Athens) had predicted concentrations exceeding 100 µg/m³ with the grid cell containing the maximum predicted concentration of 424.5 µg/m³ located in Paris. Although these cells represent a very tiny proportion of outliers, they indicate that reference method is capable of producing erroneously large predictions.

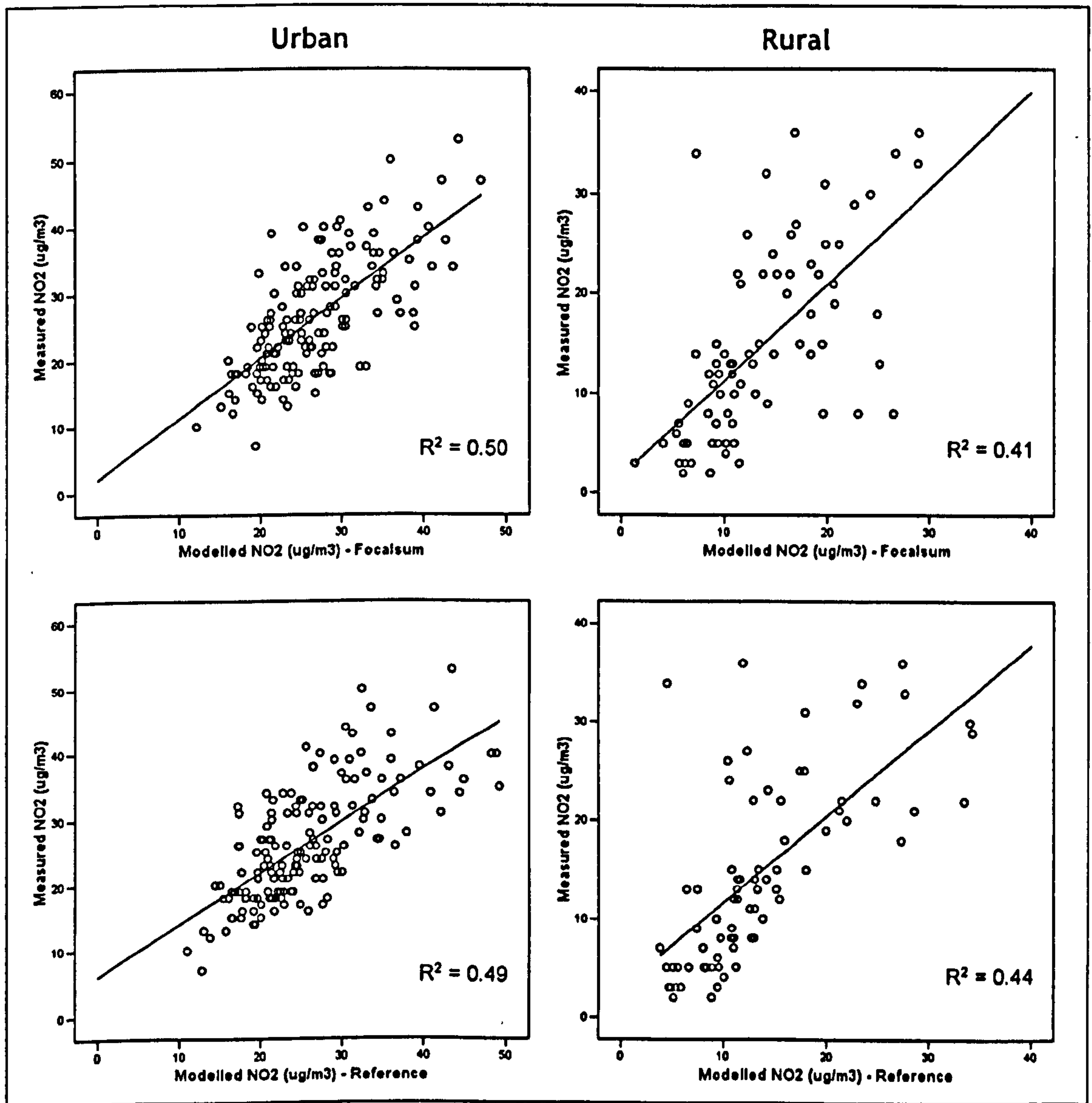


Figure 6.5 Validation stratified by urban and rural sites
Showing best focalsum model (Run 3, Analysis 3)

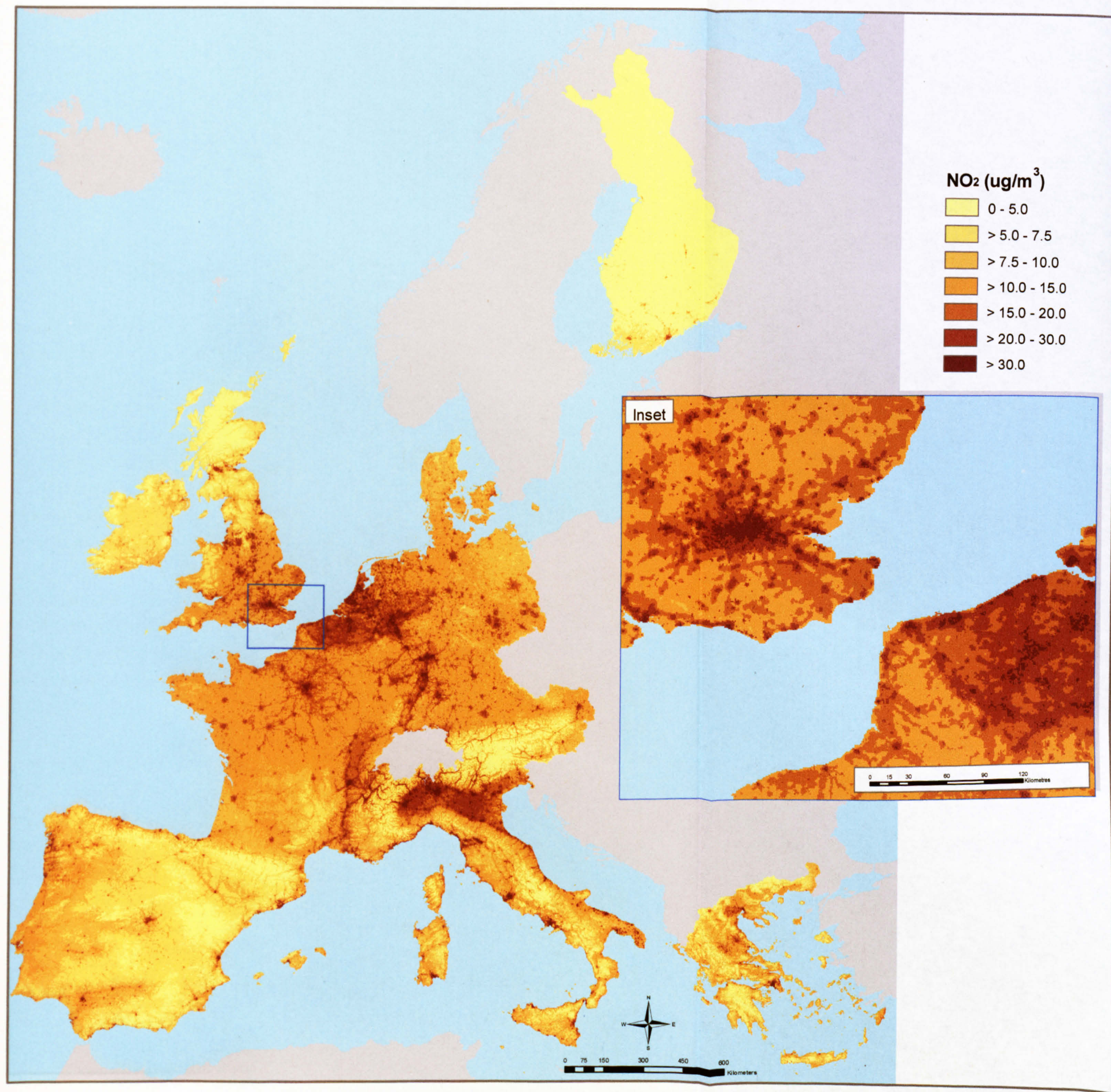


Figure 6.6
Annual mean background NO₂ (µg/m³)
Reference method

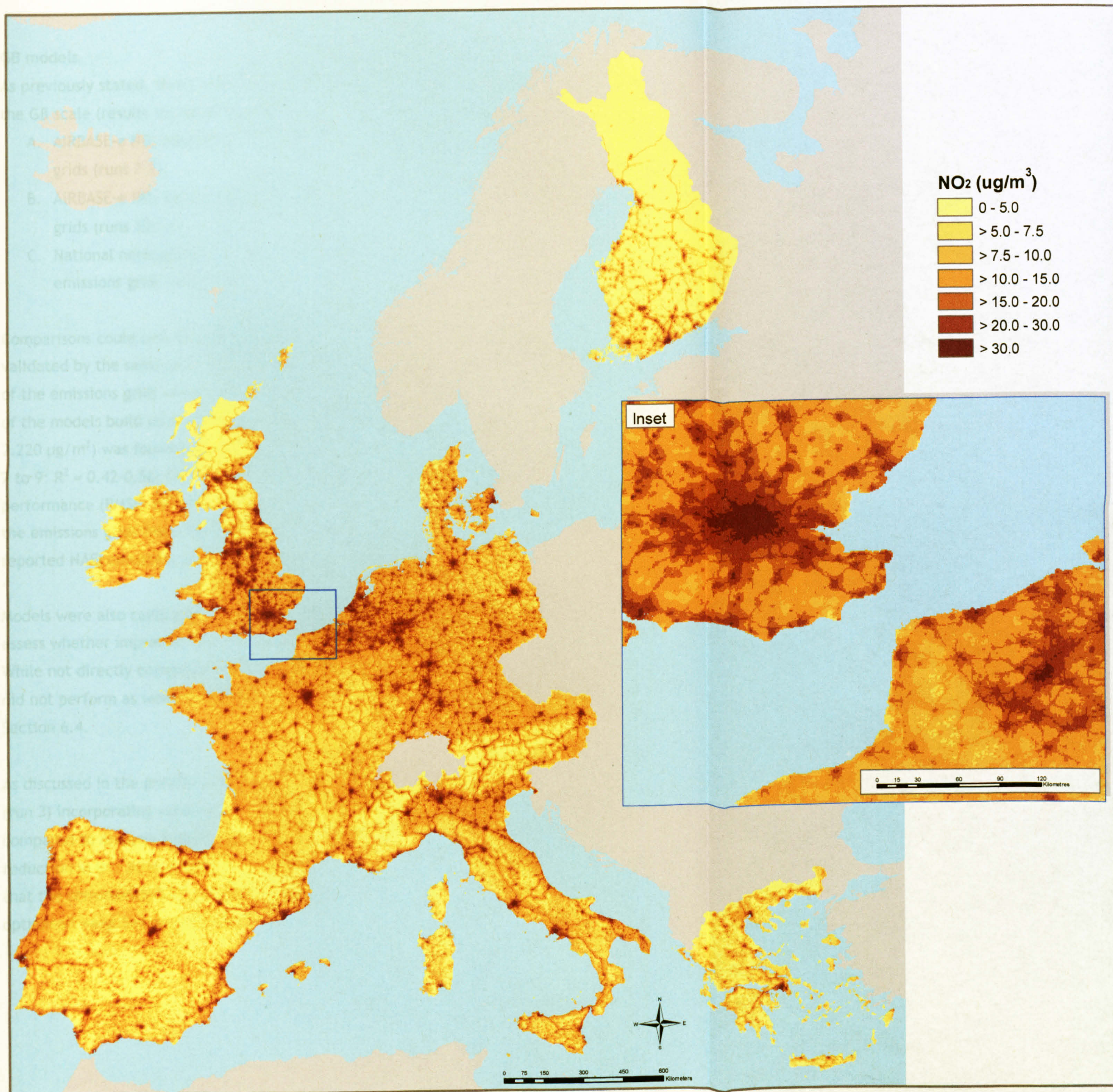


Figure 6.7
Annual mean background NO_2 ($\mu\text{g}/\text{m}^3$)
Focalsum
Run 3, Analysis 3

GB models

As previously stated, three combinations of data were available for modelling NO₂ at the GB scale (results shown in Table 6.9):

- A. AIRBASE-e NO₂ background sites in GB (n = 62) and the EU-wide emissions grids (runs 7-9)
- B. AIRBASE-e NO₂ background sites in GB (n = 62) and the national NAEI emissions grids (runs 10-12)
- C. National network NO₂ background sites (n = 580) and the national NAEI emissions grids (runs 13-15)

Comparisons could only be made between runs sharing one common data set and validated by the same procedure. Groups A and B which differed only in the source of the emissions grids were thus compared. Using jackknife techniques the robustness of the models build using the NAEI grids (runs 10 to 12: $R^2 = 0.46-0.57$; SEE = 6.464-7.220 $\mu\text{g}/\text{m}^3$) was found only marginally to exceed those based on the EU grids (runs 7 to 9: $R^2 = 0.42-0.56$; SEE = 6.543-7.504 $\mu\text{g}/\text{m}^3$). The other measures of performance (RMSE, F2 and FB) did not differ between analyses. This indicates that the emissions grids from the APMoSPHERE project were comparable to the nationally reported NAEI emissions grid in Great Britain.

Models were also calibrated on the basis of the national diffusion tube network to assess whether improved models could be derived from a greater density of monitors. While not directly comparable to other models, the group C models (runs 13 to 15) did not perform as well as expected. Reasons for this discrepancy are discussed in Section 6.4.

As discussed in the previous section, the best EU model was derived from Analysis 3 (run 3) incorporating wind speed. At the GB level, however, Analysis 3 broke down compared to Analysis 1 and 2 in each of the groups (A, B and C) as reflected by the reduction in R^2 . As also indicated by the calibration results, the validation showed that the consistently best model at the GB level was Analysis 2 based on the optimised weighted annulus windows.

6.3.2 Residual concentrations and traffic analysis

In order to gain further insight into the performance of the best NO₂ focalsum model (run 3), the residuals were examined at the validation sites across the EU. The overall distribution of the residuals from these sites is shown in Figure 6.8. The residuals exhibit a slight negatively skewed distribution with a median concentration of -0.2 µg/m³ and a standard deviation of 6.6 µg/m³.

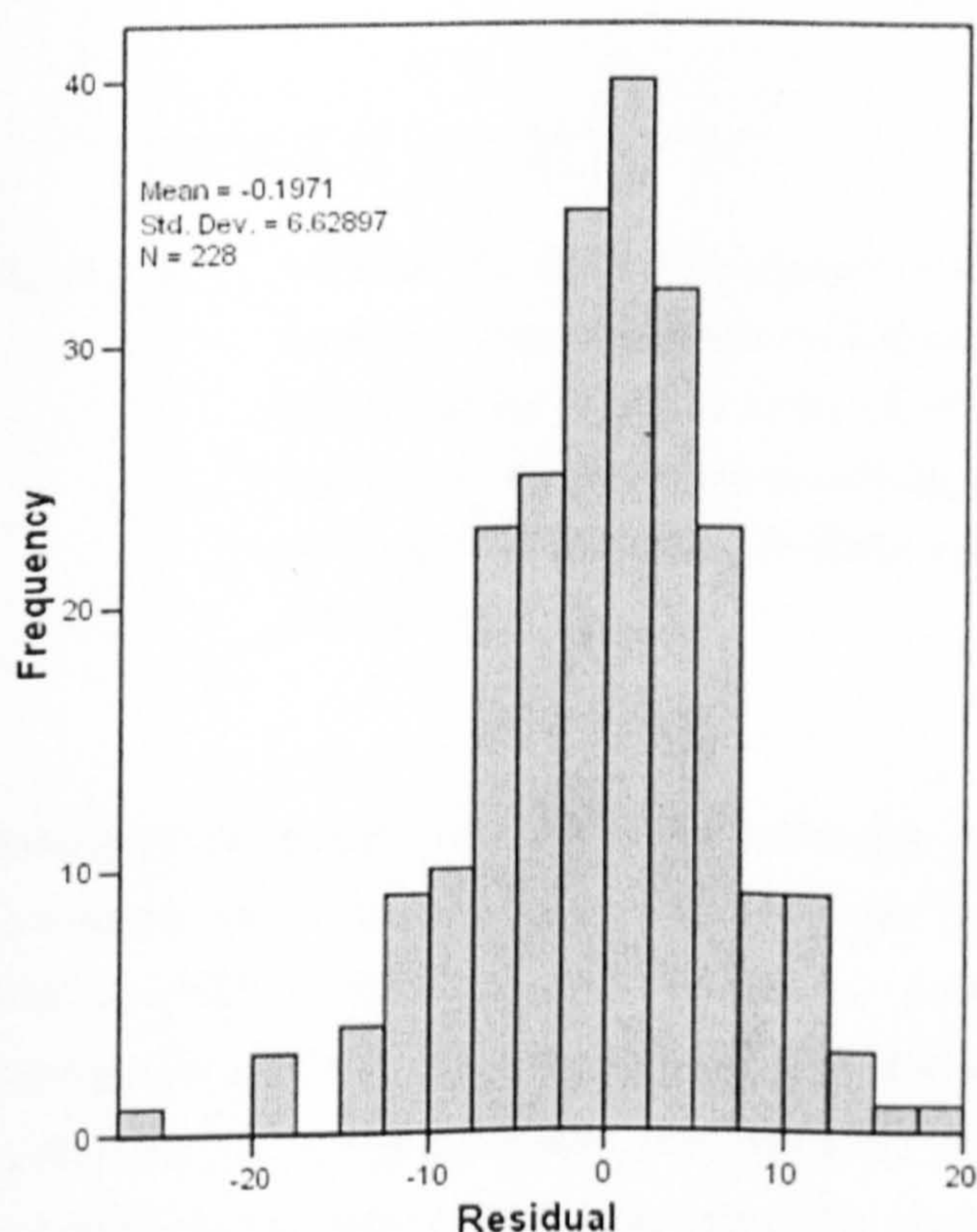


Figure 6.8 Residual NO₂ (µg/m³) at background sites

Figure 6.9 shows the residuals for the validation sites by affinity zone and country. Background concentrations tended to be under predicted in Belgium, Italy and the Netherlands and over predicted in Finland, Greece and Portugal. They also tended to be under predicted in the mixed rural, motorway and major road corridor zones (zones 1, 6 and 7) and over predicted in the forested uplands and the motorway intersections and industrial zones (zones 5, 8 and 10).

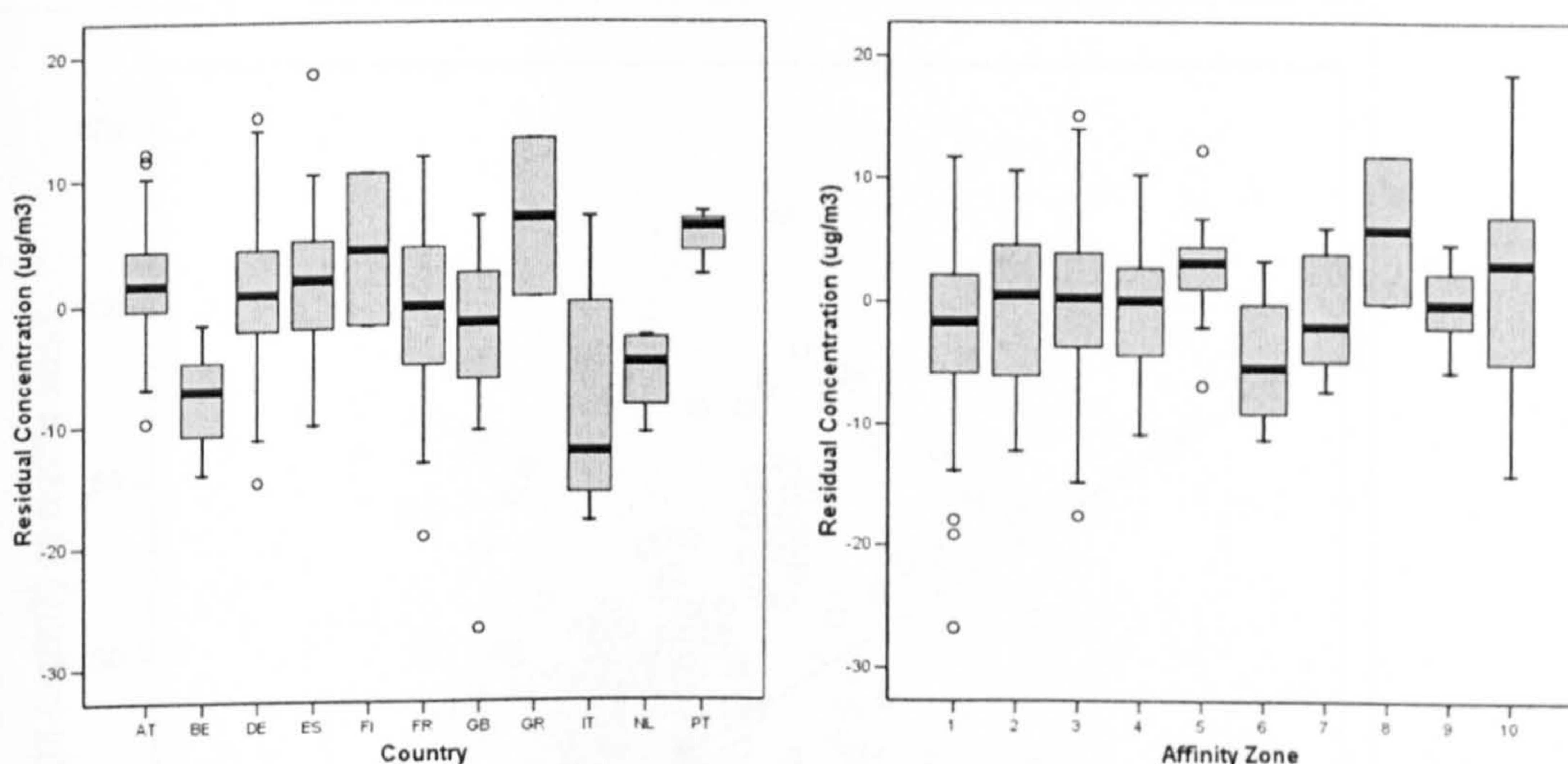


Figure 6.9 **Residual NO₂ (µg/m³) by country and affinity zone**
 Residual concentrations at background sites.
 Affinity zones: (1) Mixed rural; (2) High density residential; (3) Low density residential; (4) Metropolitan centres; (5) Forested uplands; (6) Motorways; (7) Major road corridors; (8) Motorway intersections; (9) Suburban; (10) Industrial

Modelled background concentrations were also compared with measured concentrations at the EU traffic sites, in order to determine whether they could be used as a basis for prediction at these sites. To achieve this, traffic sites were intersected with the modelled NO₂ surface, and the modelled NO₂ background concentration regressed against the measured NO₂ concentration (Figure 6.10). The slope of the regression line is approximately equal to one, and the average difference between modelled background and monitored concentrations at the traffic sites is 12.85 µg/m³. The poor fit of the regression model ($R^2 = 0.26$; $SEE = 14.092 \mu\text{g/m}^3$), however, means that concentrations at traffic sites vary independently of the background concentrations. Simply adding a constant to reflect traffic micro-environments thus produces a poor estimate of concentrations at traffic locations. As shown in Figure 6.11, this would lead to over predictions of NO₂ concentrations in several countries (notably Denmark, Spain, Finland, Ireland and Portugal) and under predictions in others (e.g. France, Great Britain, Greece and Italy). Similar variations in predictions would also occur by affinity zone: the forested uplands and motorways (zones 5 and 6) would tend to be under predicted, while the major road corridors (zone 7) would be over predicted.

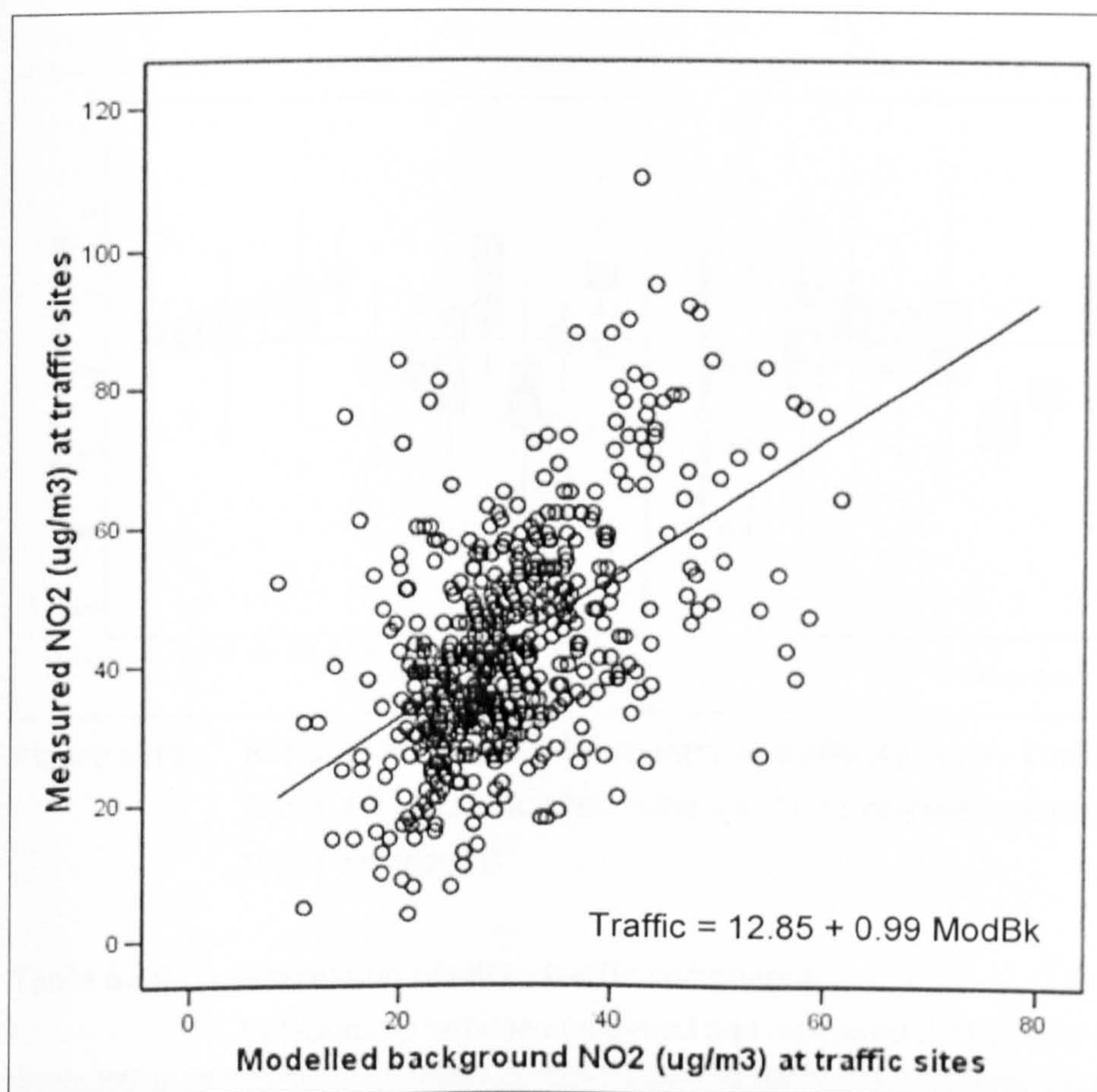


Figure 6.10 Modelled background vs. measured NO₂ at traffic sites

To further explore these national and regional variations, individual regression models were developed for each country and affinity zone (Table 6.10). The results show that the relationships vary considerably by country and affinity zone. If predictions at traffic sites are to be made, therefore, additional, higher resolution data (e.g. on traffic volumes or local street characteristics) would need to be incorporated into the models. These variables might also need to be customised by country or affinity zone, to reflect contextual differences in influences on traffic concentrations.

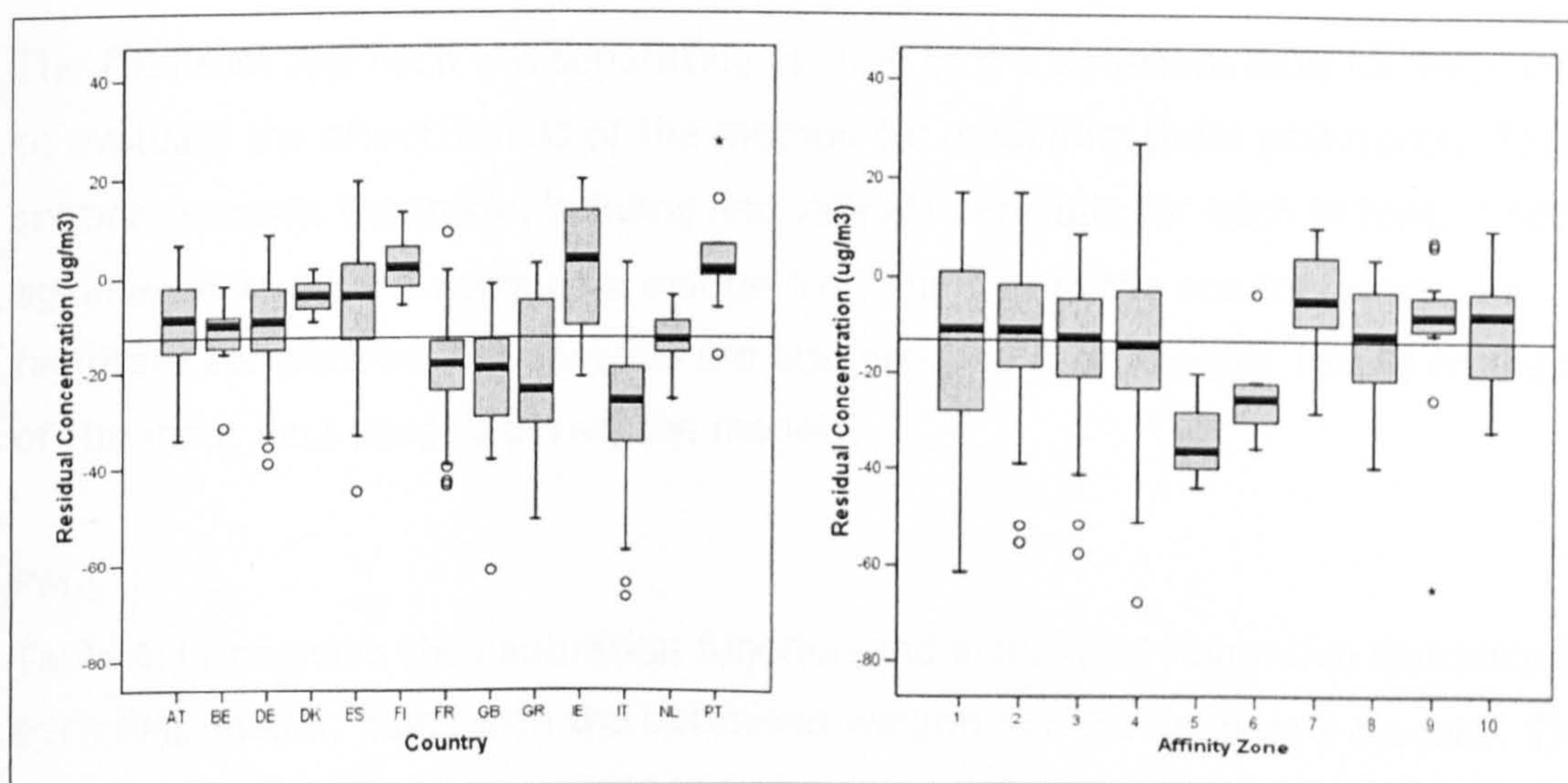


Figure 6.11 Residual NO₂ (µg/m³) by country and affinity zone - traffic
Shown for all traffic sites. Line y = -12.85 represents constant to reflect traffic component.

Table 6.10 Regression results - traffic component
Relationship between modelled and measured background NO₂ at traffic sites

Area	N	R	R ²	SEE	Unstandardised Coefficients			
					Intercept		Slope	
EU	500	0.51	0.26	14.092	12.85	*	0.99	*
AT - Austria	35	0.46	0.21	10.142	17.77	*	0.69	*
BE - Belgium	7	0.81	0.66	8.612	-16.06		1.92	*
DE - Germany	125	0.62	0.38	9.481	-2.45		1.45	*
DK - Denmark	3	0.86	0.73	3.300	21.69		0.43	
ES - Spain	122	0.68	0.47	12.200	1.08		1.11	*
FI - Finland	11	0.41	0.17	6.160	12.77		0.50	
FR - France	66	0.71	0.50	9.863	14.88	*	1.12	*
GB - Great Britain	22	0.44	0.19	13.526	29.02	*	0.78	*
GR - Greece	11	0.52	0.27	17.877	-6.84		1.68	
IE - Ireland	4	0.62	0.39	5.224	36.26	*	-0.26	
IT - Italy	72	0.48	0.23	14.153	27.57	*	0.99	*
NL - The Netherlands	13	0.36	0.13	5.823	27.52		0.51	*
PT - Portugal	9	0.30	0.09	12.196	20.67		0.42	
Affinity Zones								
1. Mixed rural	15	0.34	0.12	19.585	60.11	*	-1.38	
2. High density residential	118	0.19	0.04	12.822	23.3	*	0.56	*
3. Low density residential	148	0.52	0.27	12.141	6.66		1.25	*
4. Metropolitan centres	132	0.44	0.19	16.171	15.35	*	0.96	*
5. Forested uplands	3	0.34	0.11	17.122	36.74		0.73	
6. Motorways	8	0.04	0.00	8.655	46.37	*	0.05	
7. Major road corridors	17	0.65	0.42	10.987	-3.98		1.39	*
8. Motorway intersections	18	0.63	0.39	11.907	7.69		1.13	*
9. Suburban	13	0.13	0.02	18.571	25.33		1.04	
10. Industrial	28	0.45	0.21	11.415	7.83		1.1	*

* Significant at α = 0.05

6.3.3 Other pollutant models

The focalsum approach was separately applied to the emissions data for PM₁₀ and SO₂ to evaluate the effectiveness of the method for modelling these pollutants. This section presents the model building and validation results for each in turn. Once again each analysis is assigned a unique 'run' number in the results tables. To facilitate comparison, the analyses are also grouped (i.e. Analysis 1 to 3) on the basis of the input data used to derive the models.

PM₁₀

Table 6.11 presents the calibration function and associated regression statistics for each PM₁₀ model, along with the optimised weights for the Analysis 2 models. The EU models were built using the AIRBASE-e monitoring sites and EU-wide emissions grids, while the GB models were based on the national PM₁₀ emissions grid from NAEI and the 47 AIRBASE-e monitoring sites in Great Britain.

Although the regression statistics are generally poor, the best calibration function for both the EU and GB models was achieved from Analysis 2 in which the distance weights were optimised. Nevertheless, only a 0.3% increase in percent agreement between the log-transformed concentrations and log-transformed emissions index at the training sites was achieved for the EU Analysis 2 model (run 2 compared to run 1). Analysis 2 at the GB level resulted in a larger (6.9%) increase in agreement compared to Analysis 1 (run 6 compared to run 5). The optimised weights for both the EU and GB models show that the decay function is more gradual than $1/d^2$ for the inner-most rings (to a distance of about 2km). The weights for both models, however, decay as expected to background levels within the 11km distance.

The validation results for PM₁₀ are presented in Table 6.12. The focalsum method produced poor predictions of annual mean concentrations at the EU level ($R^2 = 0.01$; $SEE = 6.985\text{--}6.987 \mu\text{g}/\text{m}^3$), performing much worse than the reference method ($R^2 = 0.48$; $SEE = 5.037 \mu\text{g}/\text{m}^3$). This is likely to be due, in part, to inadequacies in the EU-wide emissions data for PM₁₀ in some countries because of incomplete or erroneous reporting of national emissions totals for some SNAP sectors. For this reason, the focalsum method was also applied (in run 4) to only those countries with complete emissions data. This significantly improved model performance ($R^2 = 0.11$; $SEE = 6.087 \mu\text{g}/\text{m}^3$), though the results were still poor compared to the reference method.

The GB models (runs 5 to 7, Table 6.12) were improved compared to the EU level analysis ($R^2 = 0.11-0.19$; $SEE = 2.973-3.111 \mu\text{g}/\text{m}^3$) - though the use of Jackknife techniques for validation may have enhanced the performance measures. Similar to the NO_2 GB models, Analysis 2 with the optimised weights produced the best validated model.

Due to the poor performance of the focalsum method for modelling annual mean PM_{10} concentrations across the EU, only the reference method was mapped for use in the exposure assessment (Section 7.2.3). The PM_{10} concentration map is shown in Figure 6.12.

Table 6.11 PM₁₀ model building
Including distance weights applied to annulus rings

Run/ Analysis	Method	Calibration			Distance Weights for Each Annulus Ring											
		N	Function	R ²	SEE	1	2	3	4	5	6	7	8	9	10	11
EU-wide FOCALSUM models																
1/1	FS annulus (1/d ²)	326	$\exp^{2.947 + 0.043 \ln lc}$	0.06	0.295	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
2/2	FS annulus ow	326	$\exp^{2.943 + 0.041 \ln lc}$	0.06	0.294	1.0	1.0	0.04	0.04	0.02	0.02	0.02	0.02	0.02	0.0	0.0
3/3	FS wind (1 /d ²)	326	$\exp^{3.002 + 0.037 \ln lc}$	0.05	0.296											
4/3	FS wind (1 /d)*	74	$\exp^{2.991 + 0.020 \ln lc}$	0.01	0.331											
GB-wide FOCALSUM models																
5/1	FS annulus (1/d ²)	47	$\exp^{2.548 + 0.1 \ln lc}$	0.26	0.165	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
6/2	FS annulus ow	47	$\exp^{2.501 + 0.123 \ln lc}$	0.33	0.157	1.0	0.5	0.04	0.04	0.04	0.02	0.02	0.02	0.02	0.02	0.02
7/3	FS wind (1/d ²)	47	$\exp^{2.679 + 0.081 \ln lc}$	0.25	0.166											

Notes: ow = optimised weighted; calibration based on log-transformed emissions index and log-transformed concentrations

* Run 4 was modelled only on select countries with complete emissions data (i.e. AT, BE, FI, GB, NL)

Table 6.12 Background PM₁₀ modelling results

Run/ Analysis	Method	Network	Emissions	Total N	Validation Results						
					R ²	Adj R ²	SEE	RMSE	F2	FB	Method
Reference model from APMoSPHERE (EU-wide)											
	Reference	AIRBASE-e		430	0.48	0.48	5.037	5.074	99.0%	-0.031	LS (N=104)
EU-wide FOCALSUM models											
1/1	FS annulus (1/d ²)	AIRBASE-e	2,3,4,5,6	430	0.01	0.00	6.987	7.115	98.1%	-0.063	LS (N=104)
2/2	FS annulus ow	AIRBASE-e	2,3,4,5,6	430	0.01	0.00	6.985	7.076	98.1%	-0.054	LS (N=104)
3/3	FS wind (1/d ²)	AIRBASE-e	2,3,4,5,6	430	0.01	0.00	6.986	7.077	97.1%	-0.057	LS (N=104)
4/3	FS wind (1/d ²)*	AIRBASE-e	2,3,4,5,6	96	0.11	0.06	6.087	6.486	100.0%	-0.044	LS (N=22)
GB-wide FOCALSUM models											
5/1	FS annulus (1/d ²)	AIRBASE-e	NAEI at00	47	0.12	0.11	3.093	3.045	100.0%	0.000	LOO (N=47)
6/2	FS annulus ow	AIRBASE-e	NAEI at00	47	0.19	0.17	2.973	2.916	100.0%	0.000	LOO (N=47)
7/3	FS wind (1/d ²)	AIRBASE-e	NAEI at00	47	0.11	0.09	3.111	3.078	100.0%	-0.001	LOO (N=47)

Notes: ow = optimised weighted; LS = least squares regression; LOO = leave-one-out (jackknife)

* Run 4 was modelled only on select countries with complete emissions data (i.e. AT, BE, FI, GB, NL)

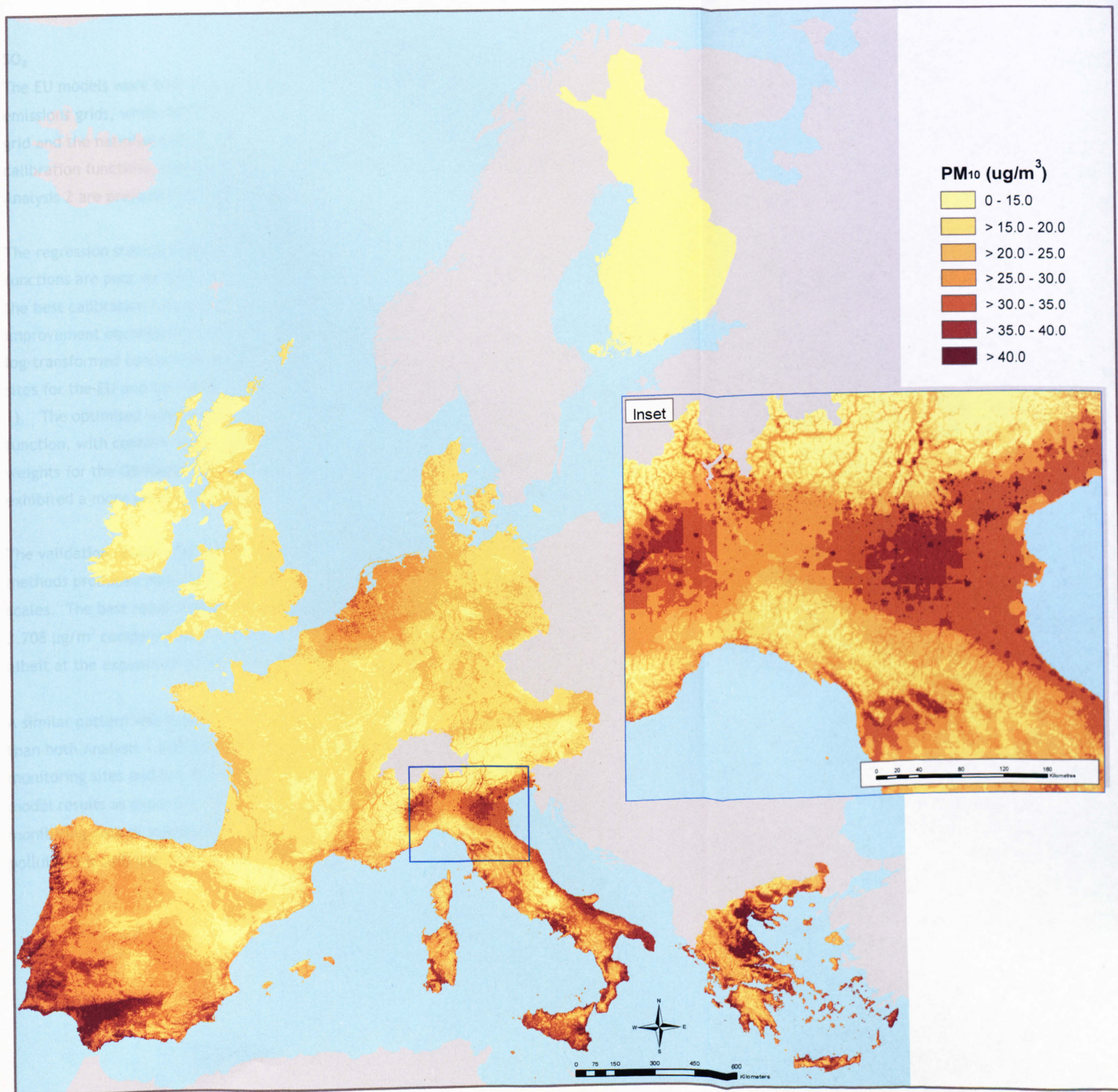


Figure 6.12
Annual mean background PM₁₀ (µg/m³)
Reference method

SO₂

The EU models were built using the AIRBASE-e monitoring sites and EU-wide emissions grids, while the GB models were based on the NAEI national SO_x emissions grid and the national monitoring sites (n = 135) in Great Britain. The SO₂ model calibration functions, associated regression statistics and the optimised weights for Analysis 2 are presented in Table 6.13.

The regression statistics shown in Table 6.13 indicate that the SO₂ calibration functions are poor for both the EU and GB scale models. As per the other pollutants, the best calibration function for both scales was achieved using Analysis 2. The improvement equated to a 0.8% and 1.7% increase in percent agreement between the log-transformed concentrations and log-transformed emissions index at the training sites for the EU and GB models, respectively (i.e. comparing Analysis 2 with Analysis 1). The optimised weights for the EU model (run 2) broadly resembled the $1/d^2$ function, with concentrations decaying to background levels within 11km. The weights for the GB model also returned to background levels within 11km, but exhibited a more gradual decay for the inner-most rings (to a distance of about 2km).

The validation results (Table 6.14) show that both the reference and focalsum methods produced poor models of annual mean SO₂ concentrations at both spatial scales. The best result was obtained for the EU level Analysis 2 ($R^2 = 0.35$; SEE = $3.708 \mu\text{g}/\text{m}^3$ compared to $R^2 = 0.20$; SEE = $4.137 \mu\text{g}/\text{m}^3$ for the reference method), albeit at the expense of a rather large negative fractional bias (i.e. -18%).

A similar pattern was found for the GB level models with Analysis 2 performing better than both Analysis 1 and the more complex Analysis 3. The greater density of monitoring sites and use of the national emissions data, however, did not improve model results as expected. This may be related to the clustered nature of these monitoring data or inadequacies in the focalsum methodology for modelling pollutants mainly derived from point sources (i.e. industrial stacks).

Table 6.13 SO₂ model building

Including distance weights applied to annulus rings

Run/ Analysis	Method	Calibration			Distance Weights for Each Annulus Ring											
		N	Function	R ²	SEE	1	2	3	4	5	6	7	8	9	10	11
EU-wide FOCALSUM models																
1/1	FS annulus (1/d ²)	665	exp ^{0.927 + 0.166 lnlc}	0.18	0.603	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
2/2	FS annulus ow	665	exp ^{0.924 + 0.170 lnlc}	0.19	0.600	1.0	0.123	0.123	0.044	0.044	0.044	0.022	0.022	0.014	0.014	0.014
3/3	FS wind (1/d ²)	665	exp ^{1.289 + 0.109 lnlc}	0.12	0.626											
GB-wide FOCALSUM models																
4/1	FS annulus (1/d ²)	102	exp ^{1.786 + 0.305 lnlc}	0.12	0.502	1.0	0.25	0.111	0.063	0.04	0.028	0.02	0.016	0.012	0.01	0.008
5/2	FS annulus ow	102	exp ^{1.543 + 0.339 lnlc}	0.14	0.497	1.0	1.0	0.078	0.078	0.078	0.078	0.078	0.015	0.015	0.015	0.015
6/3	FS wind (1/d ²)	102	exp ^{2.301 + 0.184 lnlc}	0.07	0.517											

Notes: ow = optimised weighted; calibration based on log-transformed emissions index and log-transformed concentrations

Table 6.14 Background SO₂ modelling results

Run/ Analysis	Method	Network	Emissions	Total N	Validation Results						
					R ²	Adj R ²	SEE	RMSE	F2	FB	Method
Reference model from APMoSPHERE (EU-wide)											
	Reference	AIRBASE-e		885	0.20	0.20	4.095	4.137	83.6%	-0.129	LS (N=220)
EU-wide FOCALSUM models											
1/1	FS annulus (1/d ²)	AIRBASE-e	1,2,3,4,5,6	885	0.34	0.34	3.726	4.053	82.7%	-0.188	LS (N=220)
2/2	FS annulus ow	AIRBASE-e	1,2,3,4,5,6	885	0.35	0.34	3.708	4.020	82.3%	-0.184	LS (N=220)
3/3	FS wind (1/d ²)	AIRBASE-e	1,2,3,4,5,6	885	0.27	0.27	3.916	4.262	80.9%	-0.210	LS (N=220)
GB-wide FOCALSUM models											
4/1	FS annulus (1/d ²)	UK	NAEI at00	135	0.24	0.22	7.542	7.897	78.8%	-0.151	LS (N=33)
5/2	FS annulus ow	UK	NAEI at00	135	0.22	0.20	7.640	7.925	78.8%	-0.159	LS (N=33)
6/3	FS wind (1/d ²)	UK	NAEI at00	135	0.09	0.06	8.265	8.270	75.8%	-0.130	LS (N=33)

Notes: ow = optimised weighted; LS = least squares regression

6.4 Discussion

The results presented above demonstrate that it is possible to generate EU-wide maps of annual mean background NO₂ concentrations using a moving window approach. Focalsum performed as well as the reference method, a combination of regression modelling and kriging, and produced a more realistic map reflecting expected spatial variations in concentrations. It is also informative to compare results of the focalsum modelling with that using similar methods (but more detailed emissions data) by Loibl & Orthofer (2001) in Austria. In their study, they incorporated information on stack height and emissions profiles in a similar grid-based emissions dispersion approach to model NO₂ concentrations. For NO₂, Loibl & Orthofer (2001) report that 54% (72 sites) of the measurements at monitoring sites were within $\pm 15 \mu\text{g}/\text{m}^3$ of the modelled concentrations and 76% (101 sites) were within $\pm 25 \mu\text{g}/\text{m}^3$. Using the EU run 3 model developed here, all sites are within $\pm 15 \mu\text{g}/\text{m}^3$ of the actual concentrations at the Austrian validation sites (n=26). This supports the interpretation that the performance of the focalsum methods used here is at least as good as that of other models.

The remainder of this discussion focuses on the possible sources of error and issues in applying the methodology, and potential enhancements which may be needed to facilitate the modelling of pollutants other than NO₂.

6.4.1 Possible sources of error

The quality of the emissions data

One of the key constraints of the focalsum approach is its reliance on the emissions data as a measure of source activity and intensity. These data, at a resolution fine enough to detect local variations, are difficult to source at the EU level. To address this data gap, activities in the APMoSPHERE project involved refining and applying a consistent methodology to derive a detailed emissions inventory across the EU. This analysis, however, was subject to several sources of error. In particular errors related to: (1) those in the methods used to produce the national totals; (2) the incomplete reporting of national totals for some sectors; and (3) errors in the methodology to disaggregate totals to the 1km level.

The APMoSPHERE emissions modelling protocol was based on a top-down approach in which national totals were disaggregated to the local level. The methodology depended not only on the availability of input data on source distribution, activity and high quality proxies upon which the disaggregation was based, but also on the reliable reporting of SNAP category emissions at the country level. National emissions estimates are known to suffer from major uncertainties. Dore et al. (2004) for example, quote uncertainties of -20 to +50% in national PM₁₀ estimates in the UK, though only $\pm 3\%$ for SO₂. Estimates of PM₁₀ emissions were further limited by the lack of emissions totals for some SNAP categories in some countries.

In general, there was little opportunity in the APMoSPHERE project to validate the emissions grids with independent data (comparisons were only possible in Great Britain and the Netherlands) so reliability of the emissions data is not known. In general, however, estimates of NO_x emissions were regarded as broadly accurate, while those of other pollutants (including PM₁₀ and SO_x) were considered to suffer from errors of national reporting and/or modelling. Improved reporting of PM₁₀ emissions across Europe and refinements to the APMoSPHERE emissions modelling protocol, as guided by more extensive validation, is therefore needed to provide a reliable basis for pollution modelling.

Inadequate characterisation of dispersion patterns

The second type of error associated with the focalsum approach relates to inadequacies in the characterisation of pollutant dispersion patterns. In particular, these errors include: (1) deficiencies in the shape of the moving windows; (2) the use of inaccurate distance weights; (3) inadequate allowance for meteorological factors; and (4) insufficient characterisation of the chemical processes.

Very basic window shapes (i.e. annulus and wedge) were used in the analysis. These windows, especially the wedges used in Analysis 3 which incorporated monthly wind direction data, may be too simplistic for characterising the dispersion of some pollutants. They may, for example, be too narrow or too broad which is especially relevant, perhaps, in relation to the effects of long-range transport of PM₁₀ and SO₂. The wedge shape also assumes that pollutants are dispersed only in the downwind direction, which is not necessarily valid especially under calm wind conditions. In addition to accounting for wind speed by modifying the window shape, it is possible

that customised windows and decay functions may be needed to reflect differences in emission heights and dispersion patterns depending on the pollutant and source. This topic is further explored in Section 6.4.2 related to potential enhancements in the modelling approach.

The role of meteorological data in the modelling also merits particular attention. In principle, incorporation of monthly mean wind direction into the focalsum models might have been expected to improve model performance. Although slight improvements were seen for the EU models incorporating wind direction for NO₂, this did not always happen. The best EU-wide focalsum model for SO₂, for example, was not obtained by the more advanced Analysis 3, but instead by the more basic (non-directional) Analysis 2. Additionally, when wind direction was incorporated into the NO₂ models at the GB level, a marked reduction in model performance was seen. The effect of wind speed was also explored in modelling NO₂ at the EU level by way of modified window shapes to incorporate three levels of wind speed (Section 6.2.1). This analysis (data not shown), however, resulted in reduced model performance and was therefore not extended to the other pollutants.

Several explanations for these results can be proposed. One is that the coarse spatial resolution of the wind data, which derived from a 50 x 50km grid across Europe, was insufficient to reflect the local effects of wind direction, especially in urban, upland or coastal areas, where complex wind regimes might occur. This might account for the poor performance of the models including wind direction for SO₂, emitted mainly from tall stacks and thus strongly affected by meteorology, as well as for NO₂ in the UK, which is characterised by complex local wind patterns. Equally important for SO₂ modelling, however, might be the lack of information on other meteorological factors, such as atmospheric stability and mixing heights, which have important effects on dispersion of high-level emissions.

A further issue relates to the coarse temporal resolution of the meteorological data. For Analysis 3, a measure of the number of months per year that the wind blows from a particular direction was assigned to eight non-overlapping segments. Monthly frequency was used because the wind direction data were sourced as monthly means. It is possible, however, to source more detailed meteorological data, including daily means, which would produce a more accurate depiction of the fluctuations in wind direction across the study area.

Finally, the focalsum methodology does not take account of the effects of long-range transport or atmospheric chemistry. This can be particularly problematic in modelling PM_{10} which is regional in nature and, thus, influenced by long-range transport. Particulates are also subject to deposition and chemical reactions, which complicate the dispersion process and the characterisation of PM_{10} . A large proportion of the particulate concentrations in urban air, for example, is due to chemical reactions including the photochemical oxidation of NO_2 and SO_2 to produce secondary particles such as nitrate (NO_3^-) and sulphate (SO_4^{2-}) (Koutrakis & Sioutas 1996).

Issues related to spatial scale

Another source of error concerns the spatial resolution of the emissions data and the resulting focalsum models. Use of a 1km scale for modelling emissions and air pollution patterns across the whole of the EU is, of course, extremely challenging and already stretches the available data to their limits. The monitoring data used for model validation, however, are point measurements, and as such take no account of variations in concentrations within the 1km grid cells. These may nevertheless be substantial. Traffic-related NO_2 , especially, shows steep concentration gradients over distances of a few hundred metres or less (Briggs et al. 1997; Gilbert et al. 2003). SO_2 and PM_{10} may show less local variation, but close to sources - for example, where emissions from major point sources escape as fugitives or where plumes from tall stacks touch the ground - major 'hotspots' may occur. While these may be picked up in the monitored concentration data, they cannot be detected by models of the resolution used here. Instead, they are likely to remain as noise in the monitored data. There is, as a consequence, a level of uncertainty in the modelling that cannot be resolved, unless far more detailed emissions and other data were available.

Errors in the monitoring data

Inadequacies associated with the monitoring data - for example, bias in site location, sparseness of the sites and their uneven distribution across the study area - may also hinder air pollution modelling efforts. This was particularly evident for modelling SO_2 where model performance was clearly hampered by the sparseness of the monitoring network.

The availability of national data for some pollutants was used to take advantage of a greater density of monitoring sites. Focalsum applied at the EU level, however, produced better results compared to those focused on Great Britain even when the national network was used. For NO₂, the reduction in performance may in part be explained by the fact that the national network comprises passive samplers measuring monthly averages, whereas the AIRBASE-e network is made up of fixed-location active samplers operated under EU guidelines. Despite the substantially improved network in terms of spatial coverage (i.e. 580 background monitors compared to 62 across GB), uncertainties associated with passive samplers seem to diminish their utility in focalsum modelling. These uncertainties relate to observed overestimates of concentrations measured by diffusion tubes compared to standard methods. There is some debate, for example, whether this observation is caused by chemical reactions between NO and O₃ also present in the diffusion tube (Heal et al. 1999) or due to wind turbulence effects in the tube which cause the diffusion path to shorten (Atkins & Lee 1995; Gair & Penkett 1995). To further complicate the issue, the diffusion tubes are managed by local authorities, each potentially using their own criteria for locating the monitors leading to further inconsistencies across the combined network. The tubes are also sent to different laboratories for analysis which could result in substantial differences in analytical methods and accuracy.

6.4.2 Potential enhancements to the focalsum approach

Despite the limitations outlined above, focalsum was found to be a very efficient method for modelling high resolution air pollution given the powerful ArcInfo GRID environment in which the models are run. The reference method, for example, involved approximately 5 days decision making and processing time compared to the focalsum method which took a matter of hours (in each case, excluding the time required to generate the input data). This section outlines enhancements which could be introduced to potentially improve the air pollution models derived using the focalsum approach.

The current focalsum methodology is based on a common set of annulus windows and weights applied to all of the source emissions grids used to model each pollutant. As shown in Table 6.2, the source grids relate to point sources, traffic and other transport, domestic and area sources emissions. The first enhancement, thus, relates to the need for customised window shapes and weights to reflect different

emission heights and dispersion patterns depending on the source. A smaller window may be more appropriate for low-level traffic emissions whereas a larger window may be necessary to disperse point sources to a wider area. The gradient within these windows may also need to be customised according to source. While a decay function of $1/d^2$ may be suitable for modelling traffic-related sources, a more appropriate function for industrial point sources is probably the Gaussian distribution which is commonly applied in dispersion modelling. For annulus windows, Analysis 2 (manual adjustment based on regression analysis) could be applied so that optimised weights are determined in turn for each source grid.

The second proposed enhancement involves increasing the sophistication of windows to reflect better the meteorological conditions across the study area. This, for example, would include creating windows, appropriately weighted, to reflect both wind direction and wind speed. Rather than using the wedge shape to partition wind into eight directions, elongated annulus windows may be created. More complex weighting functions are also needed to simultaneously account for distance from the target and wind direction. As previously stated, the temporal resolution of the meteorological data should also be improved so that the wind direction and wind speed variables are calculated on the basis of daily rather than monthly means. In addition, information on atmospheric stability and mixing height, together with stack height (which was not available for this study), is most likely required to enhance the methodology for SO₂ modelling. Detailed dispersion profiles for pollutants emitted from industrial stacks would also be useful in informing decisions related to the required window sizes and shapes needed to derive the most appropriate models.

As was undertaken in the research by Collins (1998a), it would also be interesting to explore focalsum in combination with another modelling technique. It is possible, for example, that a more realistic PM₁₀ map may be derived by using kriging, which tends to produce smoothed surfaces, to model the long-range global background component. The local component could then be modelled using focalsum applied to the relevant source grids (e.g. emissions from traffic and agriculture, as appropriate) and added to the background map. A finer resolution emissions grid may also be necessary to model differences occurring at the very local scale for some sources (especially for traffic-related sources).

6.5 Conclusion

Focalsum methods offer potential as a powerful way of modelling air pollution at the regional and international scale, and can match the performance of other methods where adequate emissions data exist. They also have the benefit of greatly reduced computational demands compared to other methods, such as those explored in the APMoSPHERE project. The main constraint on their use, however, is the availability of reliable emissions estimates, at a suitable resolution - though the same moving window techniques can also help to produce these data, as shown by the APMoSPHERE project. Currently, the focalsum methodology works best for low-level sources, such as those from traffic, which exhibit relatively simple and local scale dispersion characteristics. Opportunity exists, however, to improve and enhance the methodology, by using higher resolution emissions data than available here. Enhancements may also be possible by incorporating functions to reflect meteorological conditions, especially wind speed and wind direction, and the emission heights and dispersion patterns of different sources (e.g. tall stacks).

7 GIS-based Exposure Assessment – A Discussion

While severe pollution episodes such as the infamous 1952 London smog are less common today due to reductions in fossil fuel combustion, air pollutants - in particular those that are related to traffic - continue to occur at unacceptable levels in some areas of Europe. Additionally, there is a vast literature in the field of epidemiology demonstrating that both short and long-term exposures to air pollution continue to attribute to respiratory and cardiovascular mortality and other health outcomes, even at today's historically low concentrations. The EU has, thus, maintained air pollution as one of its main political concerns, addressing the issue through policy and initiatives such as the Sixth Environment Action Programme and the CAFE initiative which aim to protect human health and the environment. The EU has also imposed the development of the EUROAIRNET monitoring network and information system (AIRBASE) designed to provide data for conducting air quality assessments at the European scale in order to track air pollution problems and inform policy decisions.

Policy makers rely on epidemiological evidence to set the guideline limits to protect human health, and exposure assessment is an important tool used by epidemiologists in determining individual and population-based health effects of air pollution. Reliable data on population exposures to air pollution are also required for risk assessment, and to evaluate policy priorities and actions. The ideal exposure assessment methods involving personal or bio-monitoring, however, are often too costly and unrealistic for evaluating chronic exposure to air pollution for large populations. As a result, researchers have tended to rely on alternative methods based on measured concentrations at the nearest available monitoring sites or exposure proxies such as proximity to sources or the level of nearby source activity. With the growing use and power of GIS more sophisticated methods have been developed and applied in the field of air pollution exposure assessment. The current 'state of the art' in air pollution modelling (Chapter 2) includes spatial interpolation and regression mapping, as well as dispersion modelling.

Longitudinal studies, which follow a group of individuals over several to many years, are probably the most informative epidemiological method to study chronic exposures to air pollution. There are only a handful of large population cohort studies exploring health effects of long-term exposure to air pollution, however, and

only two of these have been focused in Europe (Hoek et al. 2002a; Nafstad et al. 2004; Nafstad et al. 2003). As yet, a standard method for assigning an exposure measure to the cohort participants has not been devised - each of the air pollution cohorts, for example, used a different method including nearest pollution monitor, proxy measures or interpolation techniques. This thesis focused on developing methods for a map-based approach to assess exposures to annual mean NO₂ and PM₁₀ in large population studies. Various statistical and GIS-based methods were explored and used to create high resolution maps of both population density and pollution levels across Europe.

Each of the thesis aims outlined in Chapter 1 are discussed in turn in the following sections, starting with an evaluation of the components used to conduct the EU-wide, GIS-based exposure assessment.

7.1 Evaluation of the exposure modelling components

A map-based exposure assessment, which will be demonstrated in Section 7.2, is a practical way in which to obtain a snapshot of European-wide exposures to long-term air pollution. The reliability of the information derived from such an approach, however, depends on the quality of the data upon which the assessment is based - especially if the results will be used to inform policy decisions. The maps used to assess exposure, therefore, have to be of high enough resolution to quantify the spatial variability of the pollution surface and to provide detailed information on the location of the population at risk. Careful consideration was, therefore, given to fundamental decisions including the input data sources and the spatial resolution at which to conduct the map-based exposure assessment. As well, a significant amount of pre-processing and modelling was undertaken to prepare the models and maps for the exposure assessment.

GIS data were sourced for a range of variables including land cover, population, meteorology, topography and transport. Each of these data sets was obtained from a single source providing EU-wide coverage to facilitate future updates of the models. In selecting a scale at which the population and pollution models would be derived, two main factors were considered. The first was the resolution of the input data sets. The scales at which the above data were available across the EU ranged from a 100m grid for land cover to 50km grid for the climate data. While a very high

resolution (e.g. 10-100 metre) would be preferred for assessing air quality in urban areas where variations in air pollution occur over fine spatial scales, the resolution of most input data sets was clearly too coarse (i.e. especially the meteorology and NUTS 5 population data). This directly relates to the second consideration - the processing time required to run the models. As the spatial resolution is increased, the processing time also tends to increase. While processing 2.8 million grid squares at the 1 x 1km level was feasible, models based on a higher resolution would have been severely limiting. Likewise, reducing the resolution would have over-smoothed the population and pollution surfaces and made the exposure estimates less reliable.

This section evaluates the methods used to derive the main components used to facilitate the GIS-based exposure assessment. The modelling methods are discussed on the basis of efficiency, ease (or lack thereof) in updating the models and maps as new data become available, as well as the gaps and potential opportunities.

7.1.1 Modelling population density

Aim: To develop an accurate, high resolution (ca. 1km²) EU-wide model and map of population density.

A GIS-based (hybrid) approach to generate a high resolution population density map for the study area was described in Chapter 4. Redistribution of census-tract population to a 1km grid was achieved by disaggregating NUTS 5 population totals on the basis of land cover and night-time light intensity variables to reflect small-area variations in population density. Similar dasymetric mapping methods using satellite-derived ancillary data have been applied at a local and regional level with success (Chen 2002; Langford & Harvey 2001; Mennis 2002; Mennis 2003; Yuan et al. 1997). Most models, however, used only a binary (populated and unpopulated) or three-level classification (high-density, low-density and non-urban) to derive the weights for disaggregation. The GIS-based approach devised here not only used a greater selection of land cover classes (i.e. up to 16 grouped classes), it also incorporated three measures of night-time light intensity per land cover class to further strengthen the correlations between land use and population density.

Country and EU-level models were developed on the basis of land cover and light intensity as well as on land cover class alone. In the absence of light intensity, more

detailed land cover classifications were needed to provide discrimination of variations in population density. When light intensity was included, however, local differences in population density were apparent even when the land cover classes were more coarsely grouped. The inclusion of light intensity resulted in noticeable improvements in the models, especially for some of the southern countries, demonstrating that the land cover classification or the relationship between land cover and night-time light intensity might be non-uniform across Europe. In general, one can conclude that the large number of predictors (and especially the use of light intensity data) resulted in a more realistic redistribution of the population. Incorporating the full range of land classifications across Europe was also an improvement, as it is clearly false to assume that people only reside in built-up areas as some of the earlier dasymetric mapping examples do.

While the large set of detailed predictors can lead to better models, the input data requirements may also be viewed as a limitation to the hybrid approach. Given that the methods and AMLs are in place, and assuming the GIS data are suitably formatted, the population modelling procedure starting at the regression modelling stage could be completed in approximately two weeks of processing time. While the population density map itself can be easily updated, the main constraint will be the time required to pre-process the population data and reconstruction of the light intensity map using newer land cover data (CLC2000). Significant pre-processing was undertaken in the MANTLE project to construct the EU-wide light intensity map, which is already out-of-date. Updated census information should also be incorporated, which, unless improvements have been made to the coding in the geographic and statistical databases, might also prove time consuming (i.e. in linking the census information to the census tract boundaries). The value of the provision of reliable, consistent and high resolution census information across the EU also cannot be understated. These data are not only important to the application presented here, but are also vital to many other policy applications and research initiatives undertaken across Europe. If temporal light emissions data (e.g. annual updates), were available, for example, these methods could be used to study population change across the EU.

To the author's knowledge, the use of sophisticated techniques to map population density across a large geographic area has yet to be comprehensively explored. The only readily available example of population density mapped across a vast area is the

world-wide map (GPW-3) by CIESIN (2005), which was calculated by an area-weighting approach. The supporting literature and success of the hybrid approach described in Chapter 4 has established that area-weighting is a poor method by which to redistribute census population to a higher resolution, at least at the European scale. The work presented here has, thus, advanced this area of research. It has clearly demonstrated that it is possible, with the appropriate input data, to generate well validated, high resolution population density maps for a large geographic area using dasymetric mapping techniques. The models should also be validated in more areas where independent data exist; perhaps in a region which might contain unique land cover classes not included in the EU-wide model (e.g. the Mediterranean or northern countries).

Despite the results showing that the country models outperformed the EU-wide models, they were not used to generate the final population density map due to the presence of step changes along national borders. A potential reason for the step changes could be that the land class definitions were inconsistently applied across Europe. Changes or improvements in the updated CLC2000, however, may have helped address this issue. For this reason, and because the models should be extended to include Sweden and the ten new Member States, the analysis should in future be repeated in full. As such, both the country level and EU-wide models should be rerun using updated input data (i.e. population, light intensity and land cover data for 2000) to produce a 1 x 1km population density map for the EU-25. This need not include a repeated analysis based only on land cover (i.e. no-light models) as the inclusion of light intensity information clearly improved the models in reflecting small-area variations in population density. In the absence of suitable input data for the new Member States, future research should first focus on determining whether recalibration of the existing model would suffice for updating the EU-15 population density map.

Reliable information on the distribution of the population is essential in many ecologic studies in order to calculate population-level exposures. While total population has been modelled here, there is often the need for sex and age stratified population counts to facilitate exposure assessments on subsets of the population. Although not provided as part of the SIRE database, this information is readily available as part of the census for most, if not all, countries in Europe. Without additional ancillary data one could, at best, evenly distribute the age and sex

stratified population across the 1km grid according to a model derived for the total population. Refinements to the stratified 1km population predictions may be possible, for example, by the incorporation of information on socioeconomic position if it was available at a resolution higher than that of the census tract.

7.1.2 Affinity zone stratification

Aim: To assess the spatial coverage of the European air pollution monitoring network as a basis for air pollution mapping and exposure assessment.

Affinity zones were originally described by McGregor (1996) as areas of similar air pollution. As was demonstrated in his work, affinity zones can be defined on the basis of the temporal and spatial variation in air pollution measured at the monitoring sites. Affinity zone stratification (AZS), focusing on the spatial variation in the AIRBASE-e monitoring network, was undertaken in Chapter 5. The AZS methods employed statistical techniques to define areas likely to be similar in air pollution based on their underlying physiographic features and land use characteristics. The use of affinity zones to provide a means of extrapolating data from monitoring stations to the wider environment was also explored.

AZS involved a multi-stage process. The first stage involved an exploratory analysis to optimally select the input variables upon which to cluster the input data, while the second stage was concerned with the selection of the most appropriate clustering algorithm. A somewhat subjective approach was taken, and it was decided *a priori* that reasonably distributed, sized and shaped clusters representing real-world geographic features was important. This analysis, therefore, involved an element of trial and error especially in identifying the most suitable input variables for clustering. Each of the iterations involving modifications to the input variables, however, could not be assessed until the data were actually clustered and viewed as a map. Using K-means CA, for example, meant that a range of solutions had to be run and the optimal number of clusters identified before the clusters could be interpreted and finally mapped.

Similar challenges in defining the most appropriate stratification were also faced by authors undertaking ecological stratification. A general theme of these papers, in the words of Bernert et al. (2003), is that the “classification of natural resources is

an abstraction, and as such the classification strategy should be optimised for a given specific application.” Carter (1997) also points out that the pragmatic solution is either to choose a manageable number of clusters or to select the spatial pattern that best reflects the specific objectives of the broader research activities. Similar to the AZS methodology, his particular aim to derive a suitable stratification for soil fertility management strategy for Western Kenya involved a subjective, iterative approach in selecting the appropriate number of clusters. In another example Briggs & France (1983) used two criteria, one statistical and one practical, to decide when to stop the clustering procedure. The practical-based criterion reflected the need to define a manageable but reasonably exclusive number of clusters for the stratification of landscapes and habitats in South Yorkshire, UK. Several authors also point out that careful attention must be paid to the initial selection of input variables as they can affect the success of the classification (Briggs & France 1983; Bunce et al. 1996a). If a particular feature is over-represented, for example by including several variables that represent the same underlying phenomena, it will dominate the resulting clusters. On the other hand, variables that are inadvertently excluded could result in clusters that do not represent reality. A balance, therefore, must be struck in comprehensively identifying the potentially important input variables without placing too much or inappropriate emphasis on any particular input. The overall purpose of the stratification must always guide these decisions. Taking this advice, a final 10 cluster solution satisfying the *a priori* criteria was eventually established.

The main drawback of AZS relates to the practical challenges of the procedure. Running repeated clustering algorithms on 2.8 million cases, 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. It should also be noted that the AZS approach is based on the assumption that the most appropriate division of the landscape and environmental features is achieved. 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 terms of updateability, therefore, 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.

AZS involved ANOVA techniques to assess whether the zones differed in pollutant concentrations as expected. This analysis for some clusters, however, was limited by the number of monitoring sites within each landscape cluster. A sufficient number of sites were available to run ANOVA on eight background and four traffic NO₂ clusters, as well as five background and four traffic PM₁₀ clusters. Three out of four of the highly populated clusters (i.e. metropolitan centres, high density residential and low density residential), along with the industrial clusters were included in each of these analyses. The suburban cluster, however, only had a sufficient number of monitoring sites to be included in the background NO₂ analysis. In most cases, the rural and transport-related clusters contained too few monitors for inclusion in the ANOVA. *Post hoc* tests demonstrated that many of the NO₂ background sites were significantly different, confirming, at least for this pollutant, that the clusters generally represented air pollution affinity zones. The determination of affinity zones from the landscape clusters can be easily repeated or updated as new pollution measurements become available or as new monitoring sites are added to the network. This would simply involve adding the updated concentration data to the site coverage, intersecting the sites with the landscape clusters, and re-running the ANOVA analysis.

In the final stage of AZS, the affinity zones were used as a basis for extrapolating data from monitoring stations to the wider environment. Simple maps of NO₂ and PM₁₀ background concentrations were produced by assigning the mean pollution concentration, derived from training sites within the affinity zones, to all 1km grid cells in each respective affinity zone. This approach, however, falsely assumes that the pollution surface within each zone is flat and completely ignores the within-zone variability. Applying more sophisticated spatial modelling techniques, in this case kriging, within the affinity zones showed promise in deriving more realistic pollution maps. Affinity zones in general offer a framework within which the monitoring data can be extrapolated to other, unmonitored, areas. They could, for example, also be used to assess exposures in different types of environments (e.g. city centre, busy streets, or suburban areas) in association with time-activity data. In addition to mapping concentrations, a potentially more important application of the affinity zones might be to use them as basis for assessing the representativeness of the monitoring network. This issue is discussed in the following section.

Evaluating the monitoring network

Air pollution monitoring objectives are typically assessed using a network of monitoring sites. Because they represent point locations and air pollution exhibits spatial variability, the monitoring sites are generally representative of only a small area immediately around the monitor. Without data enhancements, therefore, monitoring sites may only provide reliable measures of exposure for populations residing in the near vicinity. The challenge in designing monitoring networks, therefore, is in the optimal positioning of monitoring sites (often limited in number due to costs) to reflect the variability in air pollution for the area of interest. When concerned with monitoring the human health effects of air pollution the population must also be taken into account. The effectiveness of routine monitoring networks in assessing exposures, however, is often limited by the design which is usually aimed at compliance, requiring monitors to be situated where local maxima are likely to occur (Lie et al. 1986).

Many studies have demonstrated the benefit gained in using some kind of modelling or interpolation to improve the spatial resolution of monitored pollution data (Briggs 2005; Jerrett et al. 2004; Liu & Rossini 1996). The best estimates would undoubtedly be derived from a 'representative' monitoring network. In the context of exposure assessment, two key elements determine representativeness:

1. The network should contain a sufficient number of monitors to provide suitable spatial coverage to facilitate air pollution modelling with minimal prediction errors.
2. The monitoring sites must be located in or near the areas of interest, which in this case is where the majority of the population reside.

Much research has focused on the issue of how optimally to site air pollution monitors. In designing networks several influencing factors must be considered including the size of the study area, location and number of pollution sources, topography, site features, and security of and access to the equipment (Dunn & Kingham 1996). Most studies relate to positioning monitors to assess intra-urban air pollution, a few examples of which are outlined here.

Modak & Lohani (1985) describe a methodology using spatial correlation analysis (SCA) and minimum spanning tree (MST) to optimally position the most appropriate number of monitors in Taipei City, Thailand. SCA explores the dependencies or

overlaps between monitoring sites by calculating the cross-correlation coefficient between them. The locations are dependent if the cross-correlation coefficient between two sites is greater than the cut-off correlation coefficient. MST then uses the cross-correlations and area of influence to unravel the optimal configuration for the stations. This method depends on the choice of the cut-off correlation coefficient and, as one would expect, as the value is increased the optimal network density also increases. Research by Lie et al. (1986) explored the figure-of-merit (FMO) and sphere of influence (SOI) concepts. Their procedure for determining the optimal network is based on an air quality simulation model used to generate meteorological scenarios representative of the study area. To rank the most desirable locations, FMO is calculated for a specific time and location by summing the products of the pollutant concentrations and the associated probabilities of occurrence for the meteorological scenarios. The network configuration is then determined using the SOI which is dictated by a cut-off spatial correlation coefficient between concentrations at the monitoring locations. The minimum number of monitoring stations is determined by deleting lower-ranked stations with SOIs that overlap the SOIs of higher-ranked stations. While the methods of both Modak & Lohani (1985) and Lie et al. (1986) successfully identify the optimal number of monitors and configuration so that variance is explained for a maximum area of the region at a minimum possible overlap, they do not take account of the population distribution.

An approach by Kanaroglou et al. (2005) incorporated the distribution of the at-risk population in the design of an intra-urban pollution network to monitor NO₂ exposure in Toronto, Canada. The underlying concept of the approach is that areas with higher variability are sampled more intensively. First a continuous 'demand' surface is created for the study area, where areas with higher demand have an increased need for monitoring. Demand is based on two criteria: the spatial variability of the pollution surface; and the population density. In their application, the existing government monitoring network was used to generate the initial pollution surface to evaluate demand. The spatial variability of pollution at each location is determined by an equation similar to that of the semivariogram. Next the demand surface is modified by assigning greater weight, using bivariate linear rescaling (Equation 7.1), to areas with high densities of the population at risk. Finally, a location-allocation procedure in ArcInfo (i.e. attendance maximizing problem) is used to optimally locate the monitors across the candidate locations.

$$\text{Equation 7.1} \quad W_R = \frac{P_R / P_T}{\hat{\gamma}_R / \hat{\gamma}_T}$$

Where: P_R is the population of interest in region R within the study area; P_T is the same population for the entire study area; and γ is the spatial variability.

The location-allocation approach developed by Kanaroglou et al. (2005) was designed for use in epidemiological studies in Toronto to explore associations between traffic-related pollution and health outcomes ranging from childhood asthma to lung cancer mortality. For similar studies assessing exposure to air pollution at the city or regional level, this approach to locate purpose-designed monitoring sites is very promising. At the scale of the EU, however, purpose-designed monitoring is not likely to be an option and epidemiologists must rely on the existing monitoring network.

The location of monitoring stations was considered in establishing the site selection criteria for EUROAIRNET which are outlined in Table 3.4. This network, however, comprises monitoring stations from existing networks around Europe and each government would have undoubtedly used their own criteria to locate the sites. Affinity zone stratification has shown that monitoring sites not only represent point locations or the immediately surrounding area, but they actually represent areas with similar landscape elements and, in theory, similar air pollution (at least for background NO_2). In the absence of information about the sphere of influence for each of the monitoring sites in EUROAIRNET, the affinity zones thus potentially offer a way forward in assessing the representativeness of the monitoring network for exposure assessment. Preliminary results of such an assessment are presented in the remainder of this section.

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 7.1. Marked differences are found in site density both by area and population. In terms of area/site, there is an under-representation of rural areas, especially the forested uplands, and while variations in density are less obvious, suburban populations appear to be somewhat under-represented. This suggests that additional sites need to be operational in these areas, especially as they are likely to be important residential areas. Monitoring generally appears more evenly and densely distributed for NO_2 compared to PM_{10} . In the case of PM_{10} , under-representation is especially

evident in the case of rural populations, suburbs and major road corridors. Examination of the people/site values for the densely populated zones shows that suburban areas appear to be under-monitored compared to the high and low density residential and metropolitan centre zones.

Table 7.1 Site representativeness 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

The affinity zones were further divided into 50km blocks to provide more detailed view of the network coverage. This involved re-sampling the affinity zone grid, for each of the 10 zones, to a 50 x 50km grid. For each affinity zone the population density and number of monitoring sites by type (i.e. background and traffic analysed separately) was then summed in each 50km block. For each affinity zone, Table 7.2 shows the percentage of 50km blocks that are represented by one or more monitoring sites. According to this measure, the affinity zones with the highest population densities (i.e. zones 2 to 4) also tend to have the greatest percentage of blocks represented by monitoring sites. Once again, the suburban zone seems to be under represented.

Table 7.2 Percent 50km blocks containing NO₂ monitoring sites
By affinity zone

#	Affinity Zone	% 50km blocks with NO ₂ sites	
		Background	Traffic
1	Mixed rural	8.8	1.0
2	High density residential	7.6	10.7
3	Low density residential	18.5	9.6
4	Metropolitan centres	20.2	32.0
5	Forested uplands	3.4	0.2
6	Motorways	2.3	0.9
7	Major road corridors	2.9	1.1
8	Motorway intersections	1.1	2.6
9	Suburban	3.7	1.8
10	Industrial	8.5	2.8

The number of background NO₂ monitoring sites by population density for the 50km blocks for each of the affinity zones is presented in Figure 7.1. Only box plots for the NO₂ background sites are shown as the general trends for both pollutants and site types were similar (see Appendix E1 for the full set of box plots). In general, the rural areas and major road corridors show no relationship between population density and number of monitoring sites. For the remaining zones, however, there tends to be a positive association in that the number of monitoring sites increases as population density increases.

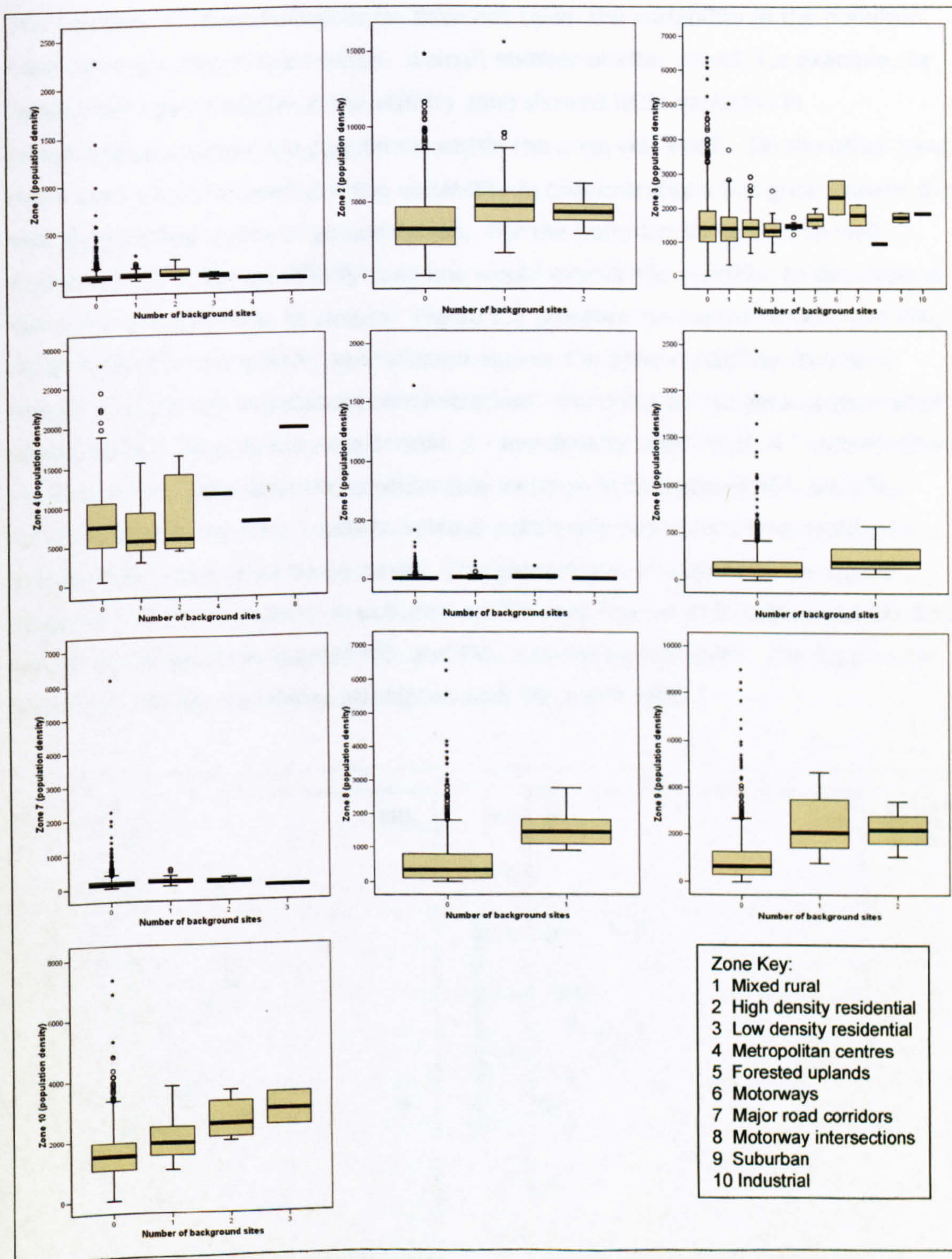


Figure 7.1 NO_2 sites by affinity zone population density (50km blocks)
Background sites

The comparisons presented thus far have not taken the variability in the pollution concentrations into consideration. A small number of sites would, for example, be considered representative if the affinity zone showed little variation in concentrations and/or the population within the zone was small. On the other hand, more sites would be needed if the variability in concentrations was great and/or the affinity zone had a very large population. For the pollution surface to be well characterised within an affinity zone one would expect the variation to decrease as the number of sites was increased. Figure 7.2 presents the number of NO₂ and PM₁₀ monitoring sites per affinity zone plotted against the zone variability (standard deviation in $\mu\text{g}/\text{m}^3$) in pollutant concentrations. Focusing on the densely populated zones (zone 2 - high density residential, 3 - low density residential, 4 - metropolitan centres and 9 - suburban) the considerable increase in background NO₂ and PM₁₀ monitoring sites for zone 3 fails to achieve noticeable reductions in variability in comparison to the other three zones. The general lack of pattern in the figure suggests that the variability in pollution within very few zones is sufficiently captured by the coverage of the current NO₂ and PM₁₀ monitoring networks. The figure also shows the greater variability associated with the traffic sites.

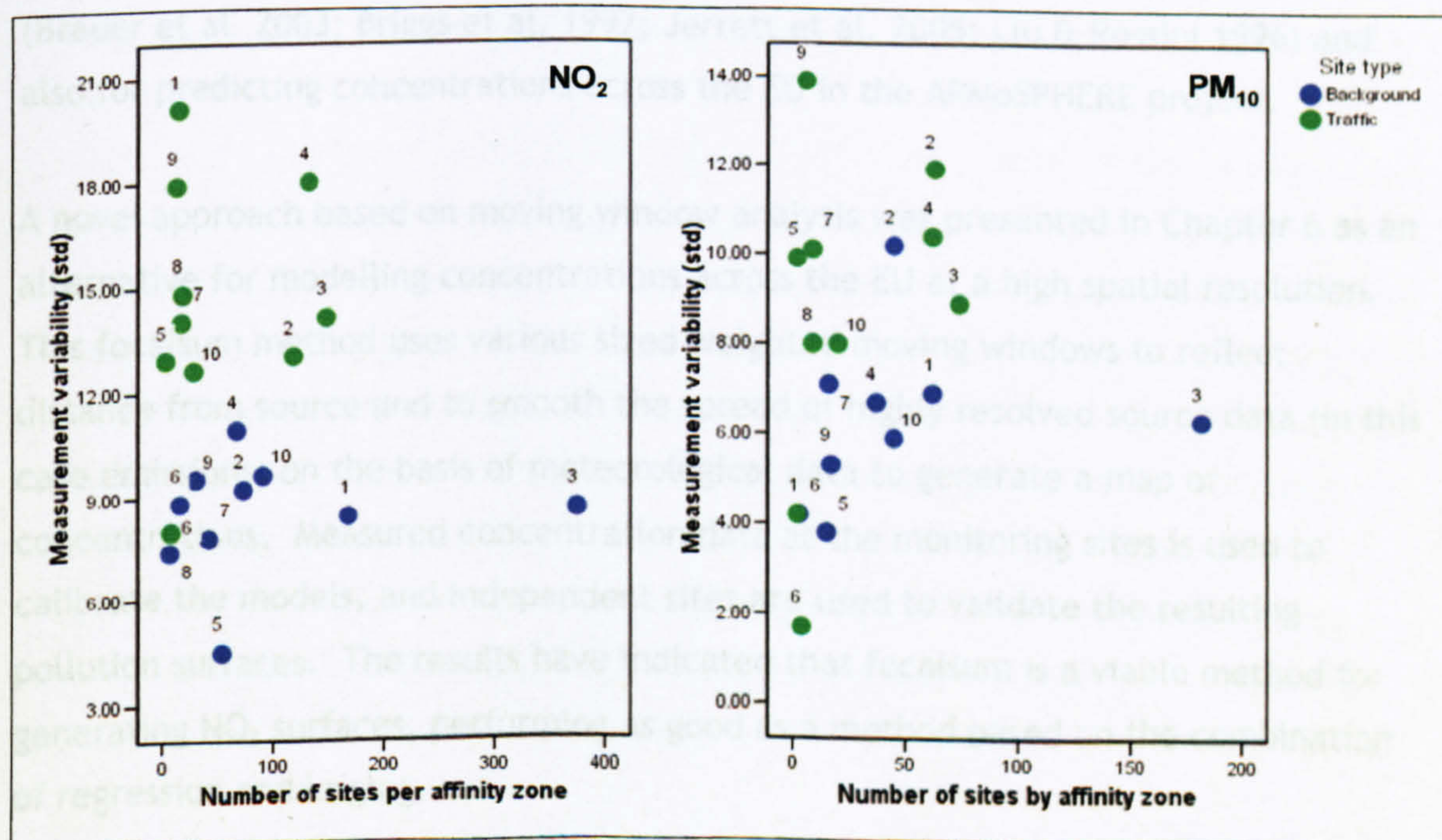


Figure 7.2 Variability of monitoring sites by affinity zone

The way forward in assessing the representativeness of the monitoring network may be to base the analysis on the principles used to build the demand surface in the location-allocation approach outlined by Kanaroglou et al. (2005). The weights could

be calculated for each affinity zone, potentially within areas smaller than the EU (e.g. country or region). Areas in which the density of monitoring sites is lacking could then be identified on the basis of the weights (i.e. larger weights indicating a greater need for additional monitoring).

7.1.3 Pollution modelling

Aim: To develop and apply a novel GIS-based interpolation technique for modelling and mapping small-area variations in annual average PM_{10} and NO_2 concentrations across the EU, and compare the performance with currently accepted methods using a reserved set of validation data.

Current research in the field of air pollution modelling has demonstrated the value of using complex dispersion or time-space models to study exposures. Models at this level of sophistication, however, are not yet feasible for assessing long-term exposure for population studies conducted at the EU-level due to the high data and processing demands. The advantage of land use regression mapping and kriging techniques have been shown by various authors in modelling intra-urban air pollution (Brauer et al. 2003; Briggs et al. 1997; Jerrett et al. 2005; Liu & Rossini 1996) and also for predicting concentrations across the EU in the APMoSPHERE project.

A novel approach based on moving window analysis was presented in Chapter 6 as an alternative for modelling concentrations across the EU at a high spatial resolution. This focalsum method uses various sized weighted moving windows to reflect distance from source and to smooth the spread of highly resolved source data (in this case emissions) on the basis of meteorological data to generate a map of concentrations. Measured concentration data at the monitoring sites is used to calibrate the models, and independent sites are used to validate the resulting pollution surfaces. The results have indicated that focalsum is a viable method for generating NO_2 surfaces, performing as good as a method based on the combination of regression and kriging.

The focalsum method offers several advantages over other regularly used approaches, including those explored in the APMoSPHERE project. First and foremost, the raster data model used in the ArcInfo GRID environment is computationally efficient and the regular shape of the spatial entities enables the

development of complex mathematical models. The raster data model also enables large data sets, such as that used in this study, to be analysed in short processing times. Significant decreases in processing time, therefore, are possible compared to alternate approaches. While the methods explored in the APMoSPHERE project involved approximately one week of processing per pollution surface, the same surface can be generated with the focalsum method in a matter of hours.

The second advantage to the focalsum method is that the model is built using explicit source data rather than proxies such as distance to road. The reliance on emissions may, however, be problematic if high quality and resolution data are difficult to obtain. The resulting concentration surface generated by this approach was shown to be limited by the quality of the input data. This was apparent in attempts to model PM_{10} which failed largely due to the poor quality of the available emissions data in some countries. In areas where local or national emissions grids are readily available, however, the focalsum method potentially provides a powerful approach to mapping air pollutant concentrations. Instead of using highly resolved emissions data, source information could be derived directly from other data sets such as point sources, population density or traffic density. The focalsum method can also be readily applied to different spatial scales, depending on the availability of the input data.

A further advantage is the ease of updating offered by the focalsum approach. As already stated, the focalsum models are quick and easy to run given readily available input grids (i.e. emissions or other source data in raster format) and point location and concentration data from the monitoring network. Even if the models need to be recalibrated each time as new input data are available, updated concentration surfaces can be quickly generated within the GRID environment. Since emissions data are routinely updated, therefore, it would be possible to remodel pollution patterns using the focalsum approach, and recalibrate and validate the maps with updated air pollution monitoring data.

As described in Chapter 6, the models were validated using an independent set of concentration data (i.e. 25% of the total sites not used in the model building). The validation sites were selected by randomly stratifying all monitoring sites by pollutant, site type, site location and country. Using this stratification, the sites were then randomly assigned to the training or validation set so that the ratio of

75:25 was maintained across these parameters (described in Section 3.3.5). Several measures of performance (i.e. R^2 , standard error, root mean square error, factor-of-two and fractional bias) calculated at the validation sites were used to compare the resulting focalsum pollution surfaces with those obtained from the APMoSPHERE project reference method. Based on these parameters, performance of the focalsum method for modelling background NO_2 was equivalent to the reference method. The NO_2 surface derived from the focalsum method was considered the better model as the predictions were within the range of concentration values upon which the model was built, and lower level variation was apparent in the maps. In the SAVIAH project, the land use regression models calibrated with 80 monitoring sites each for three European cities had an R^2 ranging from 0.79 to 0.87 (Briggs et al. 1997). Compared to the SAVIAH results, the focalsum performance ($R^2 = 0.61$, SEE $6.588 \mu\text{g}/\text{m}^3$) is respectable considering the small number of monitoring sites and vast area over which the predictions were made. Mapping PM_{10} and SO_2 with the focalsum approach, however, was less successful.

PM_{10} concentrations tend to vary over larger distances of several to tens of kilometres (Spengler & Wilson 1996). Customisation of both the window size and distance decay functions specifically to address the nature of PM_{10} variability should be explored. Or, following the example of Collins (1998a), coupling focalsum with another modelling method may be the way forward. PM_{10} is a difficult pollutant to model due to the chemical reactions creating secondary particulates, such as nitrate and sulphate. While the coarse fraction of PM_{10} decreases with distance, the fine fraction actually increases due to this conversion. Secondary particulates generally have a life-time in the atmosphere in the order of days to weeks and can travel distances from 100s to 1000s of kilometres (Spengler & Wilson 1996). The approach by Collins (1998a), which involved the use of kriging to model the global trend to which was added the local component derived from focalsum, is potentially useful in addressing this problem. Further research, therefore, is needed to incorporate and model regional factors that influence the long-range transport of particulates.

While focalsum was able to produce better validated SO_2 model compared to the reference method, the resulting surface was poor and not suitable for use in exposure assessment. The major source of SO_2 emissions in the EU is currently industrial point sources. The location of the industrial point sources were obvious in the resulting surface (map not shown) which indicates that dispersion from these

point sources is not accurately represented by the focalsum model in its current form. This is perhaps mainly because dispersion takes place at a considerable height and over wider distances than allowed for in the focalsum model. With winds of 5m/s, gases emitted from a 200m high stack, for example, can travel 200km before the spreading plume reaches the ground (Spengler & Wilson 1996). For these point sources, therefore, the focalsum method would need to be refined to take account of long distance dispersion patterns. Not only should the window size be enlarged or elongated, improvements may also be facilitated by incorporating better information on meteorology (e.g. mixing height and temperature), as well as details on stack height and emission temperature. While these data were not available for inclusion in this work, it is possible they could be sourced and incorporated into future refinements of the focalsum methodology. Alternatively, it may be feasible to use dispersion models to estimate concentration patterns resulting from a sample of these specific sites.

The approach to date was applied to an aggregated emissions grid, representing the total emissions from all SNAP categories. As stated above, the size of the windows and the weights reflecting the distance decay function should be customised for each pollutant. Opportunity also exists to explore customisations for the different emission sources which are represented by individual grids (C1 to C6, Table 6.2). Smaller windows with steep gradients are likely appropriate for traffic emissions (Briggs et al. 1997; English et al. 1999; Gilbert et al. 2003), while much larger windows in which a Gaussian distribution is used as the distance decay function could be more appropriate for point sources (Spengler & Wilson 1996). Area sources may be somewhere in between.

This research has demonstrated that the moving window approach is a robust and efficient method for generating high resolution maps of background NO₂ across Europe. In summary, the approach should be extended and refined in order to better model other pollutants such as PM₁₀, SO₂, CO or ozone. Based on these preliminary results, modelling SO₂ clearly presents an opportunity but its success might depend on more detailed information of stack height and the availability of higher resolution (both spatially and temporally) wind data. Likewise, the approach may need to be coupled with another modelling method to reflect long-range transport in order to successfully model PM₁₀. Finally, the focalsum approach should be validated against

a widely accepted dispersion model for a smaller area where suitable data are available to feasibly run both methods.

7.2 Exposure assessment

Aim: To assess European exposure to annual mean concentrations of PM₁₀ and NO₂, on the basis of the pollutant and population density maps, and identify potential populations exposed to above guideline pollution levels.

7.2.1 Methods

To evaluate potential population exposure to long-term air pollution levels in the Europe, the 1 x 1km pollutant maps were intersected with the population map in the GIS. Each grid square in the resulting map was, thus, represented by a measure of total population and pollutant concentration. Exposure classes of 5 µg/m³ bands were defined to facilitate mapping and the generation of aggregate statistics (i.e. total population per exposure class).

This exposure assessment was conducted using the NO₂ focalsum map (Figure 6.7) and the PM₁₀ reference map (Figure 6.12) to derive measures of exposure. The results of the exposure assessment are presented in Sections 7.2.2 and 7.2.3 for NO₂ and PM₁₀, respectively.

7.2.2 European exposure to NO₂

Using 5 µg/m³ bands, thirteen NO₂ exposure classes, ranging from <5 to >60 µg/m³, were defined. The percent population within each exposure class was then calculated in the GIS (shown in Figure 7.3). The figure shows that the majority of the EU population were well below the annual NO₂ guideline limit of 40 µg/m³, and that only six percent were exposed to above guideline background NO₂ concentrations in 2001. For comparison, exposures based on the nearest background monitor were also computed which resulted in a doubling (i.e. from 6 to 12%) of the persons potentially exposed to above guideline concentrations (Figure 7.4).

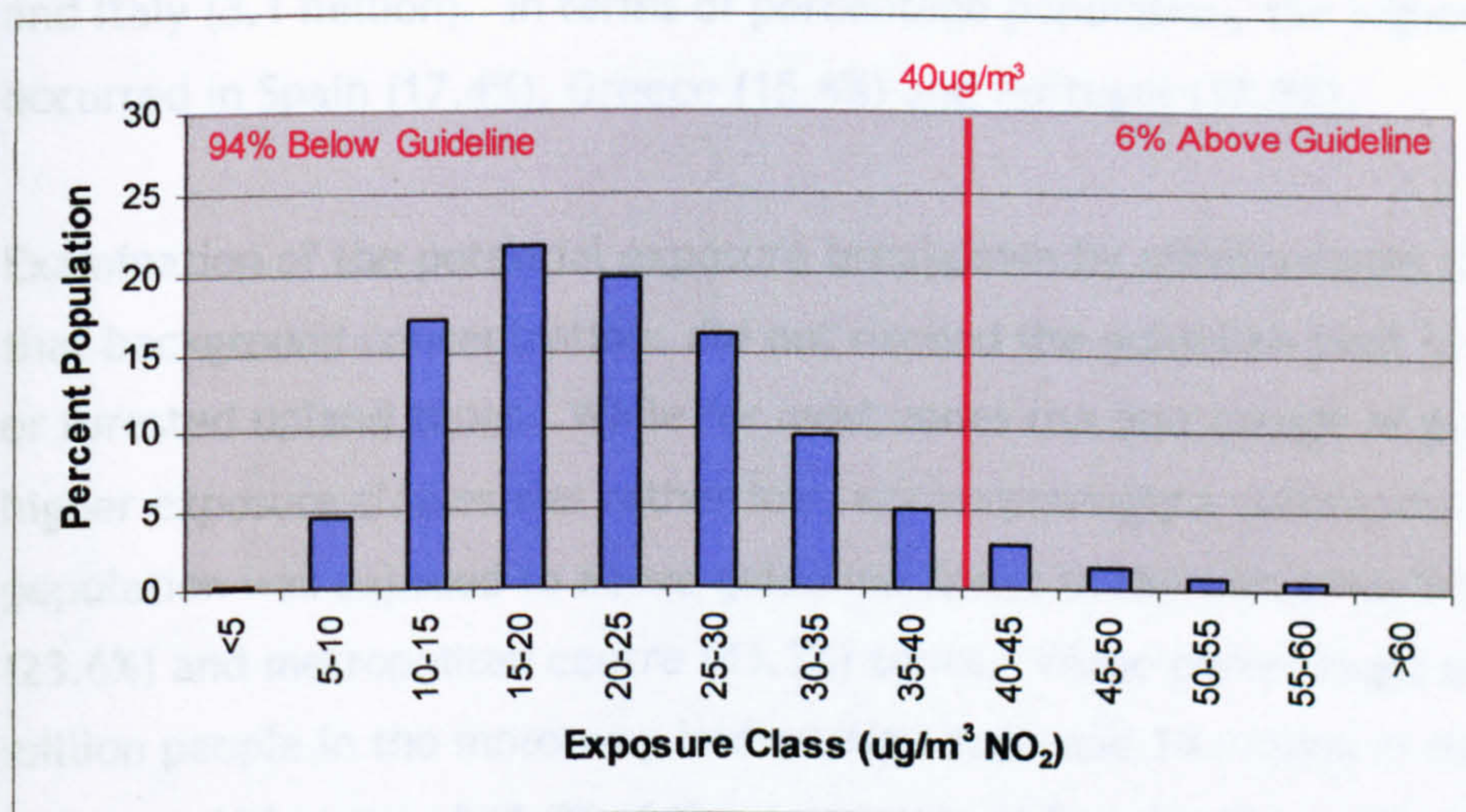


Figure 7.3 Exposure to NO₂ concentrations (µg/m³) - modelled surface
2001 annual mean concentrations - Focalsum method

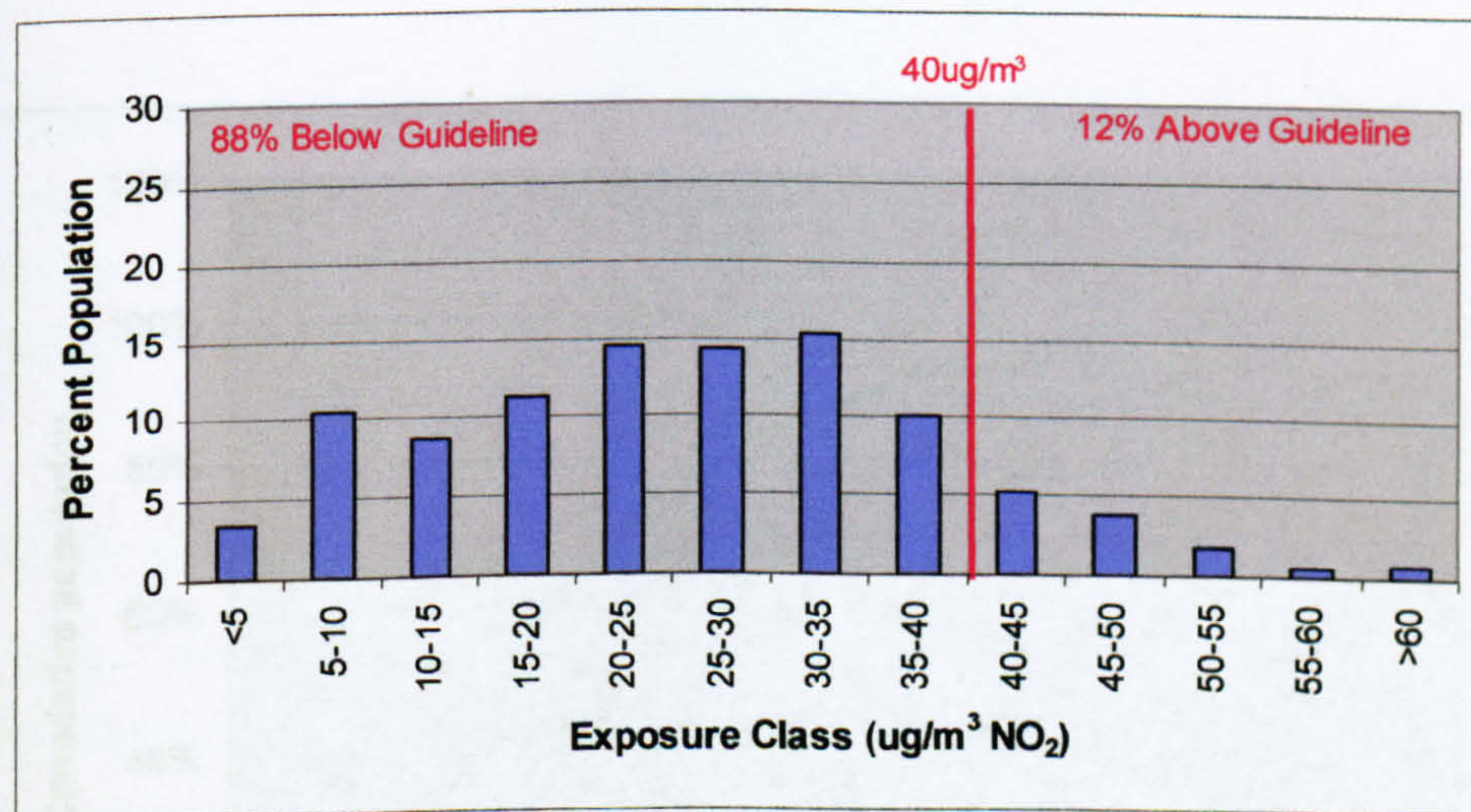


Figure 7.4 Exposure to NO₂ concentrations (µg/m³) - nearest monitor
2001 annual mean concentrations

The geographical breakdown of potential exposure to background NO₂ concentration was also examined by country and affinity zone in Figures 7.5 and 7.6, respectively (Tabular results are presented in Appendix E2, Tables E2.1 and E2.2). Figure 7.5 shows that the distribution of exposures is relatively consistent across the countries. All countries, except Luxembourg, had at least a small percentage of people residing within the higher exposure classes. Greater than 1.3 million inhabitants in each of six countries (Spain, France, Italy, Greece, the UK and Portugal) were exposed to above guideline levels of background NO₂ concentrations in 2001. The highest number of exposed individuals occurred in Spain (6.9 million), France (4.6 million)

and Italy (3.1 million). In terms of percentage population, the highest exposures occurred in Spain (17.4%), Greece (15.6%) and Portugal (12.9%).

Examination of the potential exposure breakdown by affinity zones (Figure 7.6) shows that background concentrations did not exceed the guideline limit in the mixed rural or forested upland zones. While for most zones the percentage of population in the higher exposure classes was rather low, not surprisingly a substantial part of the population was exposed to above guideline levels in the motorway intersection (23.6%) and metropolitan centre (43.3%) zones. These percentages equate to 1.4 million people in the motorway intersection zone and 14 million in metropolitan centres. Although only 1.7% of the population in low density residential areas were exposed to high pollution levels, this translates to approximately 2.2 million inhabitants across the EU.

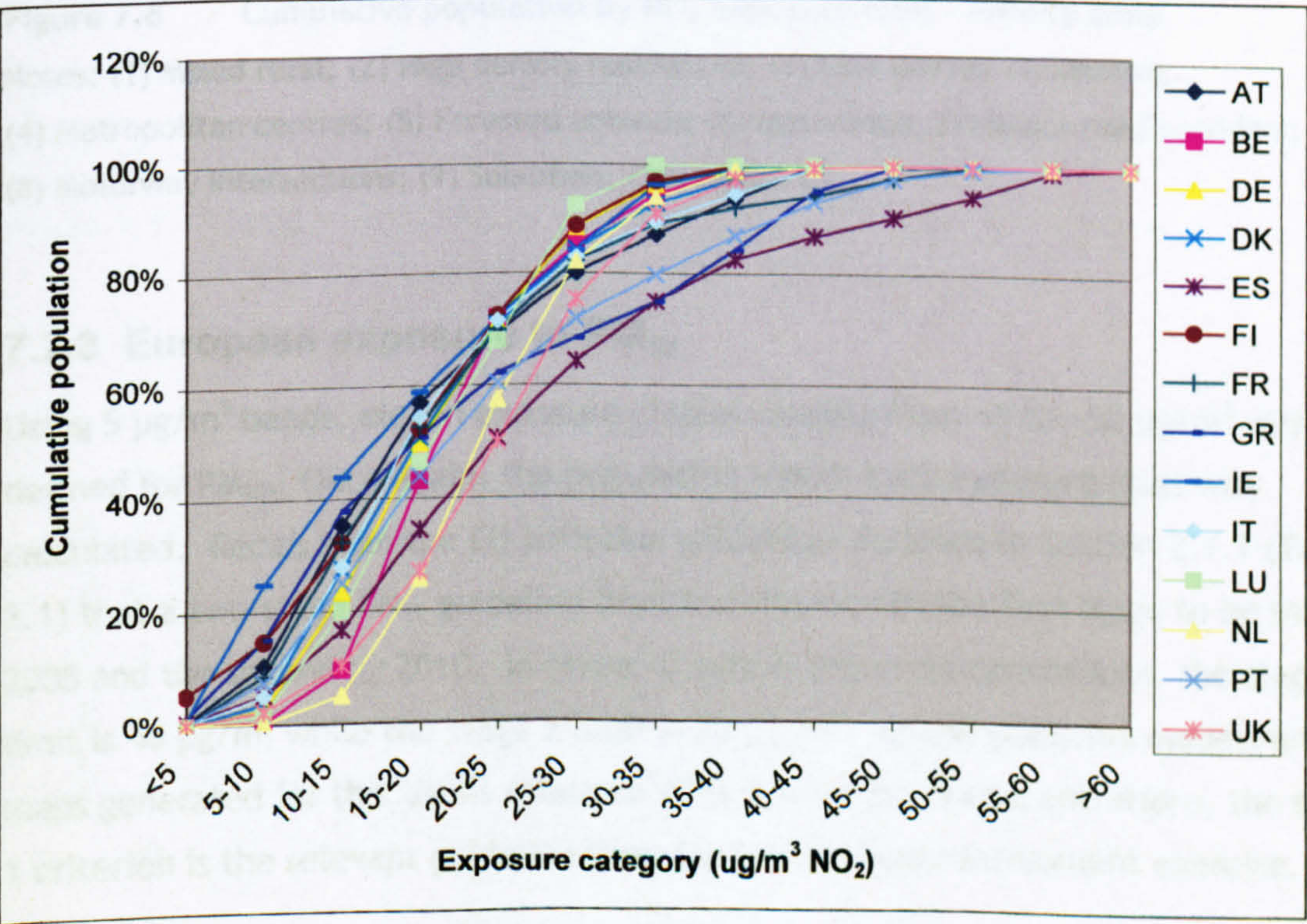


Figure 7.5 Cumulative population by NO₂ exposure class - country

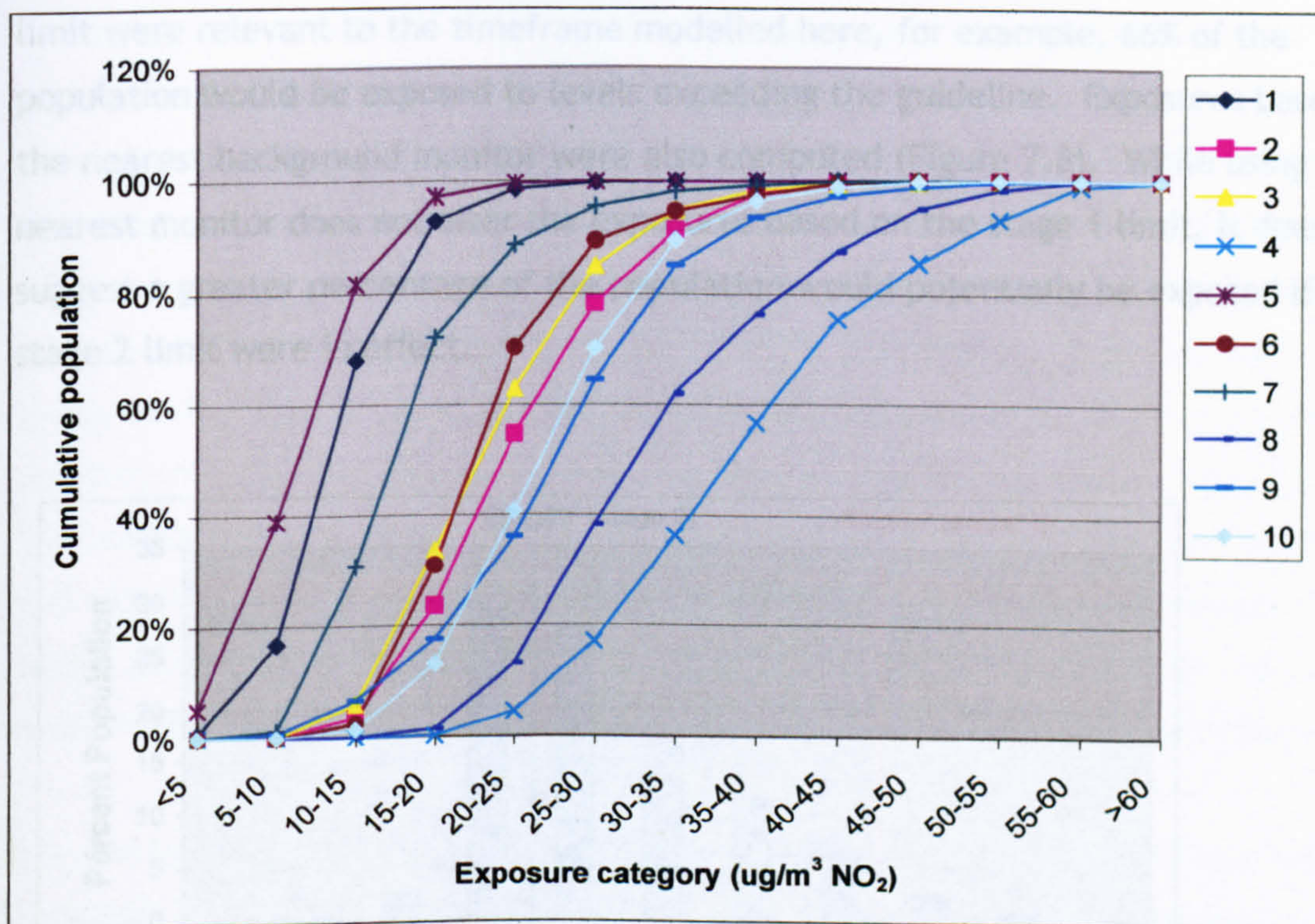


Figure 7.6 Cumulative population by NO₂ exposure class - affinity zone
 Notes: (1) Mixed rural; (2) High density residential; (3) Low density residential;
 (4) Metropolitan centres; (5) Forested uplands; (6) Motorways; (7) Major road corridors;
 (8) Motorway intersections; (9) Suburban; (10) Industrial

7.2.3 European exposure to PM₁₀

Using 5 µg/m³ bands, eleven exposure classes ranging from <5 to >50 µg/m³ were defined for PM₁₀. Once again, the population within each exposure class was calculated. Recall from the EU pollution guidelines outlined in Section 2.1.1 (Table 2.1) that a two-stage PM₁₀ guideline limit is defined with the first stage to be met by 2005 and the second by 2010. In terms of annual mean concentrations, the stage 1 limit is 40 µg/m³ while the stage 2 limit is 20 µg/m³. As the pollution models and maps generated for this thesis relate to 2001 annual mean concentrations, the stage 1 criterion is the relevant guideline limit for this exposure assessment exercise.

Figure 7.7 displays the distribution of EU population within each of the 5 µg/m³ exposure classes for PM₁₀, modelled by the reference method. According to the stage 1 limit only 4% of the EU population were exposed to above guideline limit in 2001 on the basis of the calculations derived from the reference method. The stage 2 limit was also included in the figure, as a benchmark, to highlight the extent to which PM₁₀ reductions are required to meet the more stringent guideline in 2010. If the stage 2

limit were relevant to the timeframe modelled here, for example, 66% of the population would be exposed to levels exceeding the guideline. Exposures based on the nearest background monitor were also computed (Figure 7.8). While using the nearest monitor does not alter the exposures based on the stage 1 limit, it does suggest a greater percentage of the population would potentially be exposed if the stage 2 limit were in effect.

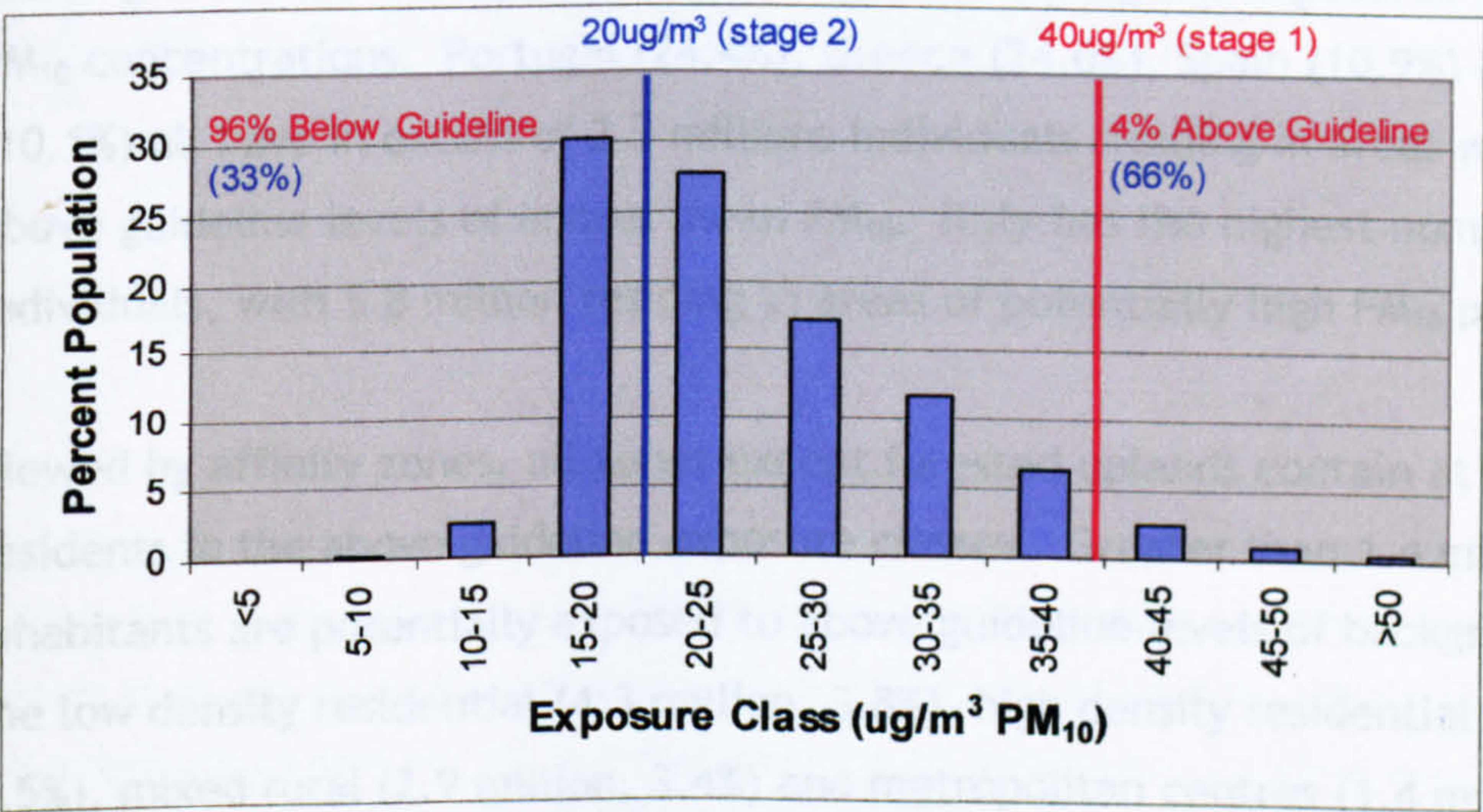


Figure 7.7 Exposure to PM₁₀ concentrations (µg/m³) - modelled surface
2001 annual mean concentrations - Reference method

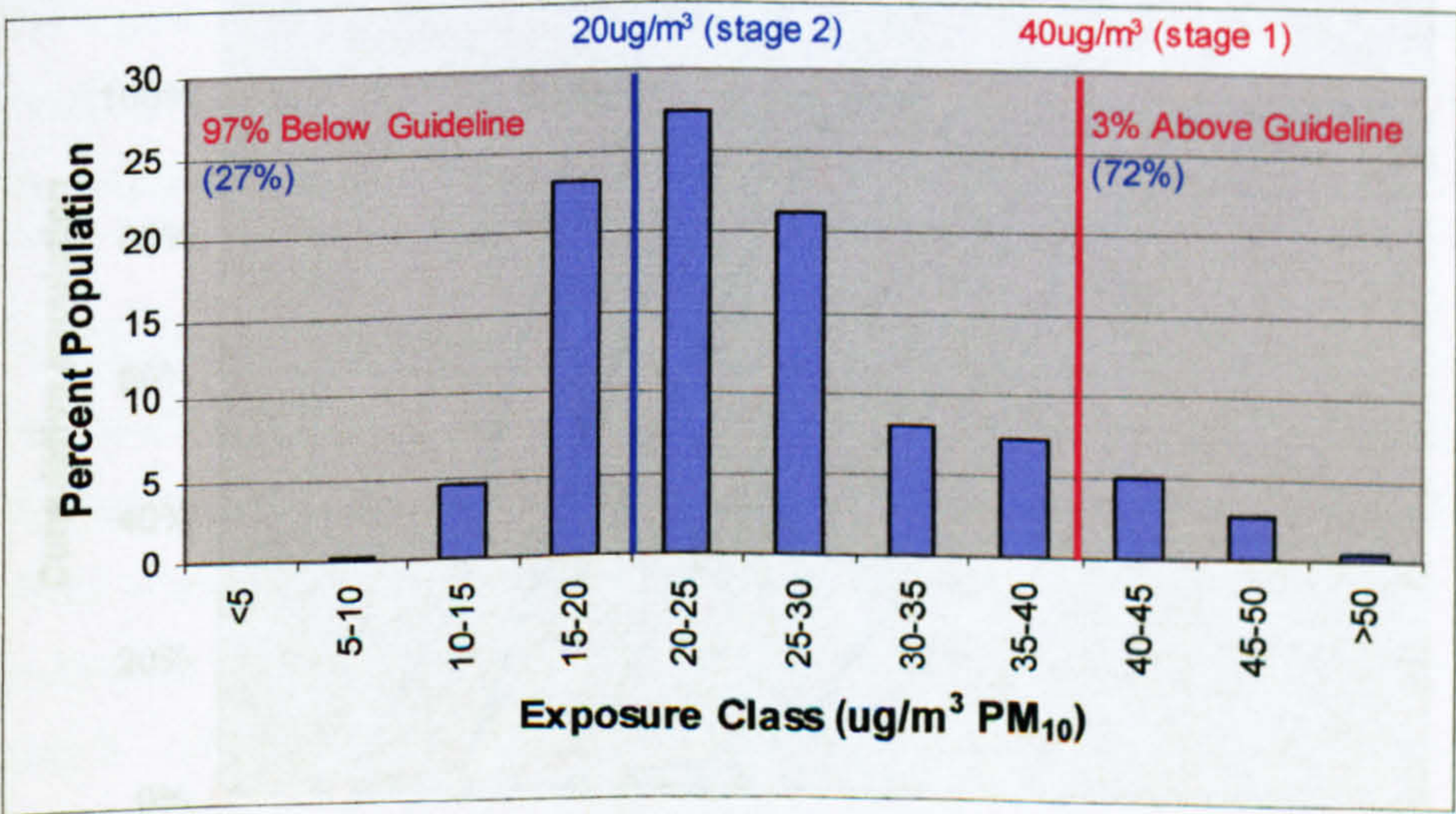


Figure 7.8 Exposure to PM₁₀ concentrations (µg/m³) - nearest monitor
2001 annual mean concentrations

The geographical breakdown of exposure to background PM₁₀ concentrations by country and affinity zone are presented in Figures 7.9 and 7.10, respectively (Tabular results are presented in Appendix E2, Tables E2.3 and E2.4). The breakdown is only discussed in relation to the relevant stage 1 guideline limit. While the distribution of NO₂ exposure was shown to be relatively consistent across countries, this is not the case for PM₁₀. In general, exposures are lower in the northern countries and greater in the south. Most of the countries with the greatest proportion of the population residing in areas with high NO₂ levels also have the highest exposures to background PM₁₀ concentrations. Portugal (24.4%), Greece (24.0%), Spain (10.9%) and Italy (10.1%) all have in excess of 2.3 millions individuals residing in areas with potentially above guideline levels of annual mean PM₁₀. Italy has the highest number of exposed individuals, with 5.8 million residing in areas of potentially high PM₁₀ pollution.

Viewed by affinity zones, all zones except forested uplands contain at least some residents in the above guideline exposure classes. Greater than 1.4 million inhabitants are potentially exposed to above guideline levels of background PM₁₀ in the low density residential (4.3 million, 3.8%), high density residential (3.2 million, 9.5%), mixed rural (2.9 million, 3.4%) and metropolitan centres (1.4 million, 5.1%) affinity zones.

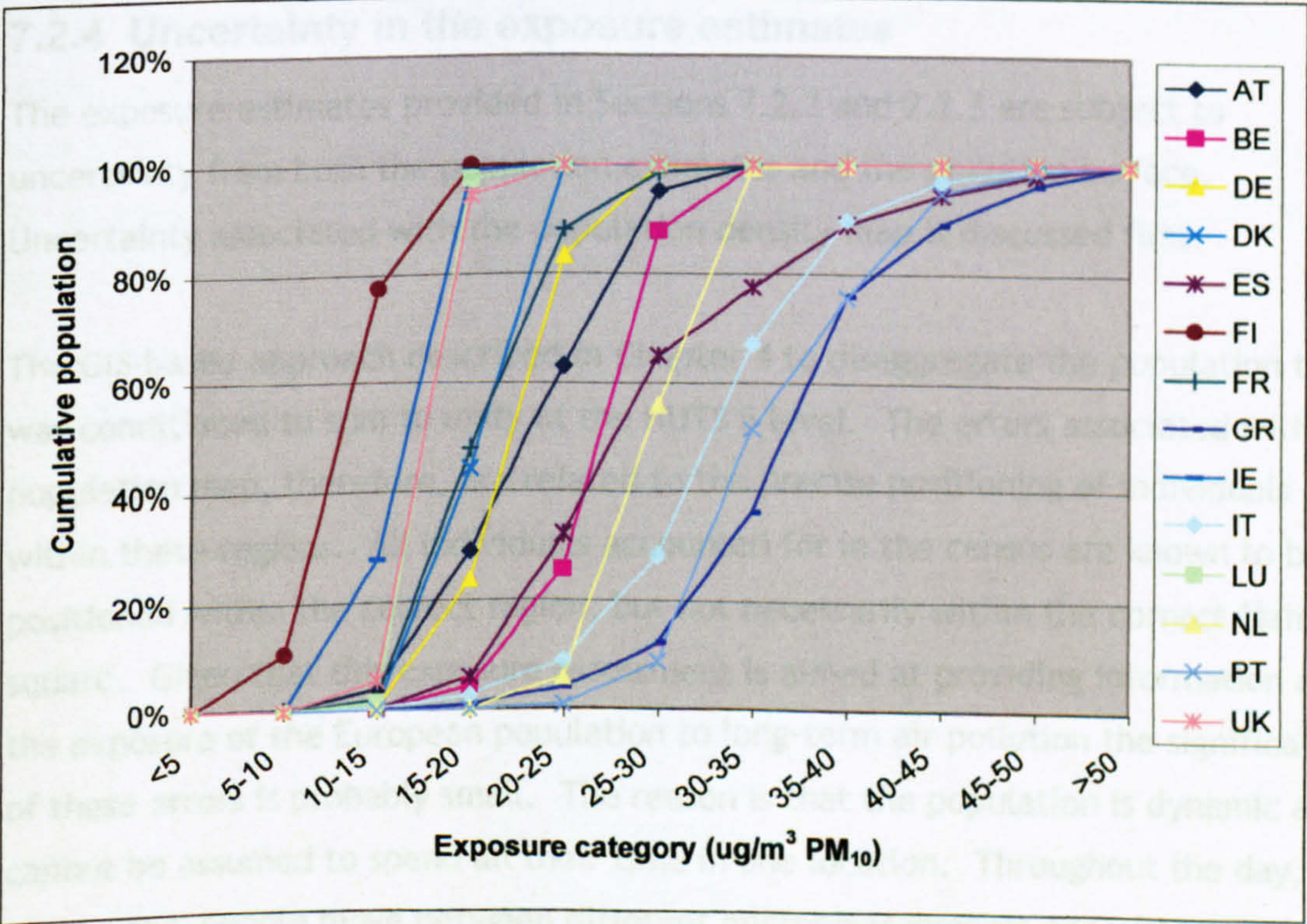


Figure 7.9 Cumulative population by PM₁₀ exposure class - country

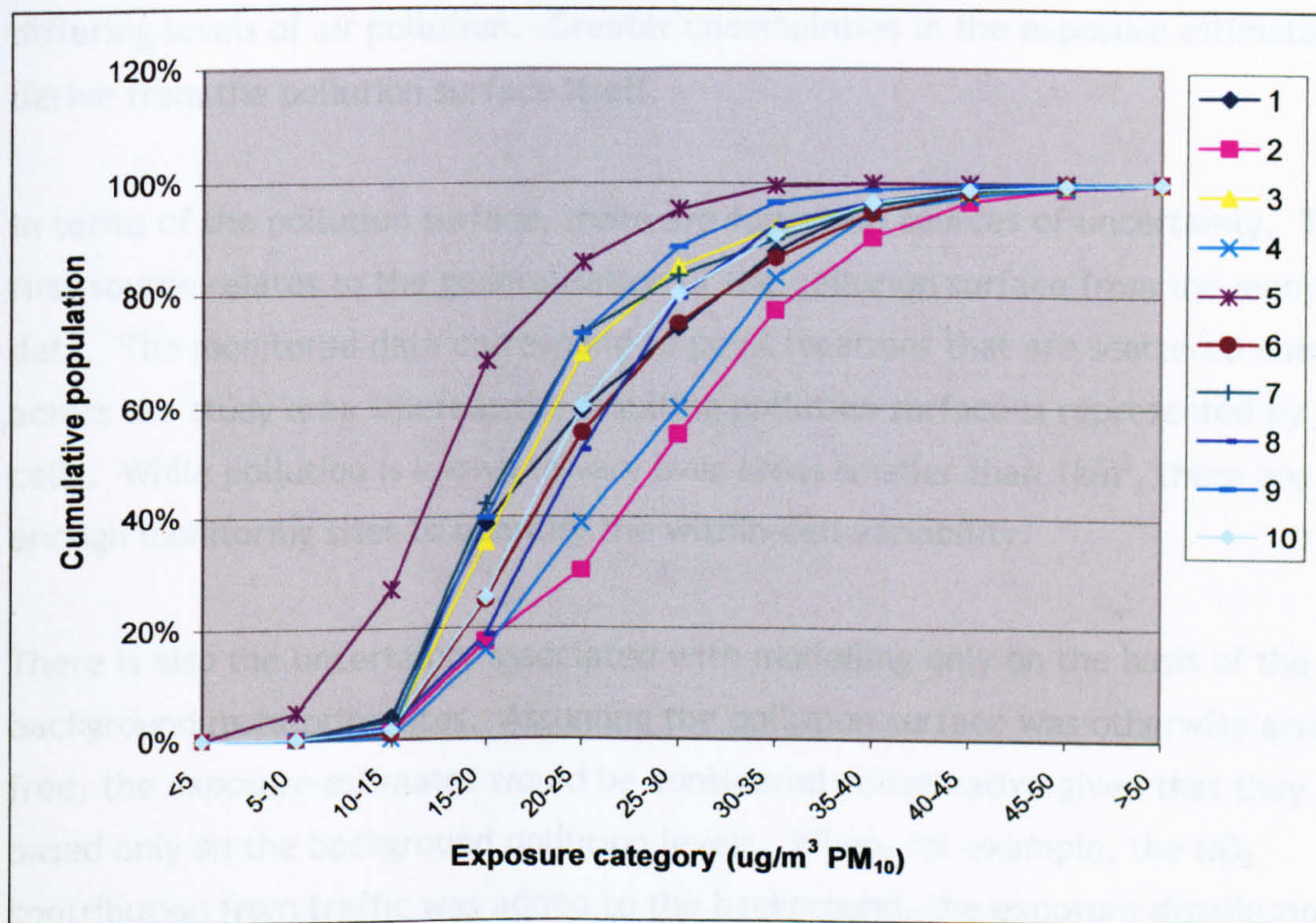


Figure 7.10 Cumulative population by PM₁₀ exposure class - affinity zone

Notes: (1) Mixed rural; (2) High density residential; (3) Low density residential; (4) Metropolitan centres; (5) Forested uplands; (6) Motorways; (7) Major road corridors; (8) Motorway intersections; (9) Suburban; (10) Industrial

7.2.4 Uncertainty in the exposure estimates

The exposure estimates provided in Sections 7.2.2 and 7.2.3 are subject to uncertainty from both the population estimates and the pollution surface. Uncertainty associated with the population density map is discussed first.

The GIS-based approach described in Chapter 4 to disaggregate the population totals was conditioned to sum to unity at the NUTS 5 level. The errors associated with the population map, therefore, are related to the precise positioning of individuals within these regions. All individuals accounted for in the census are known to be positioned within the correct region, but not necessarily within the correct 1km square. Given that this exposure assessment is aimed at providing information about the exposure of the European population to long-term air pollution the significance of these errors is probably small. The reason is that the population is dynamic and cannot be assumed to spend all their time in one location. Throughout the day, let alone year, people move between different microenvironments encountering

differing levels of air pollution. Greater uncertainties in the exposure estimates will derive from the pollution surface itself.

In terms of the pollution surface, there are four main sources of uncertainty. The first source relates to the generalisation of the pollution surface from the monitored data. The monitored data correspond to point locations that are scattered unevenly across the study area whereas the resulting pollution surface is represented by 1km cells. While pollution is known to vary over areas smaller than 1km², there are not enough monitoring sites to quantify the within-cell variability.

There is also the uncertainty associated with modelling only on the basis of the background monitoring sites. Assuming the pollution surface was otherwise error free, the exposure estimates would be considered conservative given that they are based only on the background pollution levels. When, for example, the NO₂ contribution from traffic was added to the background, the exposure distribution shifted so that over twenty percent of the EU population were shown to be exposed to above guideline levels (Figure 7.11). This result should be interpreted with caution as only a crude method, based on a poor linear relationship, was used to add a constant NO₂ concentration to reflect traffic emissions across the EU (see Section 6.3.2). It does, however, point out the importance of developing models that are as comprehensive as possible to ensure that exposure misclassification be minimised. If models are based only on background concentrations the exposures in some areas may be considerably under predicted. Models including all sources of pollution, such as the specific component from traffic-related and industrial sources, should be investigated in a more detailed manner.

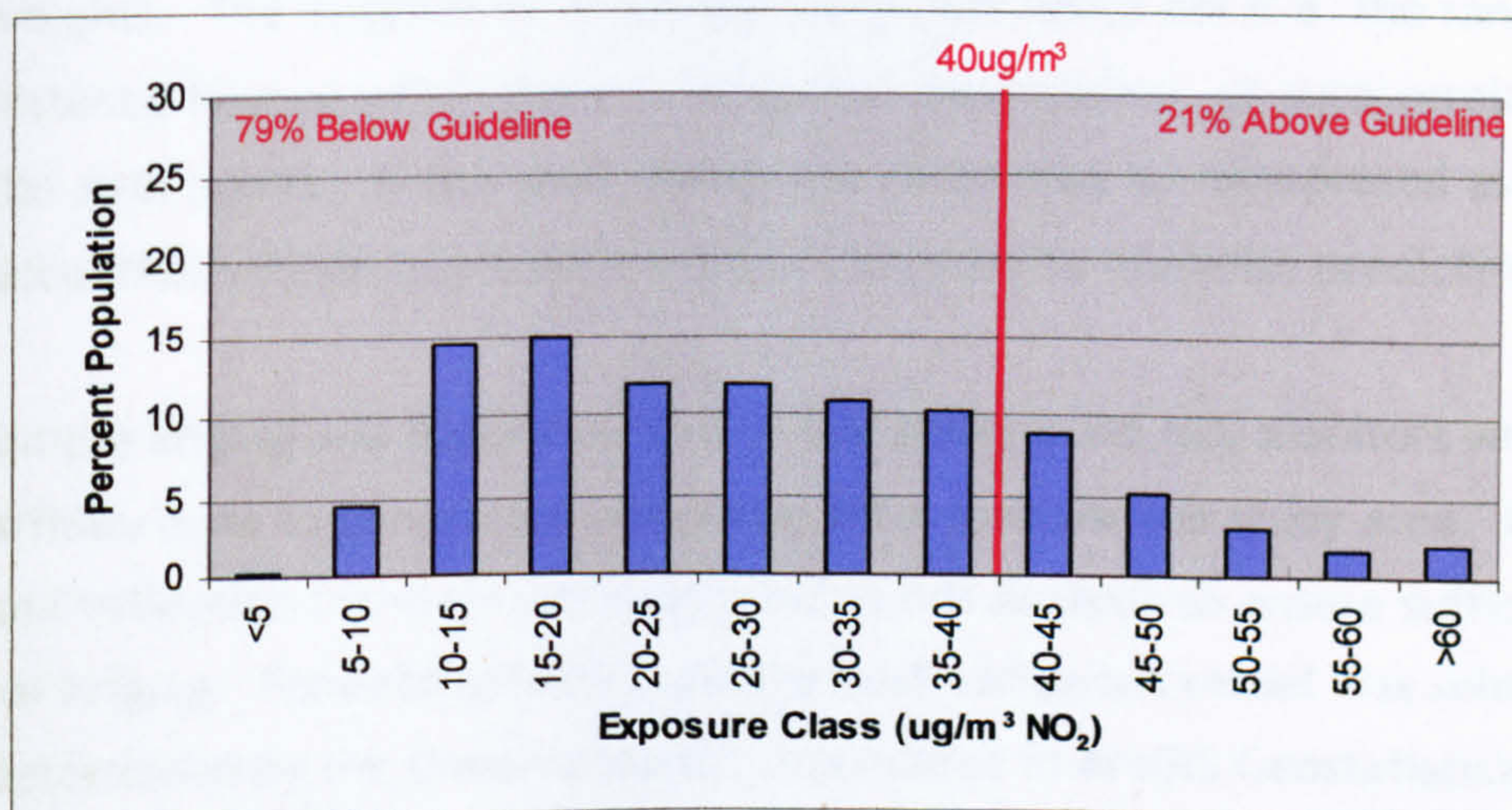


Figure 7.11 Population exposure to NO_2 concentrations ($\mu\text{g}/\text{m}^3$) - traffic
2001 annual mean concentrations - Focalsum method, includes background and traffic components

The third source of uncertainty associated with the pollution surface relates to the assumption that a single model applied across the EU is appropriate. This uncertainty was investigated by calculating the residual concentrations (i.e. modelled - measured concentration) at each of the NO_2 validation sites for the focalsum model. In Figure 7.12 the positive residuals representing over-predictions are shown in blue while the negative residuals are shown in green. From this map it is difficult to distinguish a clear pattern in the residuals. Most countries show both over and under-predictions to varying degrees. The largest over-predictions occurred at a site in Onda, Spain ($18.5 \mu\text{g}/\text{m}^3$), followed by sites in Berlin and Hamburg and Athens. Several sites in Austria and France also had high over-predictions. Approximately 60% of the highest over-predictions were at urban background sites with the remaining at rural background sites. The largest under-prediction occurred at a site in Glasgow ($26.8 \mu\text{g}/\text{m}^3$). Some sites in Belgium, Germany, France and Italy were also under predicted by more than $10 \mu\text{g}/\text{m}^3$. The under-predictions were evenly split (50%) between urban and rural background monitoring stations.

The final source of uncertainty in the pollution surface relates to the spatial distribution of the monitoring sites. The question, one that was partially addressed in Section 7.1.2, is 'how well positioned are the monitors to enable predictions across the vast area of the EU?' Geostatistical methods, which rely on spatial autocorrelation, are well suited to provide a basic answer to this issue. As was described in Section 2.3.1, kriging uses the variogram to derive the interpolation

weights. The distance at which the variogram levels off (i.e. the range) is the distance beyond which there is no spatial dependence, or autocorrelation, between the data points. In this application, the range may be interpreted as the maximum acceptable distance between monitors in order to minimise prediction errors.

Simple kriging was applied in turn to the background NO₂ monitors within each affinity zone to generate a composite error map for the study area. Both the training and validation monitors were included in this analysis to ensure sufficient numbers for kriging. For each affinity zone the best variogram model was selected as determined by the cross-validation procedures in ArcGIS Geostatistical Analyst. The mean and standard deviation of the kriging errors across the surface were calculated by affinity zone. These are reported in Table 7.3 which also shows the mean concentration at the monitoring sites. Across the study area the predictions errors ranged from 21% up to 53% of the affinity zone mean concentration.

Table 7.3 **Kriging errors by affinity zone - background NO₂**

Affinity Zone		Mean conc. at monitoring sites (µg/m ³)	Variogram range (km)	Error (µg/m ³)		% Error
				Mean	SD	
1	Mixed rural	14.2	926	6.6	1.08	47%
2	High density residential	29.5	564	9.0	0.34	30%
3	Low density residential	25.8	304	7.4	0.56	29%
4	Metropolitan centres	38.3	1268	8.0	2.17	21%
5	Forested uplands	8.5	778	4.5	0.28	53%
6	Motorways	24.5	370	9.0	0.11	37%
7	Major road corridors	16.2	769	7.9	0.13	49%
8	Motorway intersections	33.0	*	*	*	*
9	Suburban	26.3	714	7.5	1.66	29%
10	Industrial	26.4	541	8.9	0.60	34%

Notes: * Insufficient number of monitoring sites

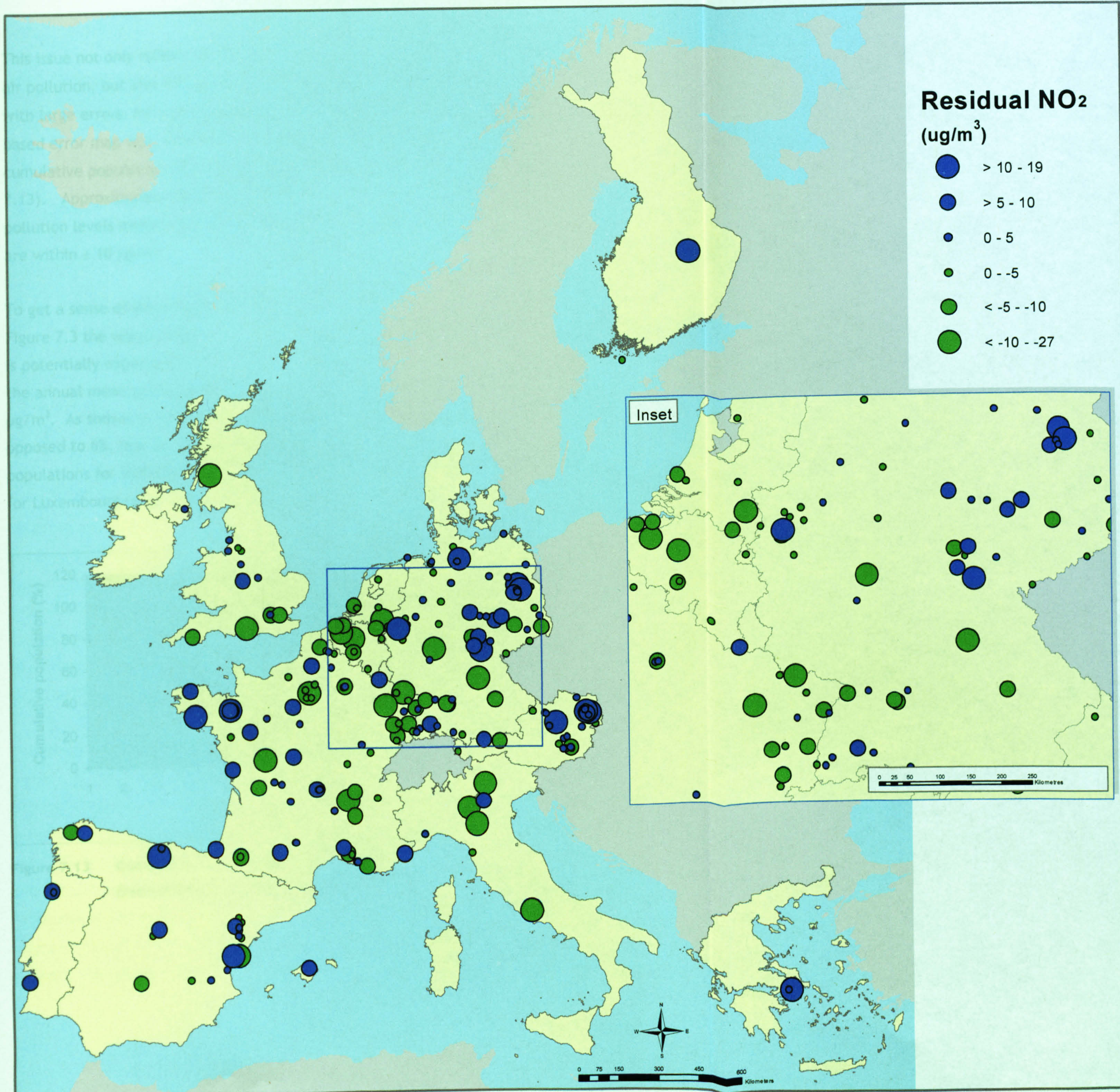


Figure 7.12
Residuals at background NO₂ sites
Focalsum

This issue not only relates to how well the monitors reflect the spatial variations in air pollution, but also the distribution of the population of interest. Are the areas with large errors, for example, densely or sparsely populated? The affinity zone based error map was, therefore, intersected with the population density map and the cumulative population distribution was plotted against the increasing errors (Figure 7.13). Approximately 95% of the population are shown to reside in areas that have pollution levels measured within a certainty of $\pm 9 \mu\text{g}/\text{m}^3$ while approximately 99% are within $\pm 10 \mu\text{g}/\text{m}^3$.

To get a sense of the effect of these errors on the exposure estimates presented in Figure 7.3 the worst case scenario was assumed - i.e. that everyone in the study area is potentially exposed to an additional $10 \mu\text{g}/\text{m}^3 \text{NO}_2$. To account for this uncertainty the annual mean guideline limit would effectively have to be reduced from 40 to $30 \mu\text{g}/\text{m}^3$. As shown in Table 7.4 this would mean that up 20% of the EU population, as opposed to 6%, may be exposed to above guideline concentrations. The exposed populations for individual countries would potentially range from approximately 7.4% for Luxembourg up to 36% in Spain.

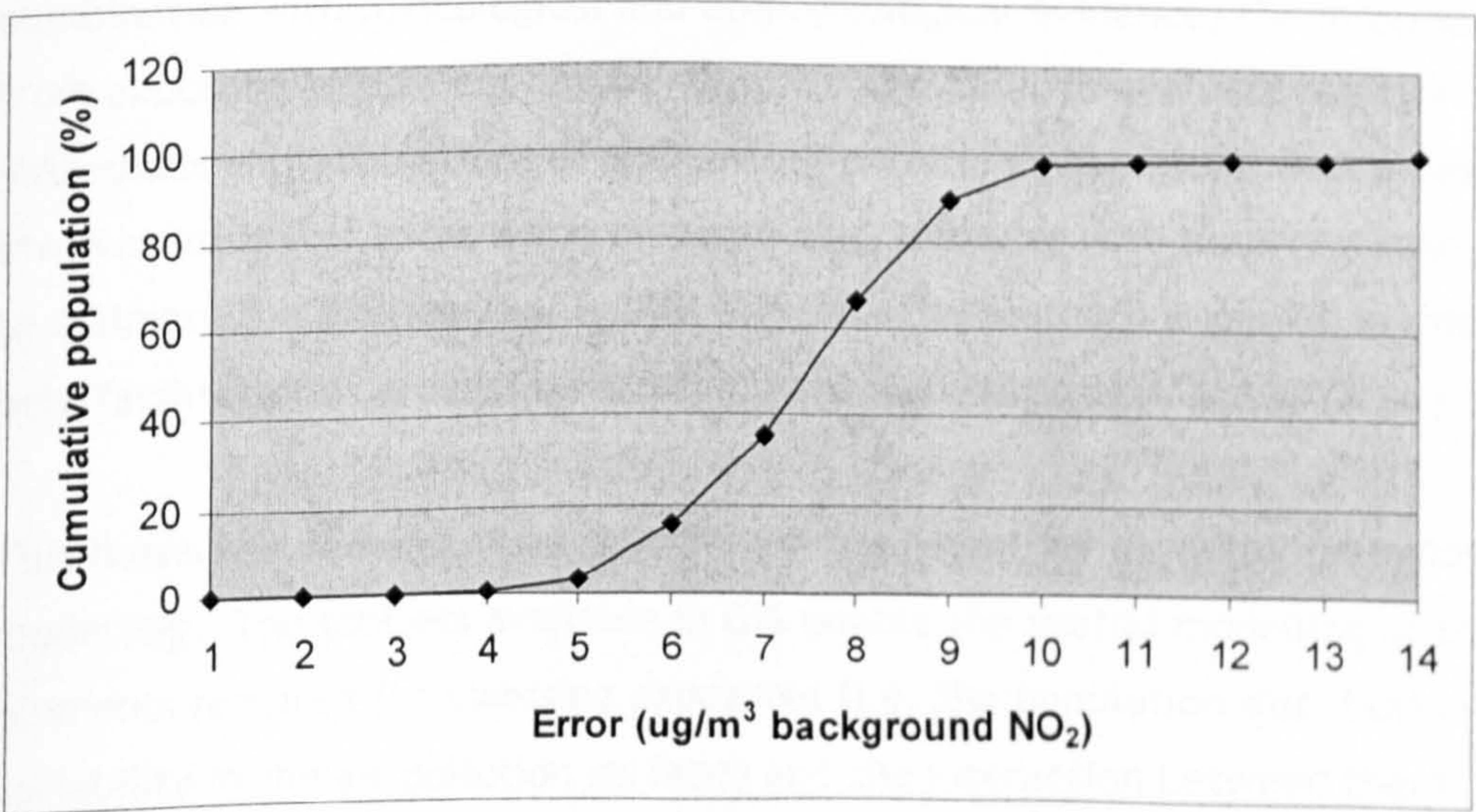


Figure 7.13 Cumulative population distribution by pollution errors
Errors derived from kriging by affinity zone

Table 7.4 Effect of uncertainty on background NO₂ exposure estimates

Country	% population residing in areas with annual mean NO ₂ above	
	40 µg/m ³	30 µg/m ³
AT	6.2	20.0
BE	0.6	14.8
DE	0.9	11.7
DK	2.4	15.7
ES	17.4	35.8
FI	0.2	11.3
FR	7.8	17.7
GR	15.6	31.6
IE	0.6	15.7
IT	5.3	17.1
LU	0.0	7.4
NL	0.8	17.9
PT	12.9	28.3
UK	2.5	24.8
EU	5.8	19.8

7.3 Exposure modelling using GIS

Exposure assessment is an important element in health risk assessment. In combination with toxicological and epidemiological evidence, the information gained from exposure assessment can be used, for example, to evaluate the risk of individuals and populations to ambient air pollution. In general, risk assessment involves using this information on exposures, together with dose-response functions, to estimate the likelihood of health effects. The methods explored in this thesis can help facilitate this process by providing exposure estimates across large populations.

This thesis has demonstrated that GIS are important for air pollution exposure modelling. The toolsets available in GIS enable the spatial modelling of the key elements required for assessing exposures (i.e. the population distribution and variability in the air pollution surface) and the interaction between them. GIS are particularly useful when there is a need to assess risks across large geographic areas and populations while maintaining a resolution high enough to detect variations in exposures at the local scale. The remainder of this discussion outlines how the exposure modelling components developed in this thesis can be used further to evaluate the risks associated with ambient air pollution in the European population.

7.3.1 Place in epidemiology

Epidemiological studies related to air pollution generally show associations between exposure and respiratory and cardiovascular health outcomes. Though the relative risks are typically small, the sheer number of people exposed means that air pollution can result in considerable risk and public health consequences. Most research to date has focused on unravelling the associations between health outcome and day-to-day variations in air pollution, particularly in susceptible groups of the population. Considerably fewer studies are conducted to explore associations between chronic air pollution exposure and health outcomes. Aside from the logistics of organising large study populations (e.g. cohorts) to facilitate this area of research, the main challenge in exposure assessment has been in characterising the spatial variations in air pollution.

The complex spatial and temporal variations in the air pollution surface, especially in urban areas, mean that exposures must be assigned at the local or individual level. The simple exposure models employed in earlier cohort studies, for example using measurements from the nearest monitor (Dockery et al. 1993; Pope III et al. 1995), are clearly inadequate for characterising individual or small-area level exposures. Additionally, a standard method for assigning exposures has not yet been devised - with each of the air pollution cohorts, for example, using a different method. This hampers comparison and pooling of the results from different studies. As was demonstrated in this thesis, GIS offer a framework in which consistent exposure models can be developed. With the advances in GIS computing technology, researchers are able to produce models of the air pollution surface that better reflect the temporal and spatial variations which in turn can reduce exposure misclassification. Furthermore, exposure models developed within a GIS framework can be consistently applied to large numbers of people across vast geographic areas, allowing epidemiologists to undertake cohort or population-based studies over larger populations and areas. This is particularly important should researchers want to compare the results of studies conducted in different regions or countries.

One of the objectives of EUROAIRNET is to allow researchers to undertake air quality assessments at the European scale. As such, the high resolution EU-wide concentration maps developed in this thesis can be of great benefit. Cohort studies, as well as being useful for monitoring health risks, are one type of epidemiological design that can be used to derive estimates of dose-response relationships. Cohorts

covering large geographic areas, however, are rare - with the exception of the ACS II study conducted across 151 cities in the United States (Pope III et al. 1995). As cohort data are time consuming and expensive to collect, the pollution maps derived here can be used with existing European cohorts to look retrospectively for associations between chronic air pollution and various health outcomes. Two studies in particular, the ECRHS (European Community Respiratory Health Survey) and EPIC (European Prospective Investigation into Cancer and Nutrition), are at present using the background NO₂ map developed with the focalsum approach in epidemiological investigations. Given that the exposures will be derived from maps based on methods applied uniformly across the EU, the results of these investigations will be comparable across all of the populations studied.

There are potentially two issues which may restrict the use of the EU-wide pollution surfaces, in their current form, in epidemiological applications: (1) the surface is, at present, only available for a single year so that long-term temporal trends cannot be studied; and (2) the surface only represents background NO₂ concentrations. The first issue is not considered a major limitation as it has been established that, using the focalsum approach, the pollution surface can be reproduced quickly given updated monitoring (and emissions) data. Reported pollution concentrations up to the year 2004 are currently available through AIRBASE. These data simply need to be downloaded, mapped and intersected with the input emissions grids (i.e. index grid), and calibrated to generate updated pollution surfaces for additional years. Incorporating updated emissions data would further enhance the resulting pollution surfaces.

The second issue, mentioned above, may be a greater limitation if the maps will be used to explore the associations between specific sources (e.g. traffic-related pollution) and health outcomes. Future work should, therefore, focus on devising a consistent method by which to incorporate the traffic (and industrial) component of air pollution into the EU background models. In the absence of more comprehensive models, there may be an opportunity to undertake an approach similar to that of Hoek et al. (2002a) whereby a proxy measure was used to assign additional exposure on the basis of local sources. Care must be taken, however, if this approach is used. Researchers would have to ensure that the proxy measures were applied uniformly so that the results of any epidemiological investigation remain broadly comparable with similar studies undertaken across Europe.

Despite this potential limitation, however, the results derived from this work are clearly important to epidemiology, filling a crucial gap by providing a background pollution surface modelled on the basis of consistent data and methods. Undoubtedly the best pollution surfaces will be those that are derived from a network of representative monitoring sites. A non-representative network with respect to population exposures, for example, can result in exposure misclassification and, thus, errors in defining the exposure-response relationships. The area-based affinity zone stratification methods developed here provide useful insight into the distances across which monitoring sites may be considered representative and, generally, in the pollution variability across the EU. They are thus potentially important for further understanding and characterising these variations - an issue identified by Hoek et al. (2002b) as being one of the main challenges in air pollution exposure assessment.

The AZS methodology may also present an opportunity for assisting with the optimal siting of purpose-designed monitoring often conducted in association with epidemiological investigations. Using the AZS approach, for example, targeted studies could be undertaken to explore more comprehensively the issue of representativeness prior to locating the monitors. Focused on the areas where the participants under study reside or spend their time (and from which their ambient exposures would be assigned) AZS could provide a clearer understanding of the potential exposure misclassifications. The delineation of air pollution affinity zones across Europe can also be used to target or focus policy, as discussed next.

7.3.2 The policy need

In order properly to estimate risk, it is crucial to know the number and location of individuals exposed to estimated levels of air pollution. This is particularly important for deciding where policy intervention is most needed, and how dramatic the interventions need to be in order to reduce risks.

The results of the GIS-based exposure assessment presented in Section 7.2 provide useful information about the current air pollution situation across Europe. Based on the EU-wide analysis, only 4% and 6% of the population were potentially exposed to above guideline levels of annual mean PM_{10} and NO_2 , respectively, in 2001. Given the $10 \mu g/m^3$ uncertainty associated with this result, however, the potential population

exposed to above guideline levels of NO₂ would increase from 6 to 20%. Transition from the PM₁₀ stage 1 criterion (40 µg/m³) to the stage 2 criterion (20 µg/m³) in 2010 may also lead to a significant increase in the number of persons residing in areas with above guideline levels of particulate pollution. While this analysis demonstrates that potentially large parts of the European population will continue to be at risk of excess exposures to air pollution, it is not very informative in terms of how to direct policy. Analysis on the basis affinity zones is possibly more revealing. The largest risks, for example, seem to be associated with the densely populated areas, including metropolitan centres, low density residential and the high density residential affinity zones. The motorway intersection and mixed rural zones also seem to be particularly important for NO₂ and PM₁₀ exposures, respectively.

The affinity zones are not only useful for identify areas where the large exposures are likely to occur, they are important for indicating where the monitoring network is too sparse for fully characterising exposures. Many of the affinity zones, for example, have an insufficient number of monitors to characterise pollution using ANOVA techniques - which in itself indicates that they are not adequately monitored. The suburban zone in particular (associated with significant populations) is identified as an area in which the monitoring network is insufficient for characterising exposure to air pollution.

The ease and speed with which air pollution maps can be generated using the focalsum approach has already been expressed. This feature is particularly advantageous to policy makers who may wish repeat risk assessments as new data become available, in order to maintain up-to-date maps for tracking the success of policy interventions. The focalsum approach also provides a way in which policy scenarios can be efficiently tested. How would the risk change, for example, if the emissions from the transport sector were cut in half? The direct link between the emissions and concentrations facilitated by focalsum, along with the efficiency of the approach, allow for this type of scenario to be investigated. Coupled with the population density map, changes in risk can be monitored as different emission-based scenarios are tested. The advantages of GIS, and in particular the exposure mapping components developed here, are thus also useful for policy makers concerned with assessing the health risks of air pollution in Europe.

8 Conclusion

The main motivation for exploring the use of GIS as a basis for exposure modelling in this thesis is the need for better characterisation of exposures to air pollution for risk assessment. The use of GIS has been shown to be beneficial in all aspects of exposure modelling focused on larger populations over vast geographic areas. It can facilitate the estimation of small-area population density, provide an efficient environment in which to model the spatial variations in long-term air pollution, and also allow for the straightforward linkage of the population and pollution data to derive exposure estimates. This has been demonstrated through modelling exposures to annual mean NO₂ and PM₁₀ concentrations for the EU population.

In any GIS-based analysis, database development plays a crucial role. The importance of a reliable, well-built GIS was particularly important in this study where it was the foundation for developing exposure estimates. Careful consideration was, thus, given to the selection and scale of appropriate input data to facilitate the spatial modelling of population density and the air pollution surfaces. GIS techniques were also used to enhance the input data to derive new variables for modelling and mapping.

GIS provided a valuable platform in which to disaggregate NUTS 5 census data to produce small-area population estimates. The hybrid approach involving dasymetric mapping and regression analysis successfully produced a consistent, well validated, high resolution population density map across the EU. This research has highlighted the importance of using ancillary data (such as land cover and light intensity data) to reflect small-area variations in population. With advances in satellite technology and the increasing availability of earth observed (EO) data, population mapping over large areas no longer needs to be limited to simplistic methods such as area-weighting. Additionally, as EO data are theoretically available in real-time, this method can also be used to study time trends in population density.

Through the development of affinity zone stratification, multivariate statistical techniques were explored to define air pollution affinity zones across the EU. The affinity zones provide a division of the monitoring sites across Europe, based on the underlying landscape and physiographic features, and delineate areas within which air pollution models may be applied. In this research, the affinity zones were used

as a basis for mapping long-term NO₂ concentrations with some success. Although only simple models were applied, more sophisticated models can be developed to take advantage of the affinity zone structure. The importance of affinity zones for evaluating the representativeness of the monitoring network for exposure assessment has also been demonstrated. The evaluation of the network showed that the areas across which sites are representative vary by affinity zone. This information is useful for identifying areas where pollution is well characterised and, therefore, where predictions from the modelled pollution surface may be considered broadly reliable. Exploring the monitoring network in association with population density also helps to identify areas where the number of sites is inadequate as a basis for exposure modelling.

The focalsum approach, also developed within GIS, has proven to be an effective and efficient method for modelling long-term air pollution across the EU. Currently, the methodology works best for low-level sources, such as those from traffic, which exhibit relatively simple and local level dispersion characteristics. The performance of this approach for modelling NO₂ matched that of other methods. The main constraint on the use of the focalsum approach, however, is the availability of reliable emissions estimates, at a suitable spatial resolution. The addition of finer scale (both temporally and spatially) meteorology data and customised windows are needed in order to refine the methodology for mapping other pollutants. The efficiency of the method for generating air pollution surfaces may be particularly useful for policy makers interested in evaluating 'what if' scenarios. Modifications to the input grids reflecting reductions in emissions could easily and efficiently be used to show the reductions in risk associated with different policy scenario proposals.

8.1 *Future work*

8.1.1 Population modelling

The data sets that were used as a basis for population modelling span more than ten years: the NUTS 5 population totals for most countries derived from the 1991 census; the land cover data (CLC90) covered a period from the mid 1980s to mid 1990s; and the light intensity map was modelled on 1999-2000 satellite imagery. As more recent data are now available (i.e. the 2001 census and land cover version CLC2000), the population model should be updated and extended to include Sweden. Ideally the

light intensity map should also be updated with newer night-time light imagery and land cover data which would allow the extension of the population model to the ten new Member States of the European Union. In the absence of suitable input data for the new Member States, future research should first focus on determining whether recalibration of the existing model would suffice for updating the EU-15 population density map. Further research in this area of population modelling should also include additional validation of the model in other countries where independent population data can be sourced, and, if possible, the provision of age and sex stratified estimates.

8.1.2 Affinity zone stratification

Further research is needed in the area of AZS to evaluate the concept of mapping air pollution on the basis of affinity zones. The nesting of more complex models - for example, those that incorporate covariate information - into the affinity zones might improve predictions by better characterising the within-zone variability of the pollution surface. In particular regression mapping, which does not rely on spatial structure of the monitoring sites, within the affinity zones should be explored.

It can also be argued that, to make use of affinity zones and to design better monitoring systems, the issue of station representativeness needs to be addressed. This implies both theoretical research, to define more clearly the concept of representativeness and how to measure it, and methodological work to develop the relevant tools. AZS as a method for evaluating the representativeness of the monitoring network should be extended by perhaps using the methods devised by Kanaroglou et al. (2005).

8.1.3 Focalsum modelling

Opportunity exists to improve and enhance the focalsum methodology in several ways, including the use of higher resolution emissions data where available (especially for traffic-related sources). Enhancements may also be possible by incorporating functions to reflect meteorological conditions, especially wind speed and wind direction, and the emission heights and dispersion patterns of different sources. This relates to the need for customised window shapes and weights to represent these differences, and for some pollutants may involve the inclusion of

detailed dispersion profiles to inform these decisions. In addition, temporally and spatially improved meteorology data should be incorporated so that the wind direction and wind speed variables can be calculated on the basis of daily rather than monthly means. Future work should also focus on devising a consistent method by which to model and incorporate the other components of air pollution (i.e. traffic and industrial) into the EU background models. Finally, the possibility of using focalsum in conjunction with other modelling methods to better reflect the long-range transport of particulates should be thoroughly investigated.

8.1.4 The way forward

The methods presented in this thesis are clearly important for characterising risks associated with long-term exposure to air pollution across large populations and geographic areas. This research demonstrates that more sophisticated exposure models are also needed to take account of the time-space interactions between air pollution and the population at risk. In particular, time-activity models, such as that explored in EXPOLIS and STEMS, are an important advancement in devising a more realistic representation of population exposures to air pollution. Despite this new direction in which future air pollution research is moving the efforts undertaken in this thesis remain highly relevant. At present the problems of air pollution continue to pose threats to human health and better knowledge about the potential risks, and better use of that knowledge, remains a top priority.

References

- Abbey, D. E., Colome, S. D., Mills, P. K., Burchette, R., Beeson, W. L., & Tian, Y. 1993, "Chronic disease associated with long-term concentrations of nitrogen dioxide", *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 3, no. 2, pp. 181-202.
- Aldenderfer, M. S. & Blashfield, R. K. 1984, *Cluster Analysis*, Sage Publications, Inc., California. pp. 1-87.
- ASTM International. 2006, *D6589-05 Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance*.
- Atkins, D. H. F. & Lee, D. S. 1995, "Spatial and temporal variation of rural nitrogen dioxide concentrations across the United Kingdom", *Atmospheric Environment*, vol. 29, no. 2, pp. 223-239.
- Bailey, T. C. & Gatrell, A. C. 1995, "Further methods for spatially continuous data," in *Interactive Spatial Data Analysis*, Longman Group Limited, Essex, pp. 208-244.
- Barr, C. J., Bunce, R. G. H., Clarke, R. T., Fuller, R. M., Furse, M. T., Gillespie, M. K., Groom, G. B., Hallam, C. J., Hornung, M., Howard, D. C., & Ness, M. J. 1993, *Countryside Survey 1990 Main Report*, Department of the Environment.
- Bayer-Oglesby, L., Briggs, D., Hoek, G., de Hoogh, C., Janssen, N., Jantunen, M., Sram, R., Sanderson, E., & Urbanus, J. 2004, *Air-pollution and the effects to human health - exposure assessment*, AIRNET Workgroup 1 Report.
- Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen, G., & Jarup, L. 2001, "Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm", *Environmental Health Perspectives*, vol. 109, no. 6, pp. 633-639.
- Bernert, J. A., Eilers, J. M., Sullivan, T. J., Freemark, K. E., & Ribic, C. 2003, "A quantitative method for delineating regions: An example for the western corn belt plains ecoregion of the USA", *Environmental Management*, vol. 21, no. 3, pp. 405-420.
- Boezen, H. M., Van Der Zee, S. C., Postma, D. S., Vonk, J. M., Gerritsen, J., Hoek, G., Brunekreef, B., Rijcken, B., & Schouten, J. P. 1999, "Effects of ambient air pollution on upper and lower respiratory symptoms and peak expiratory flow in children", *The Lancet*, vol. 353, no. 9156, pp. 874-878.
- Bracken, I. & Martin, D. 1989, "The generation of spatial population distributions from census centroid data", *Environment and Planning A*, vol. 21, pp. 537-543.
- Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., & Lewne et, al. 2003, "Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems", *Epidemiology*, vol. 14, no. 2, pp. 228-239.

- Briggs, D. J. 2003, "Environmental measurement and modelling: Geographical information systems," in *Exposure Assessment in Occupational and Environmental Epidemiology*, M. Nieuwenhuijsen, ed., Oxford University Press, Oxford, pp. 55-70.
- Briggs, D. J. 2005, "The role of GIS: Coping with space (and time) in air pollution exposure assessment", *Journal of Toxicology and Environmental Health Part A*, vol. 68, pp. 1243-1261.
- Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebreton, E., Pryl, K., van Reeuwijk, H., Smallbone, K., & van der Veen, A. 1997, "Mapping urban air pollution using GIS: a regression-based approach", *International Journal of Geographical Information Science*, vol. 11, no. 7, pp. 699-718.
- Briggs, D. J., de Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., & Smallbone, K. 2000, "A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments", *Science of the Total Environment*, vol. 253, no. 1-3, pp. 151-167.
- Briggs, D. J., Fecht, D., Gulliver, J., & Vienneau, D. 2006, "Dasymetric modelling of small-area population distribution using land cover and light emissions data", *Remote Sensing of Environment (in press)*.
- Briggs, D. J. & France, J. 1983, "Classifying landscapes and habitats for regional environmental planning", *Journal of Environmental Management*, vol. 17, pp. 249-261.
- Brunekreef, B., Dockery, D. W., & Krzyzanowski, M. 1995, "Epidemiologic studies on short-term effects of low levels of major ambient air pollution components", *Environmental Health Perspectives*, vol. 103 Supplement 2, pp. 3-13.
- Brunekreef, B. & Holgate, S. T. 2002, "Air pollution and health", *The Lancet*, vol. 360, no. 9341, pp. 1233-1242.
- Bunce, R. G. H., Barr, C. J., Clarke, R. T., Howard, D. C., & Lane, A. M. J. 1996a, "Land classification for strategic ecological survey", *Journal of Environmental Management*, vol. 47, no. 1, pp. 37-60.
- Bunce, R. G. H., Barr, C. J., Clarke, R. T., Howard, D. C., & Lane, A. M. J. 1996b, "Special Paper: ITE Merlewood land classification of Great Britain", *Journal of Biogeography*, vol. 23, no. 5, pp. 625-634.
- Bunce, R., Barr, C., Gillespie, M., & Howard, D. 1996c, "The ITE land classification: Providing an environmental stratification of Great Britain", *Environmental Monitoring and Assessment*, vol. 39, no. 1-3, pp. 39-46.
- Burrough, P. A. & McDonnell, R. A. 1998a, "Creating continuous surfaces from point data," in *Principles of Geographical Information Systems*, P. A. Burrough & R. A. McDonnell, eds., Oxford University Press, Oxford, pp. 98-131.
- Burrough, P. A. & McDonnell, R. A. 1998b, "Spatial analysis using continuous fields," in *Principles of Geographical Information Systems*, P. A. Burrough & R. A. McDonnell, eds., Oxford University Press, Oxford, pp. 183-219.

- Burrough, P. A. & McDonnell, R. A. 1998c, "Optimal interpolation using geostatistics," in *Principles of Geographical Information Systems*, P. A. Burrough & R. A. McDonnell, eds., Oxford University Press, Oxford, pp. 132-161.
- Carr, D., von Ehrenstein, O., Weiland, S., Wagner, C., Wellie, O., Nicolai, T., & von Mutius, E. 2002, "Modeling annual benzene, toluene, NO₂, and soot concentrations on the basis of road traffic characteristics", *Environmental Research*, vol. 90, no. 2, pp. 111-118.
- Carruthers, D., Blair, J., & Johnson, K. 2003, *Validation and sensitivity study of ADMS-Urban for London*, Cambridge Environmental Research Consultants Ltd, FM489/R5/03.
- Carter, S. E. 1997, "Spatial stratification of Western Kenya as a basis for research on soil fertility management", *Agricultural Systems*, vol. 55, no. 1, pp. 45-70.
- CBS. 2005. Emissieregistratie. <http://www.emissieregistratie.nl> . Centraal Bureau voor de Statistiek (CBS), Milieu en Natuurplanbureau (MNP), Wageningen Universiteit en Researchcentrum (WUR). Access date 2-8-2006.
- CEN. 1998. Determination of the PM₁₀ fraction of suspended particulate matter - Reference method and field test procedure to demonstrate reference equivalence of measurement methods (EN 12341:1998). <http://www.cenorm.be/catweb/13.040.20.htm> . European Committee for Standardization. Access date 8-1-2006.
- CERC. 2006. Software and Services for the Environment. <http://www.cerc.co.uk/index.htm> . Cambridge Environmental Research Consultants. Access date 26-5-2006.
- Chen, K. 2002, "An approach to linking remotely sensed data and areal census data", *International Journal of Remote Sensing*, vol. 23, no. 1, pp. 37-48.
- CIESIN. 2004. Gridded Population of the World (GWP), Version 3. Center for International Earth Science Information Network (CIESIN), Columbia University, and Centro Internacional de Agricultura Tropical (CIAT). Palisades, NY, Columbia University.
- CIESIN. 2005. Gridded Population of the World and the Global Rural-Urban Mapping Project. <http://sedac.ciesin.columbia.edu/gpw/index.jsp> . Access date 11-4-2006.
- Collins, S. 1998a, *A GIS approach to modelling traffic related air pollution*, The University of Huddersfield. PhD Thesis.
- Collins, S. 1998b, "Modelling spatial variations in air quality using GIS," in *GIS and Health*, M. Löytönen, ed., Taylor & Francis Ltd., London, pp. 81-95.
- Colls, J. 2002, *Air Pollution*, 2nd edn, Spon Press, London and New York.
- Colville, R. N. & Briggs, D. J. 2000, "Dispersion modelling," in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J. C. Wakefield, & N. G. Best, eds., Oxford University Press, Oxford, pp. 375-392.

Colville, R. N., Woodfield, N. K., Carruthers, D. J., Fisher, B. E. A., Rickard, A., Neville, S., & Hughes, A. 2002, "Uncertainty in dispersion modelling and urban air quality mapping", *Environmental Science and Policy*, vol. 5, no. 3, pp. 207-220.

Commission of the European Communities. 1995, *Corine Land Cover - Part 1: Methodology*. Commission of the European Communities.

Cressie, N. 2000, "Geostatistical methods for mapping environmental exposures," in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J. C. Wakefield, & N. G. Best, eds., Oxford University Press, Oxford, pp. 185-204.

de Hartog, J. J., Hoek, G., Peters, A., Timonen, K. L., Ibal-Mulli, A., Brunekreef, B., Heinrich, J., Tiittanen, P., van Wijnen, J. H., Kreyling, W., Kulmala, M., & Pekkanen, J. 2003, "Effects of fine and ultrafine particles on cardiorespiratory symptoms in elderly subjects with coronary heart disease: The ULTRA Study", *American Journal of Epidemiology*, vol. 157, no. 7, pp. 613-632.

de Kluizenaar, Y., Aherne, J., & Farrell, E. P. 2001, "Modelling the spatial distribution of SO₂ and NO_x emissions in Ireland", *Environmental Pollution*, vol. 112, no. 2, pp. 171-182.

Desqueyroux, H., Pujet, J. C., Prosper, M., Squinazi, F., & Momas, I. 2002, "Short-term effects of low-level air pollution on respiratory health of adults suffering from moderate to severe asthma", *Environmental Research*, vol. 89, no. 1, pp. 29-37.

Diem, J. E. & Comrie, A. C. 2002, "Predictive mapping of air pollution involving sparse spatial observations", *Environmental Pollution*, vol. 119, no. 1, pp. 99-117.

Dockery, D. W. 2001, "Epidemiologic evidence of cardiovascular effects of particulate air pollution", *Environmental Health Perspectives*, vol. 109 Supplement 4, pp. 483-486.

Dockery, D. W. & Pope III, C. A. 1994, "Acute respiratory effects of particulate air pollution", *Annual Review of Public Health*, vol. 15, pp. 107-132.

Dockery, D. W. & Pope III, C. A. 1996, "Epidemiology of acute health effects: Summary of time-series studies," in *Particles in Our Air*, R. Wilson & J. Spengler, eds., Harvard University Press, USA, pp. 123-147.

Dockery, D. W., Pope III, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., Jr., & Speizer, F. E. 1993, "An association between air pollution and mortality in six U.S. cities", *New England Journal of Medicine*, vol. 329, no. 24, pp. 1753-1759.

Dore, C. J., Watterson, J. D., Goodwin, J. W. L., Murrells, T. P., Passant, N. R., Hobson, M. M., Baggott, S. L., Thistlethwaith, G., Coleman, P. J., King, K. R., Adams, M., & Cumine, P. R. 2004, *UK Emissions of Air Pollutants 1970 to 2002*, AEA Technology, Netcen.

Dunn, C. & Kingham, S. 1996, "Establishing links between air quality and health: searching for the impossible?", *Social Science and Medicine*, vol. 42, no. 6, pp. 831-841.

EEA. 2005a, *EMEP/CORINAIR Emission Inventory Guidebook - 2005*, European Environment Agency, Technical report No 30.

EEA. 2005b, *The European environment - State and outlook 2005*, European Environment Agency, Copenhagen.

EEA. 2006. Air pollutant. http://glossary.eea.eu.int/EEAGlossary/A/air_pollutant . Access date 19-3-2006.

EMEP. 2003. EMEP Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe. http://www.emep.int/index_facts.html . Access date 12-6-2006.

EMEP. 2004. The EMEP measurement network. <http://www.nilu.no/projects/ccc/network/index.html> . Access date 14-2-2006.

EMEP. 2005. Expert Emissions used in EMEP models. <http://webdab.emep.int/> . Access date 15-3-2006.

English, P., Neutra, R., Scalf, R., Sullivan, M., Waller, L., & Zhu, L. 1999, "Examining associations between childhood asthma and traffic flow using a geographic information system", *Environmental Health Perspectives*, vol. 107, no. 9, pp. 761-767.

ESRI. 2001, *Using ArcGIS Geostatistical Analyst* ESRI, Redlands, CA.

ETC. 2004. AirView: Air quality Visualisation instrument for Europe on the web. <http://air-climate.eionet.eu.int/databases/airbase/airview/index.html> [AirView v3.1]. European Topic Centre on Air and Climate Change Topic Centre of European Environment Agency. Access date 15-6-2006.

Europa. 2002. The CAFE Programme. <http://europa.eu.int/comm/environment/air/ambient.htm> . Access date 1-7-2006.

Everitt, B. 2003, *Cluster Analysis*, 1st edn, Heinemann Educational Books Ltd, London.

Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., & Murray, C. J. 2002, "Selected major risk factors and global and regional burden of disease", *The Lancet*, vol. 360, no. 9343, pp. 1347-1360.

Fairbanks, D. H. K. & Benn, G. A. 2000, "Identifying regional landscapes for conservation planning: a case study from KwaZulu-Natal, South Africa", *Landscape and Urban Planning*, vol. 50, no. 4, pp. 237-257.

Finkelstein, M. M., Jerrett, M., DeLuca, P., Finkelstein, N., Verma, D. K., Chapman, K., & Sears, M. R. 2003, "Relation between income, air pollution and mortality: a cohort study", *Canadian Medical Association Journal*, vol. 169, no. 5, pp. 397-402.

Flemming, J., Stern, R., & Yamartino, R. J. 2005, "A new air quality regime classification scheme for O₃, NO₂, SO₂ and PM₁₀ observations sites", *Atmospheric Environment*, vol. 39, no. 33, pp. 6121-6129.

Gair, A. J. & Penkett, S. A. 1995, "The effects of wind speed and turbulence on the performance of diffusion tube samplers", *Atmospheric Environment*, vol. 29, no. 18, pp. 2529-2533.

- Gallego, J. & Peedell, S. 2001, "Using CORINE land cover to map population density," in *Towards agri-environmental indicators. Integrated statistical and administrative data with land cover information*, European Environment Agency, Copenhagen, pp. 94-105.
- Garber, W., Colosio, J., Grittner, S., Larssen, S., Rasse, D., Schneider, J., & Houssiau, M. 2002, *Guidance on the Annexes to Decision 97/101/EC on Exchange of Information as revised by Decision 2001/752/EC*, for the European Commission, DG Environment.
- Gilbert, N. L., Woodhouse, S., Stieb, D. M., & Brook, J. R. 2003, "Ambient nitrogen dioxide and distance from a major highway", *The Science of the Total Environment*, vol. 312, no. 1-3, pp. 43-46.
- GLA. 2005, *London atmospheric emissions inventory (LAEI 2002) User Guide*, The Environment Group, Greater London Authority, London.
- Goldberg, M. S., Burnett, R. T., Bailer, I. I. I., Brook, J., Bonvalot, Y., Tamblyn, R., Singh, R., Valois, M. F., & Vincent, R. 2001, "The association between daily mortality and ambient air particle pollution in Montreal, Quebec: 2. Cause-specific mortality", *Environmental Research*, vol. 86, no. 1, pp. 26-36.
- Gulliver, J. & Briggs, D. J. 2005, "Time-space modeling of journey-time exposure to traffic-related air pollution using GIS", *Environmental Research*, vol. 97, no. 1, pp. 10-25.
- Haines-Young, R. H., Barr, C. J., Black, H. I. J., Briggs, D. J., Bunce, R. G. H., Clarke, R. T., Cooper, A., Dawson, F. H., Firbank, L. G., Fuller, R. M., Furse, M. T., Gillespie, M. K., Hill, R., Hornung, M., Howard, D. C., McCann, T., Morecroft, M. D., Petit, S., Sier, A. R. J., Smart, S. M., Smith, G. M., Stott, A. P., Stuart, R. C., & Watkins, J. W. 2000, *Accounting for nature: assessing habitats in the UK countryside*, Department of the Environment, Transport and the Regions, London.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. 1998, *Multivariate Data Analysis*, 5th edn, Prentice-Hall International, Inc.
- Hanninen, O., Kruize, H., Lebet, E., & Jantunen, M. 2003, "EXPOLIS simulation model: PM_{2.5} application and comparison with measurements in Helsinki", *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 13, pp. 74-85.
- Harding, J. S. & Winterbourn, M. J. 1997, "An ecoregion classification of the South Island, New Zealand", *Journal of Environmental Management*, vol. 51, no. 3, pp. 275-287.
- Harris, R. J. & Longley, P. A. 2000, "New data and approaches for urban analysis: modelling residential densities", *Transactions in GIS*, vol. 4, no. 3, pp. 217-234.
- Heal, M. R., Donoghue, M. A., & Cape, J. N. 1999, "Overestimation of urban nitrogen dioxide by passive diffusion tubes: a comparative exposure and model study", *Atmospheric Environment*, vol. 33, no. 4, pp. 513-524.
- Health Effects Institute. 2000, *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality: A Special Report of the Institute's Particle Epidemiology Reanalysis Project*, Health Effects Institute, Cambridge MA.

- Ho, K. F., Lee, S. C., & Chiu, G. M. Y. 2002, "Characterization of selected volatile organic compounds, polycyclic aromatic hydrocarbons and carbonyl compounds at a roadside monitoring station", *Atmospheric Environment*, vol. 36, no. 1, pp. 57-65.
- Hochadel, M., Heinrich, J., Gehring, U., Morgenstern, V., Kuhlbusch, T., Link, E., Wichmann, H. E., & Kramer, U. 2006, "Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information", *Atmospheric Environment*, vol. 40, no. 3, pp. 542-553.
- Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., & van den Brandt, P. A. 2002a, "Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study", *The Lancet*, vol. 360, no. 9341, pp. 1203-1209.
- Hoek, G., Dockery, D. W., Pope, A., Neas, L., Roemer, W., & Brunekreef, B. 1998, "Association between PM₁₀ and decrements in peak expiratory flow rates in children: reanalysis of data from five panel studies", *European Respiratory Journal*, vol. 11, no. 6, pp. 1307-1311.
- Hoek, G., Meliefste, K., Cyrus, J., Lewne, M., Bellander, T., Brauer, M., Fischer, P., Gehring, U., Heinrich, J., van Vliet, P., & Brunekreef, B. 2002b, "Spatial variability of fine particle concentrations in three European areas", *Atmospheric Environment*, vol. 36, no. 25, pp. 4077-4088.
- Houthuijs, D., Breugelmans, O., Hoek, G., Vaskovi, E., Mihalikova, E., Pastuszka, J. S., Jirik, V., Sachelarescu, S., Lolova, D., Meliefste, K., Uzunova, E., Marinescu, C., de Leeuw, F., van de Wiel, H., Fletcher, T., Lebret, E., & Brunekreef, B. 2001, "PM₁₀ and PM_{2.5} concentrations in Central and Eastern Europe: results from the Cesar study", *Atmospheric Environment*, vol. 35, no. 15, pp. 2757-2771.
- Ionescu, A., Candau, Y., Mayer, E., & Colda, I. 2000, "Analytical determination and classification of pollutant concentration fields using air pollution monitoring network data: Methodology and application in the Paris area, during episodes with peak nitrogen dioxide levels", *Environmental Modelling and Software*, vol. 15, no. 6-7, pp. 565-573.
- ISO. 2006. 13.040.20 Ambient atmospheres.
<http://www.iso.org/iso/en/CatalogueListPage.CatalogueList?ICS1=13&ICS2=40&ICS3=20&scopelist=ALL> . International Organization for Standardization. Access date 2-8-2006.
- Janssen, N. A. H., van Vliet, P. H. N., Aarts, F., Harssema, H., & Brunekreef, B. 2001, "Assessment of exposure to traffic related air pollution of children attending schools near motorways", *Atmospheric Environment*, vol. 35, no. 22, pp. 3875-3884.
- Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., & Giovis, C. 2004, "A review and evaluation of intraurban air pollution exposure models", *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 15, no. 2, pp. 185-204.
- Jerrett, M., Buzzelli, M., Burnett, R. T., & DeLuca, P. F. 2005, "Particulate air pollution, social confounders, and mortality in small areas of an industrial city", *Social Science and Medicine*, vol. 60, no. 12, pp. 2845-2863.

JRC. 2006a. Download meteo data derived from global atmospherical model. <http://imageportal.jrc.it/foodmeteo.shtml> . Joint Research Centre. Access date 22-5-2006.

JRC. 2006b. Global Land Cover 2000. <http://ies.jrc.cec.eu.int/31.html> . Joint Research Centre. Access date 17-5-2006.

Kanaroglou, P. S., Jerrett, M., Morrison, J., Beckerman, B., Arain, M. A., Gilbert, N. L., & Brook, J. R. 2005, "Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach", *Atmospheric Environment*, vol. 39, no. 13, pp. 2399-2409.

Katsouyanni, K. 2003, "Ambient air pollution and health", *British Medical Bulletin*, vol. 68, no. 1, pp. 143-156.

Katsouyanni, K., Touloumi, G., Samoli, E., Gryparis, A., Le Tertre, A., Monopolis, Y., Rossi, G., Zmirou, D., Ballester, F., Boumghar, A., Anderson, H. R., Wojtyniak, B., Paldy, A., Braunstein, R., Pekkanen, J., Schindler, C., & Schwartz, J. 2001, "Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 European cities within the APHEA2 Project", *Epidemiology*, vol. 12, no. 5, pp. 521-531.

Katsouyanni, K., Touloumi, G., Spix, C., Schwartz, J., Balducci, F., Medina, S., Rossi, G., Wojtyniak, B., Sunyer, J., Bacharova, L., Schouten, J. P., Ponka, A., & Anderson, H. R. 1997, "Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project", *British Medical Journal*, vol. 314, no. 7095, p. 1658.

King, K., Sturman, J., & Passant, N. 2006, *NAEI UK Emission Mapping Methodology 2003: A report of the National Atmospheric Emissions Inventory*. AEAT/ENV/R/2259.

Koop, G. & Tole, L. 2004, "Measuring the health effects of air pollution: To what extent can we really say that people are dying from bad air?", *Journal of Environmental Economics and Management*, vol. 47, no. 1, pp. 30-54.

Kousa, A., Kukkonen, J., Karppinen, A., Aarnio, P., & Koskentalo, T. 2002, "A model for evaluating the population exposure to ambient air pollution in an urban area", *Atmospheric Environment*, vol. 36, no. 13, pp. 2109-2119.

Koutrakis, P. & Sioutas, C. 1996, "Physio-chemical properties and measurement of ambient particles," in *Particles in Our Air*, R. Wilson & J. Spengler, eds., Harvard University Press, USA, pp. 15-39.

Laden, F., Neas, L. M., Dockery, D. W., & Schwartz, J. 2000, "Association of fine particulate matter from different sources with daily mortality in six U.S. cities", *Environmental Health Perspectives*, vol. 108, no. 10, pp. 941-947.

Langford, M. 2006, "Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method", *Computers, Environment and Urban Systems*, vol. 30, no. 2, pp. 161-180.

Langford, M. & Harvey, J. T. 2001. "The use of remotely sensed data for spatial disaggregation of published census population counts", *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, pp. 260-264.

- Larssen, S. & Lazaridis, M. 1998, *EUROAIRNET site selection 1998*, European Environment Agency, Technical Report No. 16.
- Larssen, S., Sluyter, R., & Helmis, C. 1999, *Criteria for EUROAIRNET. The EEA Air Quality Monitoring and Information Network*, European Environment Agency, Technical Report No. 12.
- Lavecchia, C., Angelino, E., Bedogni, M., Bravetti, E., Gualdi, R., Lanzani, G., Musitelli, A., & Valentini, M. 1996, "The ozone patterns in the aerological basin of Milan (Italy)", *Environmental Software*, vol. 11, no. 1-3, pp. 73-80.
- Lie, M. K., Avrin, R. I., Pollack, J. V., Behar, J. V., & McElroy, J. L. 1986, "Methodology for designing air quality monitoring networks: I. Theoretical aspects", *Environmental Monitoring and Assessment*, vol. 6, no. 1, pp. 1-11.
- Liu, L.J. S. & Rossini, A. J. 1996, "Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto - A pilot study", *Environment International*, vol. 22, no. 6, pp. 677-692.
- Livingstone, A. E., Shaddick, G., Grundy, C., & Elliott, P. 1996, "Do people living near inner city main roads have more asthma needing treatment? Case-control study", *British Medical Journal*, vol. 312, no. 7032, pp. 676-677.
- Loibl, W. & Orthofer, R. 2001, "From national emission totals to regional ambient air quality information for Austria", *Advances in Environmental Research*, vol. 5, no. 4, pp. 395-404.
- Maffeis, G. 1999, "Prediction of carbon monoxide acute air pollution episodes. Model formulation and first application in Lombardy", *Atmospheric Environment*, vol. 33, no. 23, pp. 3859-3872.
- Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., & Gwynne, M. 1996, "Urban air pollution in megacities of the world", *Atmospheric Environment*, vol. 30, no. 5, pp. 681-686.
- Maheswaran, R. & Elliott, P. 2003, "Stroke mortality associated with living near main roads in England and Wales, A geographical study", *Stroke*, vol. 34, no. 12, pp. 2776-2780.
- Maheswaran, R., Haining, R. P., Brindley, P., Law, J., Pearson, T., Fryers, P. R., Wise, S., & Campbell, M. J. 2005, "Outdoor air pollution and stroke in Sheffield, United Kingdom: A small-area level geographical study", *Stroke*, vol. 36, no. 2, pp. 239-243.
- McGregor, G. R. 1996, "Identification of air quality affinity areas in Birmingham, UK", *Applied Geography*, vol. 16, no. 2, pp. 109-122.
- McHugh, C. A., Carruthers, D. J., Higson, H., & Dyster, S. J. 1999, *Comparison of model evaluation methodologies with application to ADMS 3 and US models*, Cambridge Environmental Research Consultants Ltd., 6 HARMO 99.
- Mennis, J. 2002, "Using geographic information systems to create and analyse statistical surfaces of population and risk for environmental justice analysis", *Social Science Quarterly*, vol. 83, no. 1, pp. 281-297.

Mennis, J. 2003, "Generating surface models of population using dasymetric mapping", *The Professional Geographer*, vol. 55, no. 1, pp. 31-42.

MET Office. 2006. Beaufort Scale.

<http://www.metoffice.com/education/secondary/students/beaufort.html>. Access date 2-8-2006.

Modak, P. M. & Lohani, B. N. 1985, "Optimization of ambient air quality monitoring networks (Part I)", *Environmental Monitoring and Assessment*, vol. 5, no. 1, pp. 1-19.

Monn, C. 2001, "Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone", *Atmospheric Environment*, vol. 35, no. 1, pp. 1-32.

Mortimer, K. M., Neas, L. M., Dockery, D. W., Redline, S., & Tager, I. B. 2002, "The effect of air pollution on inner-city children with asthma", *European Respiratory Journal*, vol. 19, no. 4, pp. 699-705.

Nafstad, P., Håheim, L. L., Oftedal, B., Gram, F., Holme, I., Hjermann, I., & Leren, P. 2003, "Lung cancer and air pollution: a 27 year follow up of 16 209 Norwegian men", *Thorax*, vol. 58, no. 12, pp. 1071-1076.

Nafstad, P., Håheim, L., Wisloff, T., Gram, F., Oftedal, B., Holme, I., Hjermann, I., & Leren, P. 2004, "Urban air pollution and mortality in a cohort of Norwegian men", *Environmental Health Perspectives*, vol. 112, no. 5, pp. 610-615.

Netcen. 2006. Non Automatic Networks - Air Quality Archive.

<http://www.airquality.co.uk/archive/nonauto.php> . UK Department for Environment, Food and Rural Affairs and the Devolved Administrations. Access date 2-8-2006.

Neuberger, M., Schimek, M. G., Horak, J., Moshhammer, H., Kundi, M., Frischer, T., Gomiscek, B., Puxbaum, H., & Hauck, H. 2004, "Acute effects of particulate matter on respiratory diseases, symptoms and functions: epidemiological results of the Austrian Project on Health Effects of Particulate Matter (AUPHEP)", *Atmospheric Environment*, vol. 38, no. 24, pp. 3971-3981.

Nieuwenhuijsen, M. 2003, "Introduction to exposure assessment," in *Exposure Assessment in Occupational and Environmental Epidemiology*, M. Nieuwenhuijsen, ed., Oxford University Press, Oxford, pp. 3-19.

Norusis, M. J. 1988, *SPSS/PC+ Advanced Statistics V2.0* SPSS Inc., Chicago.

Nyberg, F., Gustavsson, P., Jarup, L., Bellander, T., Berglind, N., Jakobsson, R., & Pershagen, G. 2000, "Urban air pollution and lung cancer in Stockholm", *Epidemiology*, vol. 11, no. 5, pp. 487-495.

Official Journal of the European Communities. 1996. Council Directive 96/62/EC of 27 September 1996 on ambient air quality assessment and management. No. L 296/55.

Official Journal of the European Communities. 1999. Council Directive 1999/30/EC of 22 April 1999 relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air. No. L 163/41.

Official Journal of the European Communities. 2003. Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS). L 154, P. 0001-0041.

ONS. 2003. Beginners' Guide to UK Geography.

http://www.statistics.gov.uk/geography/beginners_guide.asp . 2003. Access date 5-6-2006.

ONS. 2005, *All Fields Postcode Directory: 2005 User Guide*.

Penard-Morand, C., Schillinger, C., Armengaud, A., Debotte, G., Chretien, E., Pellier, S., & Annesi-Maesano, I. 2006, "Assessment of schoolchildren's exposure to traffic-related air pollution in the French Six Cities Study using a dispersion model", *Atmospheric Environment*, vol. 40, no. 13, pp. 2274-2287.

Peters, A., Dockery, D. W., Muller, J. E., & Mittleman, M. A. 2001, "Increased particulate air pollution and the triggering of myocardial infarction", *Circulation*, vol. 103, no. 23, pp. 2810-2815.

Peters, A., Liu, E., Verrier, R. L., Schwartz, J., Gold, D. R., Mittleman, M., Baliff, J., Oh, J. A., Allen, G., Monahan, K., & Dockery, D. W. 2000, "Air pollution and incidence of cardiac arrhythmia", *Epidemiology*, vol. 11, no. 1, pp. 11-17.

Pikhart, H., Bobak, M., Gorynski, P., Wojtyniak, B., Danova, J., Celko, M. A., Kriz, B., Briggs, D.J., & Elliott, P. 2001, "Outdoor sulphur dioxide and respiratory symptoms in Czech and Polish school children: a small-area study (SAVIAH). Small-Area Variation in Air Pollution and Health", *International Archives of Occupational and Environmental Health*, vol. 74, no. 8, pp. 574-578.

Ping Shi, J. & Harrison, R. M. 1997, "Regression modelling of hourly NO_x and NO₂ concentrations in urban air in London", *Atmospheric Environment*, vol. 31, no. 24, pp. 4081-4094.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. 2002, "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution", *The Journal of the American Medical Association*, vol. 287, no. 9, pp. 1132-1141.

Pope III, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., & Heath, C. W., Jr. 1995, "Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults", *American Journal of Respiratory and Critical Care Medicine*, vol. 151, no. 3 Pt 1, pp. 669-674.

PRB. 2005. 2005 World Population Data Sheet of the Population Reference Bureau. http://www.prb.org/pdf05/05WorldDataSheet_Eng.pdf . Access date 4-3-2006

Roorda-Knape, M. C., Janssen, N. A. H., de Hartog, J. J., van Vliet, P. H. N., Harssema, H., & Brunekreef, B. 1998, "Air pollution from traffic in city districts near major motorways", *Atmospheric Environment*, vol. 32, no. 11, pp. 1921-1930.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., Schwartz, J., & Zanobetti, A. 2000, "The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and mortality from air pollution in the United States", *Research Report/ Health Effects Institute*, vol. 94, no. Pt 2, pp. 5-70.

- Schwartz, J. 2000, "Fine particles are more strongly associated than coarse particles with acute respiratory health effects in schoolchildren", *Epidemiology*, vol. 11, no. 1, pp. 6-10.
- Scoggins, A., Kjellstrom, T., Fisher, G., Connor, J., & Gimson, N. 2004, "Spatial analysis of annual air pollution exposure and mortality", *Science of the Total Environment*, vol. 321, no. 1-3, pp. 71-85.
- Shaddick, G. & Wakefield, J. 2002, "Modelling daily multivariate pollutant data at multiple sites", *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 51, no. 3, pp. 351-372.
- Sluyter, R., Potma, C., Kroghes, T., Petrakis, M., & van Hooydonk, P. 1997, *AIRBASE: 1997 Development Status & Extensions Foreseen. Second European Workshop on Air Quality Monitoring and Assessment Brussels, 22-23 September 1997*.
- Spengler, J. & Wilson, R. 1996, "Emissions, dispersion, and concentration of particles," in *Particles in Our Air*, R. Wilson & J. Spengler, eds., Harvard University Press, USA, pp. 41-62.
- SPSS. 2003. SPSS 12.0 for Windows. [Release 12.0.0] Help file.
- Stedman, J. R. 1997, "A UK-wide episode of elevated particle (PM₁₀) concentration in March 1996", *Atmospheric Environment*, vol. 31, no. 15, pp. 2381-2383.
- Stedman, J. R., Vincent, K. J., Campbell, G. W., Goodwin, J. W. L., & Downing, C. E. H. 1997, "New high resolution maps of estimated background ambient NO_x and NO₂ concentrations in the U.K", *Atmospheric Environment*, vol. 31, no. 21, pp. 3591-3602.
- Sunyer, J., Ballester, F., Tertre, A. L., Atkinson, R., Ayres, J. G., Forastiere, F., Forsberg, B., Vonk, J. M., Bisanti, L., & Tenias, J. M. 2003, "The association of daily sulfur dioxide air pollution levels with hospital admissions for cardiovascular diseases in Europe (The APHEA-II study)", *European Heart Journal*, vol. 24, no. 8, pp. 752-760.
- Tayanc, M. 2000, "An assessment of spatial and temporal variation of sulfur dioxide levels over Istanbul, Turkey", *Environmental Pollution*, vol. 107, no. 1, pp. 61-69.
- Technical Working Group on Particles. 1997. *Position Paper on Ambient Air Pollution by Particulate Matter*. European Commission.
- Tobler, R. 1970, "A computer movie simulating urban growth in the Detroit region", *Economic Geography*, vol. 46, pp. 234-240.
- Tobler, R. 1979, "Smooth pycnophylactic interpolation for geographical regions", *Journal of the American Statistical Association*, vol. 74, no. 367, p. 519.
- Touloumi, G., Katsouyanni, K., Zmirou, D., Schwartz, J., Spix, C., de Leon, A. P., Tobias, A., Quenel, P., Rabaczko, D., Bacharova, L., Bisanti, L., Vonk, J. M., & Ponka, A. 1997, "Short-term effects of ambient oxidant exposure on mortality: a combined analysis within the APHEA project. Air Pollution and Health: a European Approach", *American Journal of Epidemiology*, vol. 146, no. 2, pp. 177-185.

Valkonen, E., Harkonen, J., Kukkonen, J., Rantakrans, E., Jalkanen, L., & Haarala, S. 1996, "Modelling urban air pollution in Espoo, Finland", *The Science of the Total Environment*, vol. 189-190, pp. 205-211.

Webster, R. & Burrough, P. A. 1974, "Multiple discriminant analysis in soil survey", *Journal of Soil Science*, vol. 25, no. 1, pp. 120-134.

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

WHO. 2000, *Air Quality Guidelines for Europe*, World Health Organization, Copenhagen, No. 91.

WHO. 2002, *The World Health Report 2002: Reducing Risks, Promoting Healthy Life*, World Health Organization, Geneva, Switzerland.

Wilkinson, P., Elliott, P., Grundy, C., Shaddick, G., Thakrar, B., Walls, P., & Falconer, S. 1999, "Case-control study of hospital admission with asthma in children aged 5-14 years: relation with road traffic in north west London", *Thorax*, vol. 54, no. 12, pp. 1070-1074.

Wolf, C. 2002, "Urban air pollution and health: an ecological study of chronic rhinosinusitis in Cologne, Germany", *Health and Place*, vol. 8, no. 2, pp. 129-139.

Working Group on Nitrogen Dioxide. 1997, *Position Paper on Air Quality: Nitrogen Dioxide*. European Commission.

Yu, T. Y. & Chang, L. F. 2001, "Delineation of air-quality basins utilizing multivariate statistical methods in Taiwan", *Atmospheric Environment*, vol. 35, no. 18, pp. 3155-3166.

Yuan, Y., Smith, R. M., & Limp, W. F. 1997, "Remodeling census population with spatial information from Landsat TM imagery", *Computers, Environment and Urban Systems*, vol. 21, no. 3-4, pp. 245-258.

Zelenka, M. P. 1997, "An analysis of the meteorological parameters affecting ambient concentrations of acid aerosols in Uniontown, Pennsylvania", *Atmospheric Environment*, vol. 31, no. 6, pp. 869-878.

A1. Base coverage

Parameters of Lambert Azimuthal Equal Area Projection

Coordinate System:		Extent:	
Lambert_Azimuthal_Equal_Area		EU 15 + Norway	
False_Easting:	0.000000	Left	-2738313.193264
False_Northing:	0.000000	Right	1835186.806736
Central_Meridian:	9.000000	Top	2698881.935179
Latitude_Of_Origin:	48.000000	Bottom	-1929518.064821
GSC_User_Defined		(in meters)	
Datum:	D_User_Defined		
Prime Meridian:	0		

A2. ECMWF meteorological data

Wind Direction - manual steps (part 1)

The first step in calculating wind direction is to average the U and V wind components separately for the three averaging periods (annual, summer and winter). Next the calcwind.aml is run to calculate the annual, summer and winter mean wind direction and wind speed.

```
Arc: |> additem wind_uv.pat wind_uv.pat amean_u 8 19 F 3 <|
Arc: |> additem wind_uv.pat wind_uv.pat amean_v 8 19 F 3 <|
Arc: |> additem wind_uv.pat wind_uv.pat smean_u 8 19 F 3 <|
Arc: |> additem wind_uv.pat wind_uv.pat smean_v 8 19 F 3 <|
Arc: |> additem wind_uv.pat wind_uv.pat wmean_u 8 19 F 3 <|
Arc: |> additem wind_uv.pat wind_uv.pat wmean_v 8 19 F 3 <|

Arc: |> tables <|
Tables: |> select wind_uv.pat <|
7368 Records Selected.
Tables: |> calc amean_u = ( u10_1 + u10_2 + u10_3 + u10_4 + u10_5 + u10_6 + ~
u10_7 + u10_8 + u10_9 + u10_10 + u10_11 + u10_12 ) / 12 <|
Tables: |> calc amean_v = ( v10_1 + v10_2 + v10_3 + v10_4 + v10_5 + v10_6 + ~
v10_7 + v10_8 + v10_9 + v10_10 + v10_11 + v10_12 ) / 12 <|
Tables: |> calc smean_u = ( u10_6 + u10_7 + u10_8 + u10_9 ) / 4 <|
Tables: |> calc smean_v = ( v10_6 + v10_7 + v10_8 + v10_9 ) / 4 <|
Tables: |> calc wmean_u = ( u10_1 + u10_2 + u10_3 + u10_12 ) / 4 <|
Tables: |> calc wmean_v = ( v10_1 + v10_2 + v10_3 + v10_12 ) / 4 <|
Tables: |> q <|
```

Calcwind.aml

```
/* *****
/* Calculate wind direction and wind speed from the U and V components
/* &r calcwind
/*
/* Danielle Vienneau
/* March 05
/* *****

&call calc_wind
&call join_item
&return

/*-----
&routine calc_wind
/*-----

&sv windcov = wind_uv

/* Get the size of the point coverage
ARCPLLOT
clearselect
reselect %windcov% point $recno ge 0
&sv totpts = [extract 1 [show select %windcov% point] ]
QUIT

/* Loop through seasons (annual, summer and winter), then through the points
&do month &list a s w

    /* Open a text file to hold the output for this month
    &sv fileunit = [open winddir_%month%.txt openstat -write]

    ARCPLLOT
    &sv count = 0
    &do &until %count% = %totpts%
```



```

&sv count = %count% + 1

&sv Ucomp = [show select %windcov% point %count% ITEM %month%mean_u]
&sv Vcomp = [show select %windcov% point %count% ITEM %month%mean_v]

&if [calc %Vcomp% = 0] &then
    &sv windrad = [calc 0.5 * 3.14159265]
&else
    &do
        &sv tanval = [abs [calc %Ucomp% / %Vcomp%] ]
        &sv windrad = [ATAN %tanval% ]
    &end

    &sv uv = [calc [calc %Ucomp% * %Ucomp%] + [calc %Vcomp% * ~
    %Vcomp%] ]
    &sv windspeed = [sqrt %uv%]

/* NE quadrant
&if [calc %Ucomp% >= 0] and [calc %Vcomp% >= 0] &then
    &sv winddir = [calc 57.2957795 * %windrad%]

/* SE quadrant
&if [calc %Ucomp% >= 0] and [calc %Vcomp% < 0] &then
    &sv winddir = [calc 180 - [calc 57.2957795 * %windrad%] ]

/* SW quadrant
&if [calc %Ucomp% < 0] and [calc %Vcomp% < 0] &then
    &sv winddir = [calc 180 + [calc 57.2957795 * %windrad%] ]

/* NW quadrant
&if [calc %Ucomp% < 0] and [calc %Vcomp% >= 0] &then
    &sv winddir = [calc 360 - [calc 57.2957795 * %windrad%] ]

&sv pointid = [show select %windcov% point %count% ITEM pointid]

/* Ensure variable exists, and write it to the text file
&if [variable pointid] = .FALSE. &then
    &sv windx = 0
&if [variable winddir] = .FALSE. &then
    &sv winddir = 0
&if [variable windspeed] = .FALSE. &then
    &sv windspeed = 0

&sv writestat = [WRITE %fileunit% %pointid%,%winddir%,%windspeed%]

&end

&sv closestat = [close -all]
QUIT

&sv thePath = [pathname winddir_%month%.txt -FILE]
TABLES
define winddir_%month%.lut
POINTID 8 8 F 0
WDIR_%month% 8 9 F 3
WSPEED_%month% 8 9 F 3
~
~

select winddir_%month%.lut
add from %thePath%
QUIT

&end
&return

```



```

/*-----
&routine join_item
/*-----

```

```

&do month &list a s w

```

```

    joinitem wind_ds.pat winddir_%month%.lut wind_ds.pat pointid

```

```

&end

```

```

&return

```

```

/*****

```

Wind Direction - manual steps (part 2)

The 0.5° points with wind data attached is converted to a grid then polygon coverage. It is re-projected and intersected with the base coverage.

```

Arc: |> copy wind_ds wind_dir <|

```

```

Arc: |> pointgrid wind_dir wind_dir_g pointid <|

```

```

Cell Size (square cell): |> 0.5 <|

```

```

Convert the Entire Coverage(Y/N)? : |> y <|

```

```

Enter background value (NODATA | ZERO): |><|

```

```

Arc: |> grid <|

```

```

Grid: |> wind_dir_i = int(wind_dir_g) <|

```

```

Grid: |> buildvat wind_dir_i <|

```

```

Grid: |> q <|

```

```

Arc: |> gridpoly wind_dir_i wind_dir_poly <|

```

```

Arc: |> tables <|

```

```

Tables: |> select wind_dir_poly.pat <|

```

```

Tables: |> alter <|

```

```

Enter Item Name: |> grid-code <|

```

```

Item Name: |> pointid <|

```

```

Item Output Width: |><|

```

```

Item Type: |><|

```

```

Alternate Item Name: |><|

```

```

Enter Item Name: |><|

```

```

Tables: |> q <|

```

```

Arc: |> project cover wind_dir_poly wind_dirp

```

```

o:\projects\apmosphere\projection\gcecgla.prj <|

```

```

Arc: |> build wind_dirp poly <|

```

```

Arc: |> copy wind_dirp o:\projects\apmosphere\eudata\wind_dirp <|

```

```

Arc: |> intersect eulkm_point wind_dirp eulkm_wdp point <|

```

```

Arc: |> copy eulkm eulkm_wd01 <|

```

```

Arc: |> joinitem eulkm_wd01.pat eulkm_wdp.pat eulkm_wd01.pat grid_code <|

```

```

Arc: |> dropitem eulkm_wd01.pat eulkm_wd01.pat wind_dirp-id wind_dirp#
eulkm_point# eulkm_point-id eulkm_wdp# eulkm_wdp-id <|

```

```

Arc: |> tables <|

```

```

Tables: |> select eulkm_wd01.pat <|

```

```

Tables: |> reselect grid_code = -9999 <|

```

```

Tables: |> calc am_wdir = -9999 <|

```

```

Tables: |> calc sm_wdir = -9999 <|

```

```

Tables: |> calc wm_wdir = -9999 <|

```

```

Tables: |> q <|

```


A3. Sea distance

This AML to calculate distance to sea as a measure of continentality, was written by Dr. Linda Beale from the GIS Group, Imperial College, Department of Epidemiology and Public Health.

Inside_land.aml

```

/*****
/*                               inside_landv1                               */
/*                               */
/*      Calculates the continentality of land points                       */
/*      requires: land point cover                                         */
/*                               sea point cover                           */
/*      creates: seadist.txt                                               */
/*                               seadist.dat (with x,y coords)             */
/*                               */
/*                               Linda Beale Dec 2003                      */
/*                               */
*****/

/* where 1 = NE, 2 = SE, 3 = SW, 4 = NW

&args landcover seacover

/* copy point files to new coveragrs & add x, y coordinates

copy %landcover% landpnts
  addxy landpnts point
copy %seacover% seapnts
  addxy seapnts point

/* Get the size of the two point coverages

ARCPLLOT

  clearselect
  reselect landpnts point $recno ge 0

  &sv tot_landpnts = [extract 1 [show select landpnts point] ]

  clearselect
  reselect seapnts point $recno ge 0

  &sv tot_seapnts = [extract 1 [show select seapnts point] ]

/* open text file <seadist.txt> to hold final closest distances

  &sv fileunit = [open seadist.txt openstat -write]

/* loop through the land cover point file - calc distance for each point

  &sv count = 0

  &do &until %count% eq %tot_landpnts%
    &sv count = %count% + 1
    &type calculating continentality for land cover point: %count%

    &sv landy = [show select landpnts point %count% ITEM y-coord]
    &sv landx = [show select landpnts point %count% ITEM x-coord]
    &sv num = 0

    &delvar dist*

/* loop through the sea points to find the 'closest point
```



```

&do &until %num% eq %tot_seapnts%
  &sv num = %num% + 1

  &sv seay = [show select seapnts point %num% ITEM y-coord]
  &sv seax = [show select seapnts point %num% ITEM x-coord]

  &if [calc %seay% > %landy%] and [calc %seax% >= %landx%] &then
    &sv seapnt = 1

  &if [calc %seay% <= %landy%] and [calc %seax% > %landx%] &then
    &sv seapnt = 2

  &if [calc %seay% < %landy%] and [calc %seax% <= %landx%] &then
    &sv seapnt = 3

  &if [calc %seay% >= %landy%] and [calc %seax% < %landx%] &then
    &sv seapnt = 4

/* check each new distance to see if it is closer than the previous

  &sv diffx = [abs [calc %landx% - %seax%] ]
  &sv diffy = [abs [calc %landy% - %seay%] ]
  &sv newdist = [sqrt [calc [calc %diffx% * %diffx%] + [calc %diffy% * ~
%diffy%]]]

  &if [calc %seapnt% = 1] &then
    &do
      &if [variable dist1] = .FALSE. &then
        &sv dist1 = %newdist%

      &if [calc %newdist% < %dist1%] &then
        &sv dist1 = %newdist%
    &end

  &if [calc %seapnt% = 2] &then
    &do
      &if [variable dist2] = .FALSE. &then
        &sv dist2 = %newdist%

      &if [calc %newdist% < %dist2%] &then
        &sv dist2 = %newdist%
    &end

  &if [calc %seapnt% = 3] &then
    &do
      &if [variable dist3] = .FALSE. &then
        &sv dist3 = %newdist%

      &if [calc %newdist% < %dist3%] &then
        &sv dist3 = %newdist%
    &end

  &if [calc %seapnt% = 4] &then
    &do
      &if [variable dist4] = .FALSE. &then
        &sv dist4 = %newdist%

      &if [calc %newdist% < %dist4%] &then
        &sv dist4 = %newdist%
    &end

  &end

/* ensure variable exists, and write to the txt file <seadist.txt>

  &if [variable dist1] = .FALSE. &then
    &sv dist1 = 0
  &if [variable dist2] = .FALSE. &then

```



```

        &sv dist2 = 0
    &if [variable dist3] = .FALSE. &then
        &sv dist3 = 0
    &if [variable dist4] = .FALSE. &then
        &sv dist4 = 0

    &sv writestat = [WRITE %fileunit% ~
    %landx%,%landy%,%dist1%,%dist2%,%dist3%,%dist4%]

&end

&sv closestat = [close -all]

/* build an info file that holds the x, y coords of the land cover and the
/* distance attributes

&sv thePath = [pathname seadist.txt -FILE]

QUIT

TABLES
define seadist.dat
LANDX 8 9 F 2
LANDY 8 9 F 2
DIST_NE 8 9 F 4
DIST_SE 8 9 F 4
DIST_SW 8 9 F 4
DIST_NW 8 9 F 4
~
~

select seadist.dat
add from %thePath%

QUIT

kill landpnts all
kill seapnts all

&return
/*****/

```


A4. Summary statistics for derived data

Summary Statistics for derived data

Variable	Description	N	Minimum	Maximum	Mean	Std. Deviation
RAIL	Rail length (1x1)	3314223	0.00	6.30	0.05	0.23
RAIL3	Rail length (3x3)	3314223	0.00	23.38	0.36	0.97
RAIL5	Rail length (5x5)	3314223	0.00	36.17	0.99	2.16
RAIL11	Rail length (11x11)	3314223	0.00	117.32	4.76	7.38
RAIL21	Rail length (21x21)	3314223	0.00	283.69	17.20	20.56
MOTOR	Motorway length (1x1)	3666274	0.00	12.45	0.02	0.18
MOTOR3	Motorway length (3x3)	3666274	0.00	29.16	0.13	0.72
MOTOR5	Motorway length (5x5)	3666274	0.00	42.66	0.35	1.58
MOTOR11	Motorway length (11x11)	3666274	0.00	101.03	1.68	5.36
MOTOR21	Motorway length (21x21)	3666274	0.00	235.18	6.05	14.90
AROAD	A-road length (1x1)	3666274	0.00	9.86	0.06	0.26
AROAD3	A-road length (3x3)	3666274	0.00	24.37	0.40	1.10
AROAD5	A-road length (5x5)	3666274	0.00	49.47	1.11	2.46
AROAD11	A-road length (11x11)	3666274	0.00	145.91	5.30	8.43
AROAD21	A-road length (21x21)	3666274	0.00	386.55	19.13	23.25
BROAD	B-road length (1x1)	3666274	0.00	8.01	0.24	0.49
BROAD3	B-road length (3x3)	3666274	0.00	28.01	1.66	2.18
BROAD5	B-road length (5x5)	3666274	0.00	61.76	4.61	5.09
BROAD11	B-road length (11x11)	3666274	0.00	221.18	22.11	20.00
BROAD21	B-road length (21x21)	3666274	0.00	599.50	79.65	65.67
MINOR	Minor road length (1x1)	3666274	0.00	11.44	0.47	0.70
MINOR3	Minor road length (3x3)	3666274	0.00	64.29	3.33	3.27
MINOR5	Minor road length (5x5)	3666274	0.00	150.06	9.23	7.97
MINOR11	Minor road length (11x11)	3666274	0.00	559.35	44.22	33.21
MINOR21	Minor road length (21x21)	3666274	0.00	1284.01	159.23	111.12
MAJRD1	Length of Motorway + A-roads (1x1)	3666274	0.00	12.45	0.07	0.32
MAJRD5	Length of Motorway + A-roads (5x5)	3666274	0.00	62.14	1.46	3.18
MINRD1	Length of B-roads + minor roads (1x1)	3666274	0.00	13.93	0.71	0.87
MINRD5	Length of B-roads + minor roads (5x5)	3666274	0.00	185.17	13.84	10.30
CORINE land cover						
HDRES	Percent High density residential (1x1)	2854116	0.00	100.00	0.34	4.00
HDRES_3	Percent High density residential (3x3)	2854116	0.00	99.00	0.30	2.77
HDRES_5	Percent High density residential (5x5)	2854116	0.00	95.00	0.28	2.30

Summary Statistics for derived data

Variable	Description	N	Minimum	Maximum	Mean	Std. Deviation
HDRES_11	Percent High density residential (11x11)	2854116	0.00	82.00	0.23	1.67
HDRES_21	Percent High density residential (21x21)	2854116	0.00	56.00	0.19	1.26
LDRES	Percent Low density residential (1x1)	2854116	0.00	100.00	2.43	10.28
LDRES_3	Percent Low density residential (3x3)	2854116	0.00	99.00	2.25	7.37
LDRES_5	Percent Low density residential (5x5)	2854116	0.00	99.00	2.20	6.42
LDRES_11	Percent Low density residential (11x11)	2854116	0.00	79.00	2.08	5.25
LDRES_21	Percent Low density residential (21x21)	2854116	0.00	64.00	2.01	4.50
IND	Percent Industrial (1x1)	2854116	0.00	100.00	0.35	3.60
IND_3	Percent Industrial (3x3)	2854116	0.00	96.00	0.31	2.15
IND_5	Percent Industrial (5x5)	2854116	0.00	80.00	0.29	1.66
IND_11	Percent Industrial (11x11)	2854116	0.00	38.00	0.24	1.09
IND_21	Percent Industrial (21x21)	2854116	0.00	16.00	0.19	0.77
TRANS	Percent transport (1x1)	2854116	0.00	98.00	0.04	0.85
TRANS_3	Percent transport (3x3)	2854116	0.00	55.00	0.03	0.41
TRANS_5	Percent transport (5x5)	2854116	0.00	34.00	0.03	0.29
TRANS_11	Percent transport (11x11)	2854116	0.00	8.00	0.01	0.16
TRANS_21	Percent transport (21x21)	2854116	0.00	4.00	0.01	0.08
PORT	Precent seaport (1x1)	2854116	0.00	100.00	0.02	0.98
PORT_3	Precent seaport (3x3)	2854116	0.00	82.00	0.02	0.61
PORT_5	Precent seaport (5x5)	2854116	0.00	58.00	0.02	0.46
PORT_11	Precent seaport (11x11)	2854116	0.00	25.00	0.01	0.26
PORT_21	Precent seaport (21x21)	2854116	0.00	9.00	0.01	0.14
AIRP	Percent airport (1x1)	2854116	0.00	100.00	0.07	1.96
AIRP_3	Percent airport (3x3)	2854116	0.00	95.00	0.07	1.21
AIRP_5	Percent airport (5x5)	2854116	0.00	64.00	0.06	0.86
AIRP_11	Percent airport (11x11)	2854116	0.00	17.00	0.05	0.41
AIRP_21	Percent airport (21x21)	2854116	0.00	7.00	0.02	0.19
WECS	Percent waste, construction & institutions (1x1)	2854116	0.00	100.00	0.18	2.64
WECS_3	Percent waste, construction & institutions (3x3)	2854116	0.00	99.00	0.15	1.70
WECS_5	Percent waste, construction & institutions (5x5)	2854116	0.00	95.00	0.14	1.35
WECS_11	Percent waste, construction & institutions (11x11)	2854116	0.00	51.00	0.10	0.86
WECS_21	Percent waste, construction & institutions (21x21)	2854116	0.00	28.00	0.07	0.58
URBGR	Percent urban green space (1x1)	2854116	0.00	100.00	0.19	2.50

Summary Statistics for derived data

Variable	Description	N	Minimum	Maximum	Mean	Std. Deviation
URBGR_3	Percent urban green space (3x3)	2854116	0.00	84.00	0.16	1.36
URBGR_5	Percent urban green space (5x5)	2854116	0.00	60.00	0.15	1.04
URBGR_11	Percent urban green space (11x11)	2854116	0.00	31.00	0.11	0.71
URBGR_21	Percent urban green space (21x21)	2854116	0.00	17.00	0.09	0.54
FOR	Percent forest (1x1)	3666274	0.00	100.00	34.47	39.18
FOR_3	Percent forest (3x3)	3666274	0.00	99.00	33.95	34.21
FOR_5	Percent forest (5x5)	3666274	0.00	99.00	33.91	32.40
FOR_11	Percent forest (11x11)	3666274	0.00	99.00	33.68	29.93
FOR_21	Percent forest (21x21)	3666274	0.00	99.00	33.46	28.28
AGR	Percent agriculture (1x1)	3666274	0.00	100.00	38.85	41.46
AGR_3	Percent agriculture (3x3)	3666274	0.00	99.00	38.37	37.49
AGR_5	Percent agriculture (5x5)	3666274	0.00	99.00	38.33	36.04
AGR_11	Percent agriculture (11x11)	3666274	0.00	99.00	37.99	33.83
AGR_21	Percent agriculture (21x21)	3666274	0.00	99.00	37.57	32.12
TBU	Percent total built-up (1x1)	3666274	0.00	100.00	2.96	12.39
TBU_3	Percent total built-up (3x3)	3666274	0.00	99.00	2.78	9.69
TBU_5	Percent total built-up (5x5)	3666274	0.00	99.00	2.73	8.66
TBU_11	Percent total built-up (11x11)	3666274	0.00	99.00	2.61	7.14
TBU_21	Percent total built-up (21x21)	3666274	0.00	97.00	2.52	6.05
Climatic variables for year 2001						
a_calm01	Annual calm days (1x1)	3666274	0.00	51.77	4.70	6.32
am_rad01	Annual mean radiation (1x1)	3666274	5800.11	19642.52	11016.60	3251.76
am_tmp01	Annual mean temperature (1x1)	3666274	-2.70	21.20	9.27	5.41
am_trg01	Annual mean temperature range (1x1)	3666274	3.03	13.97	8.12	1.87
am_vp01	Annual mean vapour pressure (1x1)	3666274	5.26	17.24	9.76	2.38
am_ws01	Annual mean wind speed (1x1)	3666274	1.14	8.68	3.20	0.87
as_r01	Annual mean rainfall (1x1)	3666274	78.00	1884.34	701.06	232.62
s_calm01	Summer calm days (1x1)	3666274	0.00	48.89	2.58	4.42
sm_tmp01	Summer mean temperature (1x1)	3666274	8.95	28.12	17.75	4.16
sm_ws01	Summer mean wind speed (1x1)	3666274	1.16	7.19	3.01	0.73
w_calm01	Winter calm days (1x1)	3666274	0.00	55.01	7.88	10.55
wm_tmp01	Winter mean temperature (1x1)	3666274	-14.60	16.02	1.50	7.55
wm_ws01	Winter mean wind speed (1x1)	3666274	1.11	10.03	3.35	1.07

B1. Population modelling AMLs

Data_prep.aml

```
/******  
/* data_prep.aml  
/*  
/* Danielle Vienneau (lc_split routine was written by John Gulliver)  
/* 13.02.04  
/******  
  
&call lc_split  
&call additem  
&call corine_int  
&call corine_stats  
&return  
  
/*-----  
&routine lc_split  
/*-----  
/* by John Gulliver  
  
/* Splits the grouped corine-light point coverages back out into the separate  
/* corine classes. Also adds a LIT item where points with a predicted light  
/* are set LIT = 1 and those with no light are LIT = 0  
  
&sv file_id [open coverlist.txt openstat -r]  
  
&if %openstat% ne 0 &then  
  &return &error Error in creating file  
&else &type File is created  
  
&sv line [read %file_id% readstat]  
  
&do &while %readstat% ne 102  
  &sv incov [extract 1 %line%]  
  &type incov = %incov%  
  /*&pause  
  &call classlit  
  &call makecov  
&end  
  
&s closestat [close -all]  
  
&return  
  
/*-----  
&routine classlit  
/*-----  
  
additem %incov%.pat %incov%.pat lit 2 2 i  
  
TABLES  
select %incov%.pat  
reselect predicted > 0  
calc lit = 1  
nselect  
calc lit = 0  
QUIT  
  
/*&sv line [read %file_id% readstat]  
  
&return
```

Coverlist.txt

```
group_a_idw  
group_b_idw  
group_h_idw  
group_i_idw  
group_e_idw  
group_f_idw  
group_g_idw  
group_j_idw  
group_k_idw  
group_d_idw  
group_c_idw
```



```

/*-----
&routine makecov
/*-----

&sv lclist = [listunique %incov% -point landclass]
&type lclist = %lclist%
&sv num = [token %lclist% -count]
&type number of landclasses is %num%

&do &while %num% > 0

/* extract the landclass number, set to %class%

&sv class = [extract %num% %lclist%]

/* use ARC RESELECT to make the new coverages

reselect %incov% lc%class%_idw point # point
reselect landclass = %class%
~
n
n

&sv num = %num% - 1

&end

&sv line [read %file_id% readstat]

&return

/*-----
&routine additem
/*-----
/* To add a LIT item to the corine point coverages, points with a predicted
/* light are set LIT = 1 and those with no light are LIT = 0

&sv grouplist = group_a_idw group_b_idw group_c_idw group_d_idw group_e_idw -
group_f_idw group_g_idw group_j_idw group_k_idw group_h_idw group_i_idw

&do group &list %grouplist%

&sv lclist = [listunique %grouplist% -point landclass]
&type lclist = %lclist%
&sv num = [token %lclist% -count]
&type number of landclasses is %num%

&do &while %num% > 0

/* extract the landclass number, set to %class%
&sv class = [extract %num% %lclist%]
additem lc%class%_idw.pat lc%class%_idw.pat lit 2 2 1

TABLES
select lc%class%_idw.pat
reselect predicted > 0
calc lit = 1
nselect
calc lit = 0
QUIT

&sv num = %num% - 1

&end
&end

&return

```



```

/*-----
&routine corine_int
/*-----
/* Intersect each of the corine point coverages with
/* the base coverage (eulkm + NUTS 5)

&watch corine_int.wat &append

&sv covlist = 1 2 3 4 5 6 7 8 9 10 11 12a 12b 13 14 15 16 17 18 19 20 21 22 ~
23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 49

&do count &list %covlist%
    intersect lc%count%_idw eulkm_n5 lc%count%_int point
&end

&watch &off

&return

/*-----
&routine corine_stats
/*-----
/* Sum land cover points (by class) up to NUTS 5 level to create three
/* variables: intensity, lit and unlit

/* eun5_var is a copy of the eu_n5popc coverage (NUTS 5 population). This
/* coverage will contain all the regression variables and be brought into
/* SPSS.

/* Light variables description:
/* li = intensity
/* al = lit
/* au = unlit

/*indexitem eun5_var.pat eupop-id

&sv covlist = 1 2 3 4 5 6 7 8 9 10 11 12a 12b 13 14 15 16 17 18 19 20 21 22 23
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 49

&do count &list %covlist%

    TABLES
    select lc%count%_int.pat

    statistics eupop-id lc%count%_li.stat
    sum predicted
    end

    reselect lit = 1
    statistics eupop-id lc%count%_al.stat
    n
    sum lit
    end

    nselect
    statistics eupop-id lc%count%_au.stat
    n
    sum lit
    end

    &if [exists lc%count%_li.stat -INFO] &then
        &do
            select lc%count%_li.stat
            alter
            sum-predicted
            lc%count%_li

```



```

~
~
~
~
~
&end

&if [exists lc%count%_al.stat -INFO] &then
&do
    select lc%count%_al.stat
    alter
    sum-lit
    lc%count%_al
    ~
    ~
    ~
    ~
&end

&if [exists lc%count%_au.stat -INFO] &then
&do
    select lc%count%_au.stat
    calc sum-lit = frequency
    alter
    sum-lit
    lc%count%_au
    ~
    ~
    ~
    ~
&end

QUIT

&do light &list li al au
&if [exists lc%count%_%light%.stat -INFO] &then
&do
    indexitem lc%count%_%light%.stat eupop-id
    joinitem eun5_var.pat lc%count%_%light%.stat eun5_var.pat eupop-id
    TABLES
    kill lc%count%_%light%.stat
    QUIT
&end
&end

&end

TABLES
additem eun5_var.pat lc12_li 8 18 F 6
additem eun5_var.pat lc12_al 8 18 F 6
additem eun5_var.pat lc12_au 8 18 F 6

select eun5_var.pat
calc lc12_li = lc12a_li + lc12b_li
calc lc12_al = lc12a_al + lc12b_al
calc lc12_au = lc12a_au + lc12b_au
QUIT

&return

/*****

```


SPSS Session

This section outlines how the weights are calculated at the NUTS 5 level for each variable entering the model in SPSS. Sample calculations are only shown for land class 1, area lit.

1. Calculate the population attributed to each land cover point (or road segment).

$$lc1alpop = lc1_al * B$$

Where: $lc1_al$ is each lit point in land class 1 and B is the unstandardised B coefficient from the regression model.

2. Calculate the modelled population by summing the populations attributed to each land cover point or road segment.

$$modpop = constant + lc1alpop + etc.$$

Where: etc. represents the population (also calculated in step 1) for other variables entering the model.

3. Calculate the weight for each variable entering the model.

$$lc1alwt = (lc1alpop * (pop / modpop)) / lc1_al$$

Where: pop is the true NUTS 5 population.

4. As a check, the predicted population is then determined by summing the product of each weight and its respective land cover point (or road segment). The error, which should equal zero, is found by subtracting the true NUTS 5 population from the predicted population.

$$predpop = (lc1alwt * lc1_al) + etc.$$

$$error = preddpop - pop$$

Pop_model.aml

```
/* *****  
/* pop_model2_gb.aml  
/*  
/* Apply the weights from the SPSS regression model on a point basis  
/* to redistribute population to the 1km level  
/*  
/* Danielle Vienneau  
/* Modified 03.04  
/* *****  
  
/* Set the initial variables  
&sv country = GB  
&sv nut5cov = britain_end  
&sv nut5link = ward91  
&sv popitem = pop_new  
&sv worksp = f:\pop_dec04\gb  
&sv worksp2 = f:\pop_dec04  
  
&sv spssfile = gb_spss  
  
&sv covlist = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24  
25 26 27 28 29 30 31 32 33 34 35 36 37 38 49  
  
/* the weights included in spss file - full list, and broken down into  
variable type
```



```

&sv li_list = 2 1 23 10 20 21
&sv al_list = 2 1 49 3 18 23 10 11 20
&sv au_list = 2 1 49 3 18 23 11
&sv arab_list = 12 13 14 15 16 17 19
&sv trans_list = 4 5 6
&sv wecs_list = 7 8 9
&sv otfor_list = 22 24 25
&sv unimp_list = 26 27 28 29
&sv otunc_list = 30 31 32 33 34 35 36 37 38 39

&call part1
&call part2
&call part3
&call part4
/*&call part4a
/*&call part5

&return

/*-----
&routine part1
/*-----

copyinfo %worksp2%\eulkm.pat eulkm.lut

/* Get the "section file" and only keep records for the country of interest
copyinfo %worksp2%\eulkm_n5.pat eulkm_n5.lut
TABLES
select eulkm_n5.lut
reselect eupop-id cn [quote %country%]
nselect
purge
y
QUIT
indexitem eulkm_n5.lut eupop-id
indexitem eulkm_n5.lut eulkm_n5#
dropitem eulkm_n5.lut eulkm_n5.lut population

/* Copy the NUT5 coverage of interest and drop population. This file
/* will be joined to all others to assign the NUT5 code (ward91 or shn)
copyinfo %worksp2%\%nut5cov%.pat %nut5cov%.lut
dropitem %nut5cov%.lut %nut5cov%.lut %popitem%
indexitem %nut5cov%.lut eupop-id
joinitem eulkm_n5.lut %nut5cov%.lut eulkm_n5.lut eupop-id eupop-id
indexitem eulkm_n5.lut %nut5link%

/* Bring in the spss file which contains NUTS5 population and weights for
/* each NUTS5 area
dbaseinfo %spssfile%.dbf %spssfile%.lut
indexitem %spssfile%.lut %nut5link%

&return

/*-----
&routine part2
/*-----

/* Copy in all of the land cover classes that may contain people.
/* Only keep records for the country of interest.

&do cov &list %covlist%

copyinfo %worksp2%\lc%cov%_int.pat lc%cov%_%country%.lut

TABLES
select lc%cov%_%country%.lut
reselect eupop-id cn [quote %country%]
nselect

```



```

purge
y
QUIT

/* check if records. If yes, additems. If not, kill.

ARCPLLOT
clearselect
reselect lc%cov%_%country%.lut info lc%cov%_int# > 0
&sv recs = [extract 1 [show select lc%cov%_%country%.lut info ]]
QUIT
&if %recs% > 0 &then
  &do
    TABLES
    select lc%cov%_%country%.lut
    alter
    predicted
    lc%cov%_predicted
    ~
    ~
    ~
    ~
    lit
    lc%cov%_lit
    ~
    ~
    ~
    ~
    QUIT

    /* Add the NUTS 5 code and join spss file (contains weights)
    indexitem lc%cov%_%country%.lut eupop-id
    joinitem lc%cov%_%country%.lut %nuts5cov%.lut lc%cov%_%country%.lut ~
    eupop-id
    indexitem lc%cov%_%country%.lut %nuts5link%
    joinitem lc%cov%_%country%.lut %spssfile%.lut lc%cov%_%country%.lut ~
    %nuts5link%

    additem lc%cov%_%country%.lut lc%cov%_%country%.lut lc%cov%li_pop ~
    8 9 F
    additem lc%cov%_%country%.lut lc%cov%_%country%.lut lc%cov%al_pop ~
    8 9 F
    additem lc%cov%_%country%.lut lc%cov%_%country%.lut lc%cov%au_pop ~
    8 9 F
    additem lc%cov%_%country%.lut lc%cov%_%country%.lut lc%cov%_pop 8 9 F
  &end
&else
  &do
    TABLES
    kill lc%cov%_%country%.lut
    QUIT
  &end
&end

&return

/*-----
&routine part3
/*-----

/* Apply the weights to each point

TABLES
&do li &list %li_list%
  &if [exists lc%li%_%country%.lut -info] &then
    &do
      select lc%li%_%country%.lut
      reselect lc%li%_lit = 1

```



```

        calc lc%li%li_pop = lc%li%_predicted * lc%li%liwt
        &type lc%li%li_pop = lc%li%_predicted * lc%li%liwt
    &end
&end

&do al &list %al_list%
    &if [exists lc%al%_country%.lut -info] &then
    &do
        select lc%al%_country%.lut
        reselect lc%al%_lit = 1
        calc lc%al%al_pop = lc%al%alwt
        &type lc%al%al_pop = lc%al%alwt
    &end
&end

&do au &list %au_list%
    &if [exists lc%au%_country%.lut -info] &then
    &do
        select lc%au%_country%.lut
        reselect lc%au%_lit = 0
        calc lc%au%au_pop = lc%au%auwt
        &type lc%au%au_pop = lc%au%auwt
    &end
&end

/* Section for grouped land classes - May need modification each run

&do al &list %arab_list%
    &if [exists lc%al%_country%.lut -info] &then
    &do
        select lc%al%_country%.lut
        reselect lc%al%_lit = 1
        calc lc%al%al_pop = arabalwt
        &type lc%al%al_pop = arabalwt
    &end
&end

&do au &list %arab_list%
    &if [exists lc%au%_country%.lut -info] &then
    &do
        select lc%au%_country%.lut
        reselect lc%au%_lit = 0
        calc lc%au%au_pop = arabauwt
        &type lc%au%au_pop = arabauwt
    &end
&end

&do li &list %arab_list%
    &if [exists lc%li%_country%.lut -info] &then
    &do
        select lc%li%_country%.lut
        reselect lc%li%_lit = 1
        calc lc%li%li_pop = lc%li%_predicted * arabliwt
        &type lc%li%li_pop = lc%li%_predicted * arabliwt
    &end
&end

&do al &list %trans_list%
    &if [exists lc%al%_country%.lut -info] &then
    &do
        select lc%al%_country%.lut
        reselect lc%al%_lit = 1
        calc lc%al%al_pop = transalwt
        &type lc%al%al_pop = transalwt
    &end
&end

&do au &list %wecs_list%

```



```

        &if [exists lc%au%_%country%.lut -info] &then
        &do
            select lc%au%_%country%.lut
            reselect lc%au%_lit = 0
            calc lc%au%au_pop = wecsauwt
            &type lc%au%au_pop = wecsauwt
        &end
    &end

    &do li &list %wecs_list%
        &if [exists lc%li%_%country%.lut -info] &then
        &do
            select lc%li%_%country%.lut
            reselect lc%li%_lit = 1
            calc lc%li%li_pop = lc%li%_predicted * wecsliwt
            &type lc%li%li_pop = lc%li%_predicted * wecsliwt
        &end
    &end

    &do au &list %otfor_list%
        &if [exists lc%au%_%country%.lut -info] &then
        &do
            select lc%au%_%country%.lut
            reselect lc%au%_lit = 0
            calc lc%au%au_pop = otforauwt
            &type lc%au%au_pop = otforauwt
        &end
    &end

    &do li &list %unimp_list%
        &if [exists lc%li%_%country%.lut -info] &then
        &do
            select lc%li%_%country%.lut
            reselect lc%li%_lit = 1
            calc lc%li%li_pop = lc%li%_predicted * unimpliwt
            &type lc%li%li_pop = lc%li%_predicted * unimpliwt
        &end
    &end

    &do au &list %otunc_list%
        &if [exists lc%au%_%country%.lut -info] &then
        &do
            select lc%au%_%country%.lut
            reselect lc%au%_lit = 0
            calc lc%au%au_pop = otuncauwt
            &type lc%au%au_pop = otuncauwt
        &end
    &end

    QUIT
&return

/*-----
&routine part4
/*-----

/* Sum up the populations for each point, and attach to wards and 1km
&do cov &list %covlist%
    &if [exists lc%cov%_%country%.lut -info] &then
    &do
        TABLES
        select lc%cov%_%country%.lut
        calc lc%cov%_pop = lc%cov%li_pop + lc%cov%al_pop + lc%cov%au_pop + -
        constwt
        &type lc%cov%_pop = lc%cov%li_pop + lc%cov%al_pop + lc%cov%au_pop + -
        constwt
        QUIT

        /* sum point population to wards - should equal

```



```

statistics lc%cov%_%country%.lut lc%cov%_ward91.stat %nut5link%
sum lc%cov%_pop
~
n
n

TABLES
select lc%cov%_ward91.stat
alter
sum-lc%cov%_pop
lc%cov%_pop
~
~
~
~
~
QUIT

indexitem lc%cov%_ward91.stat %nut5link%
joinitem %spssfile%.lut lc%cov%_ward91.stat %spssfile%.lut %nut5link%

/* sum point population to 1km
statistics lc%cov%_%country%.lut lc%cov%_1km.stat grid_code
sum lc%cov%_pop
~
n
n

TABLES
select lc%cov%_1km.stat
alter
sum-lc%cov%_pop
lc%cov%_pop
~
~
~
~
~
QUIT

indexitem lc%cov%_1km.stat grid_code
joinitem eulkm.lut lc%cov%_1km.stat eulkm.lut grid_code

&end
&end

&type CHECK ITEMS FOR CALCULATION
&return

/*-----
&routine part4a
/*-----

additem %spssfile%.lut %spssfile%.lut tmppopF 8 18 F 6
additem %spssfile%.lut %spssfile%.lut modpopF 8 18 F 6

TABLES
select %spssfile%.lut
calc tmppopF = lc1_pop + lc2_pop + lc3_pop + lc4_pop + lc5_pop + lc6_pop + ~
lc7_pop + lc8_pop + lc9_pop + lc10_pop + lc11_pop + lc12_pop + lc16_pop + ~
lc18_pop + lc20_pop
calc modpopF = tmppopF + lc21_pop + lc23_pop + lc24_pop + lc25_pop + lc26_pop
+ ~
lc27_pop + lc29_pop + lc30_pop + lc31_pop + lc35_pop + lc36_pop + lc37_pop + ~
lc49_pop
QUIT

```



```

additem eulkm.lut eulkm.lut tmppopF 8 18 F 6
additem eulkm.lut eulkm.lut modpopF 8 18 F 6

```

```

TABLES

```

```

select eulkm.lut
calc tmppopF = lc1_pop + lc2_pop + lc3_pop + lc4_pop + lc5_pop + lc6_pop + ~
lc7_pop + lc8_pop + lc9_pop + lc10_pop + lc11_pop + lc12_pop + lc16_pop + ~
lc18_pop + lc20_pop
calc modpopF = tmppopF + lc21_pop + lc23_pop + lc24_pop + lc25_pop + ~
lc26_pop + lc27_pop + lc29_pop + lc30_pop + lc31_pop + lc35_pop + lc36_pop + ~
lc37_pop + lc49_pop
QUIT

```

```

&type CHECK - WARD TOTALS AND 1KM TOTALS

```

```

&return

```

```

/*-----
&routine part5
/*-----

```

```

&data arc INFO
ARC
SELECT EU1KM.LUT
RESELECT MODPOPF > 0
EXPORT %WORKSP%\%COUNTRY%_1KMPop SDF NORMAL APPEND GRID_CODE,MODPOPF
Q STOP
&end

```

```

&sv thePath = [pathname %country%_1kmpop -FILE]

```

```

TABLES
define %country%_1kmpop.lut
grid_code 4 8 B
%country%_modpop 8 18 F 6
~
~
select %country%_1kmpop.lut
add from %thePath%
QUIT

```

```

indexitem %country%_1kmpop.lut grid_code

```

```

TABLES
kill lc* noprompt
QUIT

```

```

w %worksp2%
copyinfo %worksp%\%country%_1kmpop.lut %country%_1kmpop.lut
joinitem eulkm_pop.pat %country%_1kmpop.lut eulkm_pop.pat grid_code

```

```

&return

```

```

/*****

```

After this AML is run for each country, an item called modpop is added to the eu1km_pop coverage. The country values are summed across the study area on a cell basis to attribute modpop.

Lut_2_Point.aml

```
/* *****  
/* lut_2_point.aml  
/*  
/* Attach the 200m population values back to the LC points.  
/*  
/* Danielle Vienneau  
/* 03.01.05  
/* *****  
  
&call join_pt  
&call drop  
&call appnd  
&call int_ed  
&return  
  
/*-----  
&routine join_pt  
/*-----  
  
&do count &list 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 ~  
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 49  
  &if [exists lc%count%_gb.lut -info] &then  
    &do  
  
      w f:\pop_dec04  
  
      /* Put the GB points for the LC class into new coverage  
      additem lc%count%_int.pat lc%count%_int.pat int# 4 5 B  
      TABLES  
      select lc%count%_int.pat  
      calc int# = lc%count%_int#  
      QUIT  
  
      ARCEDIT  
      ec lc%count%_int  
      ef point  
      select all  
      reselect eupop-id cn [quote GB]  
      put f:\pop_dec04\gb\lc%count%_gb  
      QUIT  
  
      w f:\pop_dec04\gb  
  
      TABLES  
      select lc%count%_gb.pat  
      alter  
      int#  
      lc%count%_int#  
      ~  
      ~  
      ~  
      ~  
      QUIT  
  
      /* Join the 200m population to the points  
  
      indexitem lc%count%_gb.lut lc%count%_int#  
      indexitem lc%count%_gb.pat lc%count%_int#  
  
      joinitem lc%count%_gb.pat lc%count%_gb.lut lc%count%_gb.pat  
      lc%count%_int#  
      &end  
  
&end  
&return
```



```

/*-----
&routine drop
/*-----
w f:\pop_dec04\gb
&do count &list 2 3 4 5 6 7 8 9 10 11 16 18 20 21 23 24 25 26 27 29 30 31 35 ~
36 37 49

dropitem lc%count%_gb.pat lc%count%_gb.pat lc%count%_int# ~
lc%count%_int-id lc%count%_idw# lc%count%_idw-id measured predicted ~
nurgcd nurgcdl0 code nurgcdl1 nurgcdl2 arrgcdl0 group pool measscale ~
predscale lit

dropitem lc%count%_gb.pat lc%count%_gb.pat eulkm_n5# eulkm_n5-id eulkm# ~
eulkm-id light1 light3 light5 eu_n5popc# eu_n5popc-id lc%count%_predicted ~
lc%count%_lit britain_end# britain_end-id nuts3v7eu# nuts3v7eu-id

dropitem lc%count%_gb.pat lc%count%_gb.pat lc2alwt lc2auwt lc2liwt ~
lc1alwt lc1auwt lc1liwt lc49alwt lc49auwt arabalwt arabauwt arabliwt ~
lc3alwt lc3auwt lc18alwt lc18auwt lc23alwt lc23auwt lc23liwt lc10alwt ~
lc10liwt

dropitem lc%count%_gb.pat lc%count%_gb.pat lc11alwt lc11auwt transalwt ~
lc20alwt lc20liwt lc21liwt wecsauwt wecsliwt otforauwt unimpliwt ~
otuncauwt constwt lc%count%al_pop lc%count%au_pop lc%count%li_pop

TABLES
select lc%count%_gb.pat
alter
lc%count%_pop
modpopF
~
~
~
~
~
QUIT
&end

&do count &list 1
dropitem lc%count%_gb.pat lc%count%_gb.pat lc%count%_int# ~
lc%count%_int-id lc%count%_idw# lc%count%_idw-id measured predicted ~
nurgcd nurgcdl0 code nurgcdl1 nurgcdl2 arrgcdl0 group pool measscale ~
predscale lit

dropitem lc%count%_gb.pat lc%count%_gb.pat eulkm_n5# eulkm_n5-id eulkm# ~
eulkm-id light1 light3 light5 eu_n5popc# eu_n5popc-id lc%count%_predicted ~
lc%count%_lit britain_end# britain_end-id

dropitem lc%count%_gb.pat lc%count%_gb.pat lc2alwt lc2auwt lc2liwt ~
lc1alwt lc1auwt lc1liwt lc49alwt lc49auwt arabalwt arabauwt arabliwt ~
lc3alwt lc3auwt lc18alwt lc18auwt lc23alwt lc23auwt lc23liwt lc10alwt ~
lc10liwt

dropitem lc%count%_gb.pat lc%count%_gb.pat lc11alwt lc11auwt transalwt ~
lc20alwt lc20liwt lc21liwt wecsauwt wecsliwt otforauwt unimpliwt ~
otuncauwt constwt lc%count%al_pop lc%count%au_pop lc%count%li_pop

TABLES
select lc%count%_gb.pat
alter
lc%count%_pop
modpopF
~
~
~
~
~
QUIT

```



```

&end

&do count &list 12
  dropitem lc%count%_gb.pat lc%count%_gb.pat lc%count%_int# ~
  lc%count%_int-id lc%count%a_idw# lc%count%a_idw-id measured ~
  predicted nurgcd nurgcd10 code nurgcd11 nurgcd12 arrgcd10 group ~
  pool measscale predscale lit new-id

  dropitem lc%count%_gb.pat lc%count%_gb.pat eulkm_n5# eulkm_n5-id eulkm# ~
  eulkm-id light1 light3 light5 eu_n5popc# eu_n5popc-id lc%count%_predicted ~
  lc%count%_lit britain_end# britain_end-id nuts3v7eu# nuts3v7eu-id

  dropitem lc%count%_gb.pat lc%count%_gb.pat lc2alwt lc2auwt lc2liwt ~
  lc1alwt lc1auwt lc1liwt lc49alwt lc49auwt arabalwt arabauwt arabliwt ~
  lc3alwt lc3auwt lc18alwt lc18auwt lc23alwt lc23auwt lc23liwt lc10alwt ~
  lc10liwt

  dropitem lc%count%_gb.pat lc%count%_gb.pat lc11alwt lc11auwt transalwt ~
  lc20alwt lc20liwt lc21liwt wecsauwt wecsliwt otforauwt unimpliwt ~
  otuncauwt constwt lc%count%al_pop lc%count%au_pop lc%count%li_pop

  TABLES
  select lc%count%_gb.pat
  alter
  lc%count%_pop
  modpopF
  ~
  ~
  ~
  ~
  ~
  QUIT
&end

&return

/*-----
&routine appnd
/*-----
w f:\pop_dec04\gb

append pop200 lc1_gb
lc1_gb
lc2_gb
lc3_gb
lc4_gb
lc5_gb
lc6_gb
lc7_gb
lc8_gb
lc9_gb
lc10_gb
lc11_gb
lc12_gb
lc16_gb
lc18_gb
lc20_gb
lc21_gb
lc23_gb
lc24_gb
lc25_gb
lc26_gb
lc27_gb
lc29_gb
lc30_gb
lc31_gb
lc35_gb
lc36_gb

```



```
lc37_gb
lc49_gb
end
```

```
&return
```

```
/*-----
&routine int_ed
/*-----
```

```
/* Project 200m points to BNG and intersect with EDs.
```

```
project cover pop200 pop200p o:\projects\apmosphere\projection\la-bng2.prj
intersect pop200p ed91 ed200m_int point 0.001
```

```
statistics ed200m_int.pat ed200m_pop.stat ed91
sum modpopF
end
```

```
TABLES
select ed200m_pop.stat
alter
sum-modpopF
modpopF
```

```
~
~
~
~
~
```

```
QUIT
```

```
indexitem ed200m_pop.stat ed91
indexitem ed91.pat ed91
indexitem ed_pop91_gb ed91
```

```
joinitem ed91.pat ed200m_pop.stat ed91.pat ed91
joinitem ed91.pat ed_pop91_gb ed91.pat ed91
```

```
statistics ed91.pat ed_ptvalid.stat ed91
mean modpopF
mean pop91
sum area
end
```

```
&return
```

```
/******
```

Population Validation - Postcodes

1. The total ED population was distributed evenly among all postcodes in each ED.

```
Arc: |> statistics pc91_edpop.pat freq_pc91.stat ed91 <|
Statistics: |> sum pop91 <|
Statistics: |> end <|
Arc: |> dropitem freq_pc91.stat freq_pc91.stat sum-pop91 <|
Arc: |> tables <|
Tables: |> alter <|
Enter Item Name: |> frequency <|
Item Name: |> freq_pc91 <|
Item Output Width: |><|
Item Type: |><|
Alternate Item Name: |><|
Enter Item Name: |><|
Tables: |> q <|
```



```

Arc: |> indexitem freq_pc91.stat ed91 <|
Arc: |> joinitem pc91_edpop.pat freq_pc91.stat pc91_edpop.pat ed91 <|
Arc: |> additem pc91_edpop.pat pc91_edpop.pat pcpop 8 9 F <|

Arc: |> tables <|
Tables: |> select pc91_edpop.pat <|
Tables: |> reselect freq_pc91 > 0 <|
Tables: |> calc pcpop = ( 1 / freq_pc91 ) * pop91 <|
Tables: |> statistics <|
Statistics: |> sum pcpop <|
Statistics: |> end <|
Tables: |> q <|

```

2. The postcodes (with population attached) and the 200m modelled population points were intersected with the 1km British national grid. For both, the statistics command was used to calculate the total population within each 1km square in Britain.

```

Arc: |> intersect pc91_edpop bng_1km pc91_1km point 0.001 <|
Arc: |> intersect pop200p_gb bng_1km pop200_1km point 0.001 <|
Arc: |> statistics pc91_1km.pat pc91_1km.stat cellid_1km <|
Statistics: |> sum pcpop <|
Statistics: |> end <|

Arc: |> statistics pop200_1km.pat pop200_1km.stat cellid_1km <|
Statistics: |> sum gb_modpop <|
Statistics: |> sum gb91_modpop <|
Statistics: |> end <|

Arc: |> indexitem pc91_1km.stat cellid_1km <|
Arc: |> indexitem pop200_1km.stat cellid_1km <| .

```

3. The statistics files were joined back to the British national grid. This file was then exported from ArcGIS and imported into SPSS where the correlations were examined.

```

Arc: |> joinitem bng_1km.pat pc91_1km.stat bng_1km.pat cellid_1km <|
Arc: |> joinitem bng_1km.pat pop200_1km.stat bng_1km.pat cellid_1km <|

Arc: |> items bng_1km.pat <|

```

COLUMN INDEXED?	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC	ALTERNATE NAME
1	AREA	4	12	F	3	-
5	PERIMETER	4	12	F	3	-
9	BNG_1KM#	4	5	B	-	-
13	BNG_1KM-ID	4	5	B	-	-
17	CELLID_1KM	4	10	B	-	Indexed
21	SUM-PCPOP	8	18	F	6	-
29	SUM-GB_MODPOP	8	18	F	6	-
37	SUM-GB91_MODPOP	8	18	F	6	-

Population Validation – Enumeration Districts

1. The 200m modelled population points were intersected with the EDs then the total modelled population for each ED was calculated using the statistics command. The statistics file was exported from ArcGIS and imported into SPSS for analysis of the correlations. NOTE: not all EDs are represented by the 200m points.

```

Arc: |> intersect pop200p_gb ed91 pop200_ed point 0.001 <|
Arc: |> joinitem pop200_ed.pat ed_pop91_gb pop200_ed.pat ed91 <|
Arc: |> statistics pop200_ed.pat pop200_eds.stat ed91 <|
Statistics: |> mean pop91 <|
Statistics: |> sum gb_modpop <|
Statistics: |> sum gb91_modpop <|
Statistics: |> end <|

```


B2. Population modelling results

Table 1
Strategy 1 Country & EU Models

Land Cover Class		Adj. R ²	Std Error	Constant	lc1_al	lc1_au	lc1_li	lc2_al	lc2_au	lc2_li	lc3_al	lc3_au	lc3_li	trans_al	trans_au	trans_li	wecs_al	wecs_au	wecs_li	lc10_al	lc10_au	lc10_li	lc11_al	lc11_au	lc11_li	arab_al	arab_au
		0.82	4518	234.4	1744.6	187.0	0.023	98.9	85.4	0.007	141.2	95.7		54.0			13.3		0.009	128.1		0.023	62.6	19.7		0.4	0.2
		0.90	2581	7255	4246	1256.4		155.3	73.9	0.003	265.1	231.0								2001.6	94.3			147.9			
		0.87		1744.6	67.4	491.4					221.3								0.028	376.5							
		0.97	4246	860.4	1085.0			99.4		0.019	86.6		0.026							39.3	1108.2	0.033					3.3
		0.84	25414	934.9				33.4	37.1	0.034	97.3									42.7		0.036					
		0.95	4071	533.8				50.9	25.9	0.012	97.3	209.4								145.8	18.0			2559.6			
		0.81	7886					102.3	77.0		288.5	167.0								102.0							
		0.83	2794						52.6	0.027	151.0	108.1									395.5	38.2					
		0.73	2033					114.3	70.8	0.008				50.9													
		0.82	6644					66.4	94.1	0.012												0.019	288.1				
		0.73											0.005														
		0.82																									
		0.73																									
		0.83																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									
		0.82																									
		0.73																									

crops	arab_li	0.005				0.011		0.007		0.005		0.008		0.011	
18: Pasture	lc18_al	1.6	4.7							0.9					
	lc18_au	0.6	2.9			2.8	4.6						0.3		
	lc18_li			0.027				0.035					0.001		0.003
20: Complex cultivation patterns	lc20_al	1.0	3.0			3.6		5.1				17.0	1.8		
	lc20_au					4.4				1.5				3.8	
	lc20_li	0.013	0.049			0.068		0.013				0.010		0.018	
21: Land principally occupied by agriculture	lc21_al		30.2				4.8		20.7	16.9	21.6				
	lc21_au					14.6						2.3	0.6		
	lc21_li	0.017		0.051		0.143		0.016				0.054		0.016	
22,24,25: Other woodland	otfor_al					2.0			0.6						
	otfor_au	0.0	0.2			0.4		0.5	0.1						
	otfor_li		0.016		0.102	0.025				0.004					
23: Broad-leaved woodland	lc23_al	8.8													34.9
	lc23_au	0.2				3.8									
	lc23_li	0.005				0.067	0.087			0.005			0.646	0.012	
26-29: Unimproved grass and scrub	unimp_al							4.3		1.1				2.0	
	unimp_au														
	unimp_li	0.009							0.027				0.050		
30-38: Other uncultivated	otunc_al														
	otunc_au	0.4													
	otunc_li							0.020							
49: Unclassified surfaces	lc49_al	19.1								41.1		26.9	55.3		
	lc49_au	1.0									12.5	0.5		53.4	
	lc49_li							0.422							

Table 2 Strategy 2 Country & EU Models

Land Cover Class	Adj. R ²	EU	AT	BELU	west DE	east DE	DK	ESPT	FI	FR	GB	GR	IENI	IT	NL
Constant	Std Error	4551	2702	7451	4670	26200	4097	8117	5555	2848	2032	6835	801	7449	7576
1: Continuous urban	lc1_al	87.3	140.2	2368.3	427.5	42977.5	3413.9	125.0	2150.2	82.4	1597.2	1586.7	414.0	30.0	1087.6
	lc1_au	187.5		470.9	988.4	770.2	545.7	251.7	366.2	394.0	167.3	664.9	116.3		192.0
	lc1_li	0.023	1288.1		1205.9			0.049		0.013	0.003		0.002	200.0	178.6
2: Discontinuous urban	lc2_al	100.7	176.6		106.5	39.1	49.9	115.6			115.0	66.7	98.7	129.9	52.3
	lc2_au	84.8	71.1			37.8	25.0	75.5		52.8	71.1	94.8	100.7	121.7	112.2
	lc2_li	0.007	0.003	0.026	0.018	0.036	0.013		0.005	0.027	0.008	0.012	0.006	0.012	
3: Industrial / commercial	lc3_al	145.5	229.9	212.1	86.9	110.8	97.5	303.9	65.3	162.3				63.3	129.9
	lc3_au	95.0	234.8	327.4			213.8	197.4		99.3		49.9		42.2	
	lc3_li				0.064								0.005		
4-9: Other urban	oturb_al	9.1		26.4				94.2	32.9	19.6			15.1		
	oturb_au	11.0			61.1								13.3	92.6	
	oturb_li	0.006					0.010		0.011					0.054	0.046
10,11: Open urban	opurb_al	119.3		45.3				582.1				209.6			
	opurb_au	24.0		149.9	59.7				2941.5	115.4				129.2	
	opurb_li		0.170		0.050		0.024			0.020	0.006		0.001	0.074	0.049
	arab_al	0.6						3.1	2.3			5.1			
12-17,19: Arable and permanent crops	arab_au	0.2											1.1		
	arab_li	0.005			0.010			0.008		0.005		0.012		0.013	
	lc18_al	1.3	5.7							0.9					
18: Pasture	lc18_au	0.5	3.3		2.8		4.9						0.2		
	lc18_li			0.035				0.049					0.001		
20,21: Other	otcul_al		5.9									12.1	1.3		

cultivated	otcul_au						8.0					1.3		1.0	0.9	
	otcul_li	0.015					0.093				0.094	0.003		0.014		0.018
	wood_al								10.0	1.7		1.2				5.2
22-25: Woodland	wood_au	0.0					1.8				0.1					
	wood_li		0.010				0.059					0.006				0.007
26-29: Unimproved grass and scrub	unimp_al									6.5		1.2	1.1			
	unimp_au															
	unimp_li	0.009									0.027		0.012		0.047	
30-38: Other uncultivated	otunc_al															
	otunc_au	0.4								0.8						
	otunc_li									0.021	0.564					
49: Unclassified surfaces	lc49_al	18.2										54.4		28.8	53.1	
	lc49_au	1.0											12.3			52.9
	lc49_li									0.484						

Table 3 Strategy 3 Country & EU Models

Land Cover Class		EU	AT	BELU	west DE	east DE	DK	ESPT	FI	FR	GB	GR	IENI	IT	NL
	Adj. R ²	0.81	0.92	0.81	0.97	0.83	0.95	0.80	0.97	0.82	0.73	0.81	0.73	0.85	0.92
Constant	Std Error	4634	2362	8708	4611	26002	4367	8140	5367	2862	2037	6889	813	7611	7632
1: Continuous urban		60.6	374.6	1804.6	27.8	44722.7	2819.8	58.3	2989.4	32.1	1602.3	2017.5	432.9	84.6	1036.3
	lc1_al	191.9	662.1	616.7	927.1	694.3	579.4		385.9	402.0	167.4	667.1	114.0	196.3	195.6
	lc1_au	145.2	1404.3		910.7			253.0			161.1	20.5			175.9
2: Discontinuous urban	lc1_li	0.023						0.047		0.013	0.003		0.002	0.042	0.006
	lc2_al	104.9	148.7		106.5	47.2	50.7	114.3			115.2	64.5	100.7	124.6	48.4
	lc2_au	89.3	64.2	25.6		33.1	29.5	88.7		67.4	70.8	99.4	100.2	112.8	107.1
3-11: Non-residential urban	lc2_li	0.007		0.033	0.027	0.029	0.011		0.002	0.028	0.009	0.012	0.006	0.016	
	nrurb_al	91.2			66.0		56.2	236.1		116.4				65.3	62.2
	nrurb_au	22.3		140.0	69.1	15.5			52.5				12.5	78.9	
12-21: Cultivated	nrurb_li	0.002	0.017	0.011	0.018	0.046							0.003	0.009	0.020
	cul_al	0.7		3.2					1.5	0.7		8.6			
	cul_au	0.2	1.5		3.4		0.4	0.3		0.5			0.3		
22-38: Uncultivated	cul_li	0.007			0.014			0.012		0.004		0.013	0.001	0.012	
	ucul_al						7.0	2.7	0.5		1.1			1.6	
	ucul_au	0.1	0.3		0.8										
49: Unclassified surfaces	ucul_li			0.031	0.071				0.006	0.003	0.005		0.014	0.005	
	lc49_al	20.1								49.1		27.8	59.9		
	lc49_au	1.0									12.3			51.2	
	lc49_li							0.528							

C1. Cluster analysis – map granularity index

Cluster_index.aml

```
/* *****  
**  
/* Cluster_index.aml  
/*  
/* Calculates map granularity by counting the number of same cells within a  
/* rectangle window. The remap tables simply change each of the cluster  
/* numbers  
/* to 1 (eg. clus6.remap -> 6 : 1).  
/*  
/* &r cluster_index <pollutant> <>window size>  
/*  
/* Danielle Vienneau  
/* 08.05  
/* *****  
**  
&args poll neigh  
  
&sv fileunit = [open cluster_info_&neigh.txt openstat -write]  
  
GRID  
&do clust &list 6 8 10 12 14 16 18 20 22 24 26  
  
  &poll_&clust = polygrid(&poll_group, qcl_&clust, N, N, 1000)  
  &sv count = &clust  
  &do &while &count > 0  
  
    &poll_&clustsol_&count = reclass(&poll_&clust, clustcount.remap, -  
    nodata)  
    docell  
    dsum += &poll_&clustsol_&count  
    end  
    &sv sumD = [show dsum]  
  
    fs_&clustsol_&count = focalsum(&poll_&clustsol_&count, rectangle, -  
    &neigh, &neigh)  
    num_&clustsol_&count = fs_&clustsol_&count - -  
    &poll_&clustsol_&count  
  
    docell  
    nsum += num_&clustsol_&count  
    end  
    &sv sumN = [show nsum]  
  
    &type The numerator is &sumN and the denominator is &sumD  
    &sv writestat = [write &fileunit &clust, &count, &sumN, &sumD]  
  
    kill &poll_&clustsol_&count all  
    kill fs_&clustsol_&count all  
    kill num_&clustsol_&count all  
  
    &sv count = &count - 1  
  &end  
&end  
QUIT  
  
&sv closestat = [close -all]  
  
&sv thePath = [pathname cluster_info_&neigh.txt -FILE]  
TABLES  
define cluster_info_&neigh.lut  
Solution 3 3 N 0  
Clus_num 3 3 N 0  
Numer 12 12 N 0  
Denom 12 12 N 0
```


~
~

```
select cluster_info_%neigh%.lut  
add from %thePath%
```

```
additem cluster_info_%neigh%.lut cindex 8 9 F 2  
calc cindex = Numer / Denom  
QUIT
```

```
statistics cluster_info_%neigh%.lut cluster_info_%neigh%.stat solution  
min cindex  
max cindex  
mean cindex  
std cindex  
sum cindex  
end
```

```
&return  
/*****
```


D1. Weighted wind direction

Calcwind_month.aml

```
/* *****  
/* Calculate monthly wind direction and wind speed from the U and  
/* V components  
/* &r calcwind_month  
/*  
/* Danielle Vienneau  
/* March 05  
/* *****  
  
&call calc_wind_month  
&call wind_code  
&call sum_code  
&call mk_grds  
&return  
  
/*-----  
&routine calc_wind_month  
/*-----  
/* Calculates monthly wind direction and speed from the U and V components  
  
&sv windcov = wind_uv  
  
/* Get the size of the point coverage  
ARCPLLOT  
clearselect  
reselect %windcov% point $recno ge 0  
&sv totpts = [extract 1 [show select %windcov% point] ]  
QUIT  
  
/* Loop through months, then through the points  
&do month &list 1 2 3 4 5 6 7 8 9 10 11 12  
  
    /* Open a text file to hold the output for this month  
    &sv fileunit = [open winddir_month%month%.txt openstat -write]  
  
    ARCPLLOT  
    &sv count = 0  
    &do &until %count% = %totpts%  
  
        &sv count = %count% + 1  
  
        &sv Ucomp = [show select %windcov% point %count% ITEM U10_%month%]  
        &sv Vcomp = [show select %windcov% point %count% ITEM V10_%month%]  
  
        &if [calc %Vcomp% = 0] &then  
            &sv windrad = [calc 0.5 * 3.14159265]  
        &else  
            &do  
                &sv tanval = [abs [calc %Ucomp% / %Vcomp%] ]  
                &sv windrad = [ATAN %tanval% ]  
            &end  
  
            &sv uv = [calc [calc %Ucomp% * %Ucomp%] + ~  
                [calc %Vcomp% * %Vcomp%] ]  
            &sv windspeed = [sqrt %uv%]  
  
            /* NE quadrant  
            &if [calc %Ucomp% >= 0] and [calc %Vcomp% >= 0] &then  
                &sv winddir = [calc 57.2957795 * %windrad%]  
  
            /* SE quadrant  
            &if [calc %Ucomp% >= 0] and [calc %Vcomp% < 0] &then  
                &sv winddir = [calc 180 - [calc 57.2957795 * %windrad%] ]
```



```

/* SW quadrant
&if [calc %Ucomp% < 0] and [calc %Vcomp% < 0] &then
    &sv winddir = [calc 180 + [calc 57.2957795 * %windrad%] ]

/* SNE quadrant
&if [calc %Ucomp% < 0] and [calc %Vcomp% >= 0] &then
    &sv winddir = [calc 360 - [calc 57.2957795 * %windrad%] ]

&sv pointid = [show select %windcov% point %count% ITEM pointid]

/* Ensure variable exists, and write it to the text file
&if [variable pointid] = .FALSE. &then
    &sv windx = 0
&if [variable winddir] = .FALSE. &then
    &sv winddir = 0
&if [variable windspeed] = .FALSE. &then
    &sv windspeed = 0

    &sv writestat = [WRITE %fileunit% %pointid%,%winddir%,%windspeed%]

&end

&sv closestat = [close -all]
QUIT

&sv thePath = [pathname winddir_month%month%.txt -FILE]
TABLES
define winddir_month%month%.lut
POINTID 8 8 F 0
WDIR_%month% 8 9 F 3
WSPEED_%month% 8 9 F 3
~
~

select winddir_month%month%.lut
add from %thePath%
QUIT

&end

&do month &list 1 2 3 4 5 6 7 8 9 10 11 12

    joinitem wind_ds2.pat winddir_month%month%.lut wind_ds2.pat pointid

&end
&return

/*-----
&routine wind_code
/*-----
/* Assigns a wind direction code to each cell. There are 8 possible
/* wind direction segments.

&sv windcov = wind_ds2

/* Get the size of the point coverage
ARC PLOT
clearselect
reselect %windcov% point $recno ge 0
&sv totpts = [extract 1 [show select %windcov% point] ]
QUIT

/* Loop through months, then through the points
&do month &list 1 2 3 4 5 6 7 8 9 10 11 12
    &sv fileunit2 = [open winddir_cd%month%.txt openstat -write]
    ARC PLOT

        &sv count = 0
        &do &until %count% = %totpts%

```



```

&sv count = %count% + 1
&sv pointid = [show select %windcov% point %count% ITEM pointid]
&sv dircd = [show select %windcov% point %count% ITEM wdir_%month%]

&if [calc %dircd% > 22.5] and [calc %dircd% <= 67.5] &then
  &sv wdircd = 1
&else &if [calc %dircd% > 67.5] and [calc %dircd% <= 112.5] &then
  &sv wdircd = 2
&else &if [calc %dircd% > 112.5] and [calc %dircd% <= 157.5] &then
  &sv wdircd = 3
&else &if [calc %dircd% > 157.5] and [calc %dircd% <= 202.5] &then
  &sv wdircd = 4
&else &if [calc %dircd% > 202.5] and [calc %dircd% <= 247.5] &then
  &sv wdircd = 5
&else &if [calc %dircd% > 247.5] and [calc %dircd% <= 292.5] &then
  &sv wdircd = 6
&else &if [calc %dircd% > 292.5] and [calc %dircd% <= 337.5] &then
  &sv wdircd = 7
&else &sv wdircd = 8
&sv writestat = [WRITE %fileunit2% %pointid%,%wdircd%]
&end
&sv closestat = [close -all]

QUIT

&sv thePath = [pathname winddir_cd%month%.txt -FILE]
TABLES
define winddir_cd%month%.lut
POINTID 8 8 F 0
WDIRCD_%month% 2 2 I
~
~

select winddir_cd%month%.lut
add from %thePath%
QUIT

joinitem %windcov%.pat winddir_cd%month%.lut %windcov%.pat pointid

&end
&return

/*-----
&routine sum_code
/*-----
/* Counts up the number of months for which wind is in each of the 8 segments.

&sv windcov = wind_ds2

/* Get the size of the point coverage
ARC PLOT
clearselect
reselect %windcov% point $recno ge 0
&sv totpts = [extract 1 [show select %windcov% point] ]
QUIT

&do seg &list 1 2 3 4 5 6 7 8

&sv fileunit3 = [OPEN sum_seg%seg%.txt openstat -write]
ARC PLOT
  &sv count = 0
  &do &until %count% = %totpts%
    &sv count = %count% + 1

    &sv pointid = [show select %windcov% point %count% ITEM pointid]

    &sv seg_tot = 0
    &do month &list 1 2 3 4 5 6 7 8 9 10 11 12

```



```

        &sv seg_num = [show select %windcov% point %count% ITEM ~
        wdircd_%month%]
        &if [calc %seg_num% = %seg%] &then
            &sv seg_tot = %seg_tot% + 1
        &end

        &sv writestat = [WRITE %fileunit3% %pointid%,%seg_tot%]
    &end
QUIT
&sv closestat = [close -all]

&sv thePath = [pathname sum_seg%seg%.txt -FILE]
TABLES
define sum_seg%seg%.lut
POINTID 8 8 F 0
SEG_%seg% 2 2 I
~
~

select sum_seg%seg%.lut
add from %thePath%
QUIT

joinitem %windcov%.pat sum_seg%seg%.lut %windcov%.pat pointid
&end
&return

/*-----
&routine mk_grds
/*-----
/* Create 8 grids that represent monthly weighted wind direction (one for
/* each segment)

&sv windcov = wind_ds2

joinitem wind20_dirp.pat %windcov%.pat wind20_dirp.pat pointid

&sv cnt = 8

&do &while %cnt% > 0
    polygrid wind20_dirp dirseg_%cnt% seg_%cnt%
    1000
    y

    &sv cnt = %cnt% - 1
&end
&return
/*****

```


D2. Focalsum AMLs

Focalsum_annulus.aml

```
/******  
/* Create doughnut windows using kernels based on the ANNULUS option  
/* &r focalsum_annulus  
/*  
/* Danielle Vienneau  
/* March 05  
/******  
  
&sv path m:\projects\apmosphere\focalsum  
  
&call total_grd  
&call make_grds  
&call site_intersect  
&call create_map  
&return  
  
/*-----  
&routine total_grd  
/*-----  
/* Calculate total grid (excludes point sources)  
  
&do poll &list nox pm10  
  GRID  
  %poll%_tot1 = %poll%c21 + %poll%c31 + %poll%c41 + %poll%c51 + %poll%c61  
  QUIT  
  
&return  
  
/*-----  
&routine make_grds  
/*-----  
&do poll &list nox pm10  
  w %path%\%poll%\annulus  
  copy %path%\nox_tot1 nox_tot1  
  &do ker &list 3 5 7 9 11 13 15 17 19 21  
    GRID  
    %poll%_tot%ker% = focalsum (nox_tot1, irregular,  
%path%\ann%ker%.ker)  
    QUIT  
  &end  
&end  
&return  
  
/*-----  
&routine site_intersect  
/*-----  
/* Assign values to training & validation points using grid_spot70.aml  
/* which uses  
/* the cellvalue command to intersect sites with the grids  
&do poll &list nox pm10  
  &do ker &list 1 3 5 7 9 11 13 15 17 19 21  
    &r %path%\gridspot70 %path%\%poll%\annulus\%poll%_tot%ker% ~  
    %path%\%poll%\annulus\%poll%_sites %poll%_tot%ker%  
  &end  
&end  
&return  
/* Use SPSS, to calibrate model  
  
/*-----  
&routine create_map  
/*-----  
/* Apply weights from regression analysis to the emission grids, and then  
/* use calibration equation to create a final concentration grid
```


GRID

/* Customise weights!

```
predno2 = (%poll%_tot1 + %poll%_tot3 + (0.5 * %poll%_tot5) + ~
(0.028 * (%poll%_tot7 + %poll%_tot9 + %poll%_tot11)) + ~
(0.01 * (%poll%_tot13 + %poll%_tot15 + %poll%_tot17 + %poll%_tot19 + ~
%poll%_tot21)))
```

/* Customise calibration function!

```
modno2_bk = exp (1.820 + (0.241 * ln(predno2)))
```

QUIT

&return

/******

Focalsum_wwind.aml

/******

/* Create wedge windows, weighted by monthly wind direction frequency

/* &r focalsum_wwind

/*

/* Danielle Vienneau

/* March 05

/******

&sv poll = nox

&call total_grd

&call wind_weight

&call weight_1/d2

/* only one weight routine is called

/*&call weight_1/d

/*&call weight_1/ln(d)

/*&call weight_1/d3

&call site_intersect

&call create_map

&return

/*-----

&routine total_grd

/*-----

/* Calculate total grid (excludes point sources)

GRID

```
%poll%_nptot = %poll%c21 + %poll%c31 + %poll%c41 + %poll%c51 + %poll%c61
```

QUIT

&return

/*-----

&routine wind_weight

/*-----

GRID

```
&do rad &list 1 2 3 4 5 6 7 8 9 10 11
```

/* Calculate each wedge for all radii (fixed wind directions)

```
seg5_%rad% = focalsum(%poll%_nptot,wedge,%rad%,22.5,67.5)
```

```
seg4_%rad% = focalsum(%poll%_nptot,wedge,%rad%,67.5,112.5)
```

```
seg3_%rad% = focalsum(%poll%_nptot,wedge,%rad%,112.5,157.5)
```

```
seg2_%rad% = focalsum(%poll%_nptot,wedge,%rad%,157.5,202.5)
```

```
seg1_%rad% = focalsum(%poll%_nptot,wedge,%rad%,202.5,247.5)
```

```
seg8_%rad% = focalsum(%poll%_nptot,wedge,%rad%,247.5,292.5)
```

```
seg7_%rad% = focalsum(%poll%_nptot,wedge,%rad%,292.5,337.5)
```

```
seg6_%rad% = focalsum(%poll%_nptot,wedge,%rad%,337.5,22.5)
```

/* Weight segments by wind direction frequency

```
seg1_%rad%w = seg1_%rad% * dirseg_1 / 12
```

```
seg2_%rad%w = seg2_%rad% * dirseg_2 / 12
```



```

    seg3_%rad%w = seg3_%rad% * dirseg_3 / 12
    seg4_%rad%w = seg4_%rad% * dirseg_4 / 12
    seg5_%rad%w = seg5_%rad% * dirseg_5 / 12
    seg6_%rad%w = seg6_%rad% * dirseg_6 / 12
    seg7_%rad%w = seg7_%rad% * dirseg_7 / 12
    seg8_%rad%w = seg8_%rad% * dirseg_8 / 12
&end
QUIT
&return

/*-----
&routine weight_1/d2
/*-----
GRID
&do seg &list 1 2 3 4 5 6 7 8
    /* Subtract the rings and apply distance weighting (1/d2)
    seg%seg%_2r = (seg%seg%_2w - seg%seg%_1w) * 0.25
    seg%seg%_3r = (seg%seg%_3w - seg%seg%_2w) * 0.111
    seg%seg%_4r = (seg%seg%_4w - seg%seg%_3w) * 0.0625
    seg%seg%_5r = (seg%seg%_5w - seg%seg%_4w) * 0.04
    seg%seg%_6r = (seg%seg%_6w - seg%seg%_5w) * 0.028
    seg%seg%_7r = (seg%seg%_7w - seg%seg%_6w) * 0.02
    seg%seg%_8r = (seg%seg%_8w - seg%seg%_7w) * 0.0156
    seg%seg%_9r = (seg%seg%_9w - seg%seg%_8w) * 0.0123
    seg%seg%_10r = (seg%seg%_10w - seg%seg%_9w) * 0.01
    seg%seg%_11r = (seg%seg%_11w - seg%seg%_10w) * 0.0083

    seg%seg%_tot = seg%seg%_1w + seg%seg%_2r + seg%seg%_3r + seg%seg%_4r + ~
    seg%seg%_5r + seg%seg%_6r + seg%seg%_7r + seg%seg%_8r + seg%seg%_9r + ~
    seg%seg%_10r + seg%seg%_11r
&end
QUIT
&return

/*-----
&routine weight_1/d
/*-----
GRID
&do seg &list 1 2 3 4 5 6 7 8
    /* Subtract the rings and apply distance weighting (1/d)
    seg%seg%_2r = (seg%seg%_2w - seg%seg%_1w) * 0.5
    seg%seg%_3r = (seg%seg%_3w - seg%seg%_2w) * 0.333
    seg%seg%_4r = (seg%seg%_4w - seg%seg%_3w) * 0.25
    seg%seg%_5r = (seg%seg%_5w - seg%seg%_4w) * 0.2
    seg%seg%_6r = (seg%seg%_6w - seg%seg%_5w) * 0.167
    seg%seg%_7r = (seg%seg%_7w - seg%seg%_6w) * 0.143
    seg%seg%_8r = (seg%seg%_8w - seg%seg%_7w) * 0.125
    seg%seg%_9r = (seg%seg%_9w - seg%seg%_8w) * 0.111
    seg%seg%_10r = (seg%seg%_10w - seg%seg%_9w) * 0.1
    seg%seg%_11r = (seg%seg%_11w - seg%seg%_10w) * 0.091

    seg%seg%_tot = seg%seg%_1w + seg%seg%_2r + seg%seg%_3r + seg%seg%_4r + ~
    seg%seg%_5r + seg%seg%_6r + seg%seg%_7r + seg%seg%_8r + seg%seg%_9r + ~
    seg%seg%_10r + seg%seg%_11r
&end
QUIT
&return

/*-----
&routine weight_1/ln(d)
/*-----
GRID
&do seg &list 1 2 3 4 5 6 7 8
    /* Subtract the rings and apply distance weighting (1/ln(d))
    seg%seg%_2r = (seg%seg%_2w - seg%seg%_1w) * 1.443
    seg%seg%_3r = (seg%seg%_3w - seg%seg%_2w) * 0.910
    seg%seg%_4r = (seg%seg%_4w - seg%seg%_3w) * 0.721
    seg%seg%_5r = (seg%seg%_5w - seg%seg%_4w) * 0.621

```



```

    seg%seg%_6r = (seg%seg%_6w - seg%seg%_5w) * 0.558
    seg%seg%_7r = (seg%seg%_7w - seg%seg%_6w) * 0.514
    seg%seg%_8r = (seg%seg%_8w - seg%seg%_7w) * 0.481
    seg%seg%_9r = (seg%seg%_9w - seg%seg%_8w) * 0.455
    seg%seg%_10r = (seg%seg%_10w - seg%seg%_9w) * 0.434
    seg%seg%_11r = (seg%seg%_11w - seg%seg%_10w) * 0.417

    seg%seg%_tot = seg%seg%_1w + seg%seg%_2r + seg%seg%_3r + seg%seg%_4r + ~
        seg%seg%_5r + seg%seg%_6r + seg%seg%_7r + seg%seg%_8r + seg%seg%_9r + ~
        seg%seg%_10r + seg%seg%_11r
&end
QUIT
&return

/*-----
&routine weight_1/d3
/*-----
GRID
&do seg &list 1 2 3 4 5 6 7 8
    /* Subtract the rings and apply distance weighting (1/d3)
    seg%seg%_2r = (seg%seg%_2w - seg%seg%_1w) * 0.125
    seg%seg%_3r = (seg%seg%_3w - seg%seg%_2w) * 0.037
    seg%seg%_4r = (seg%seg%_4w - seg%seg%_3w) * 0.0156
    seg%seg%_5r = (seg%seg%_5w - seg%seg%_4w) * 0.008
    seg%seg%_6r = (seg%seg%_6w - seg%seg%_5w) * 0.0046
    seg%seg%_7r = (seg%seg%_7w - seg%seg%_6w) * 0.0029
    seg%seg%_8r = (seg%seg%_8w - seg%seg%_7w) * 0.002
    seg%seg%_9r = (seg%seg%_9w - seg%seg%_8w) * 0.0014
    seg%seg%_10r = (seg%seg%_10w - seg%seg%_9w) * 0.001
    seg%seg%_11r = (seg%seg%_11w - seg%seg%_10w) * 0.0008
    seg%seg%_tot = seg%seg%_1w + seg%seg%_2r + seg%seg%_3r + seg%seg%_4r + ~
        seg%seg%_5r + seg%seg%_6r + seg%seg%_7r + seg%seg%_8r + seg%seg%_9r + ~
        seg%seg%_10r + seg%seg%_11r
&end
QUIT
&return

/*-----
&routine index_grid
/*-----
GRID
index_grd = seg1_tot + seg2_tot + seg3_tot + seg4_tot + seg5_tot + seg6_tot ~
+ seg7_tot + seg8_tot

QUIT
&return

/*-----
&routine site_intersect
/*-----

&r M:\Projects\APMoSPHERE\focalsum\gridspot70 index_grd %poll%_allsite75 ~
    index_grd

&return
/* Use SPSS, to calibrate model

/*-----
&routine create_map
/*-----
GRID

/* Customise calibration function!
modno2_bk = exp (2.232 + (0.234 * ln(index_grd)))

QUIT
&return
/*****

```


E1. Evaluation of the monitoring network - 50km analysis

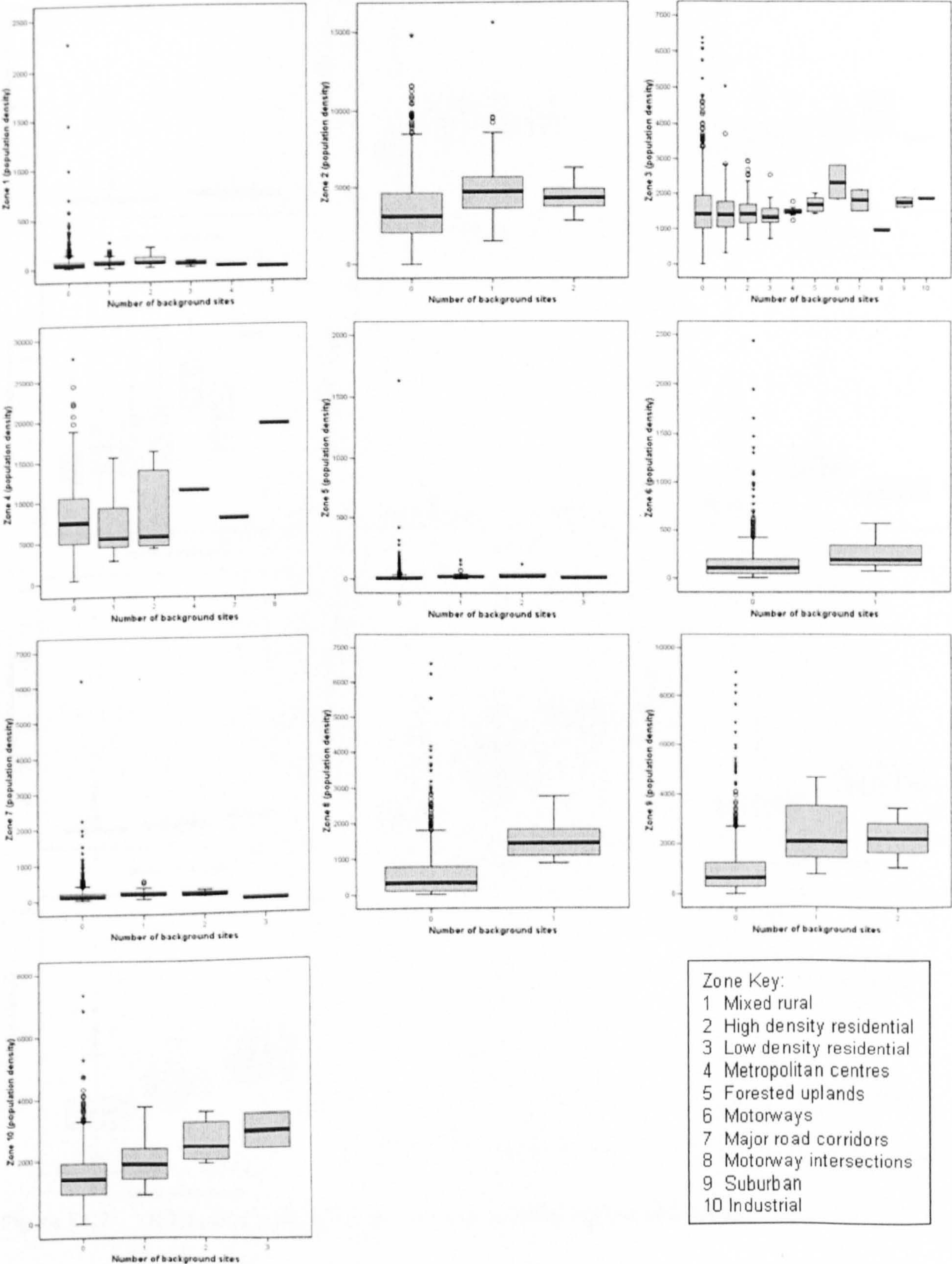


Figure E1.1 NO₂ background sites by affinity zone population density (50x50 km)

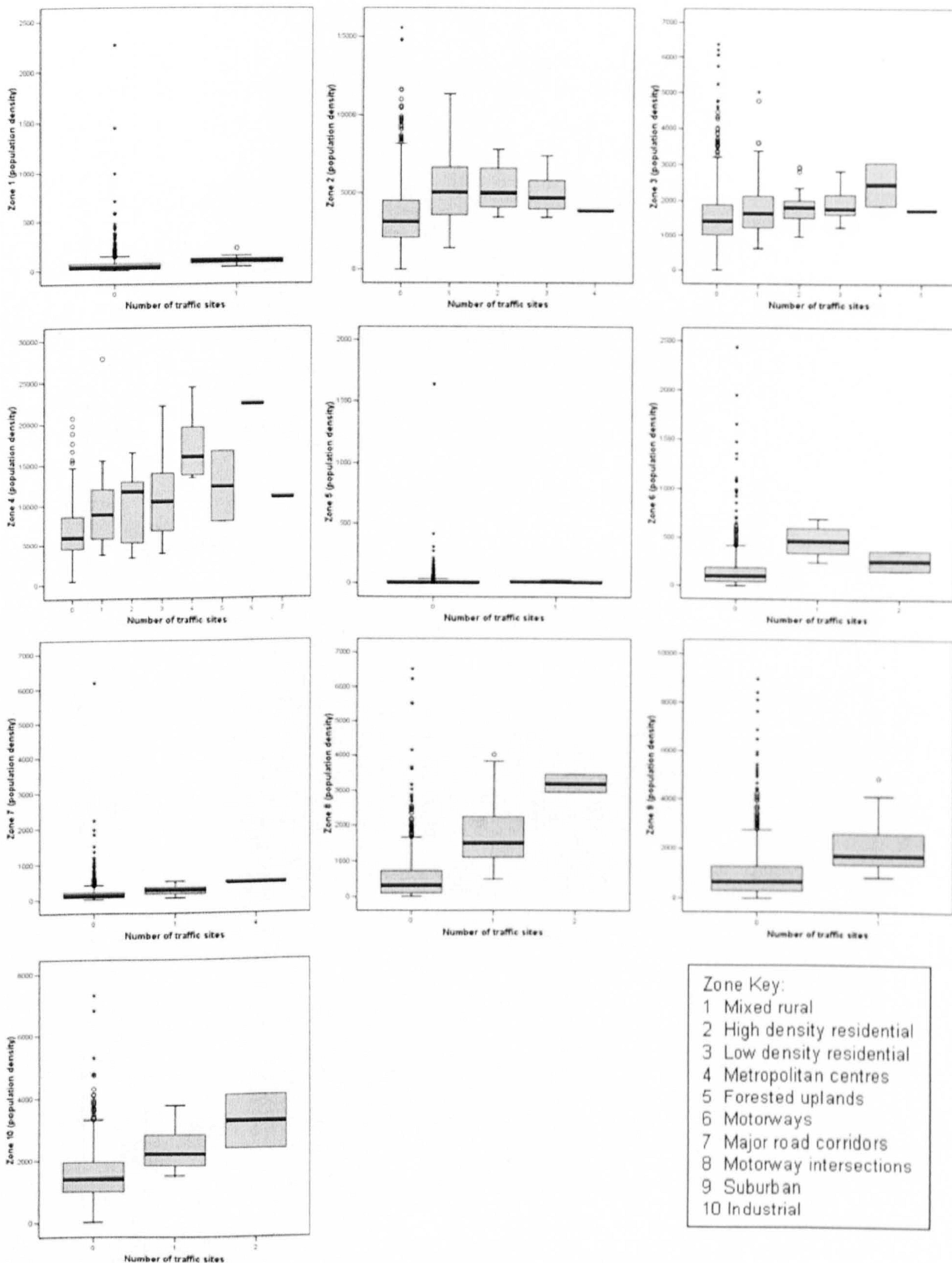


Figure E1.2 NO₂ traffic sites by affinity zone population density (50x50 km)

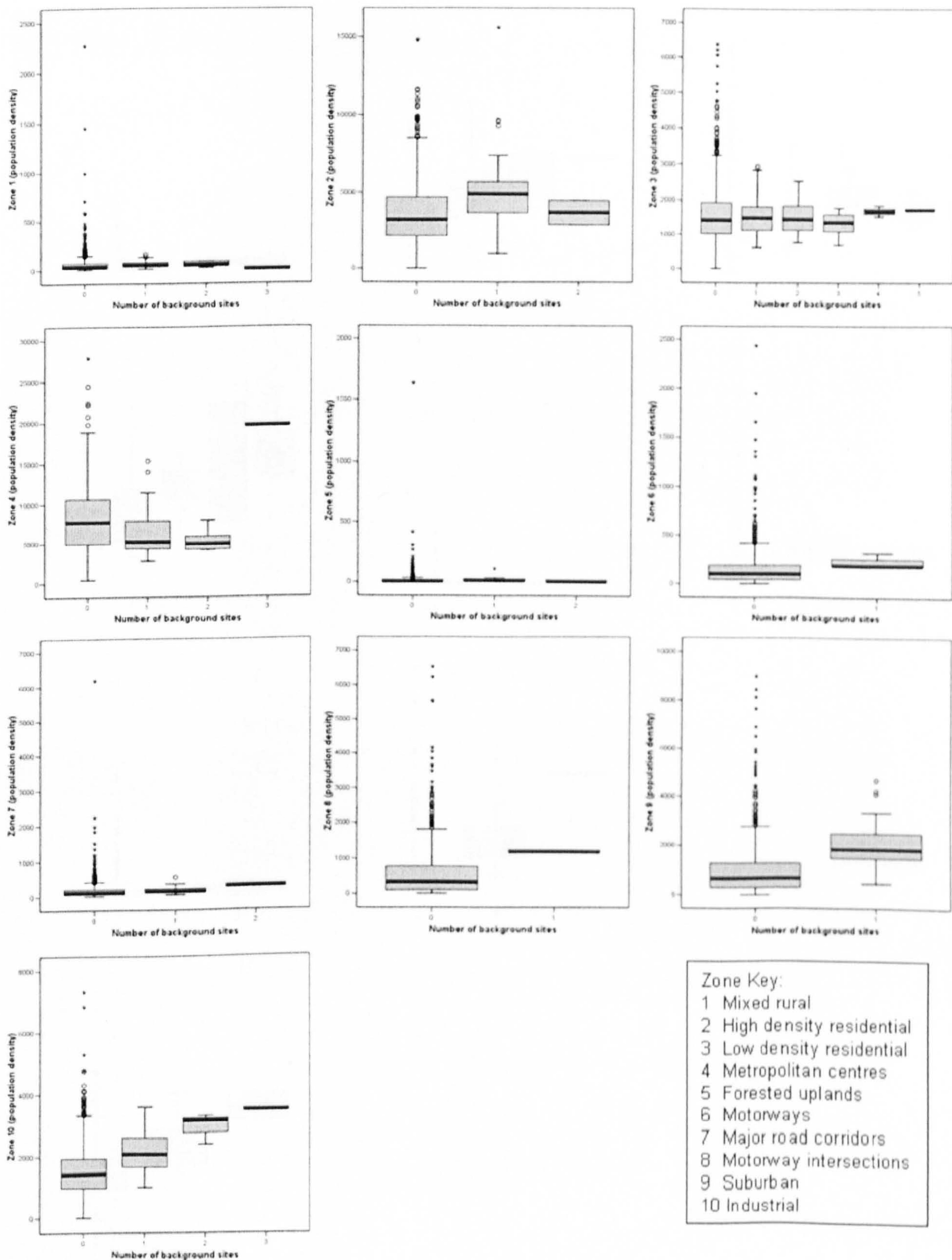


Figure E1.3 PM₁₀ background sites by affinity zone population density (50x50 km)

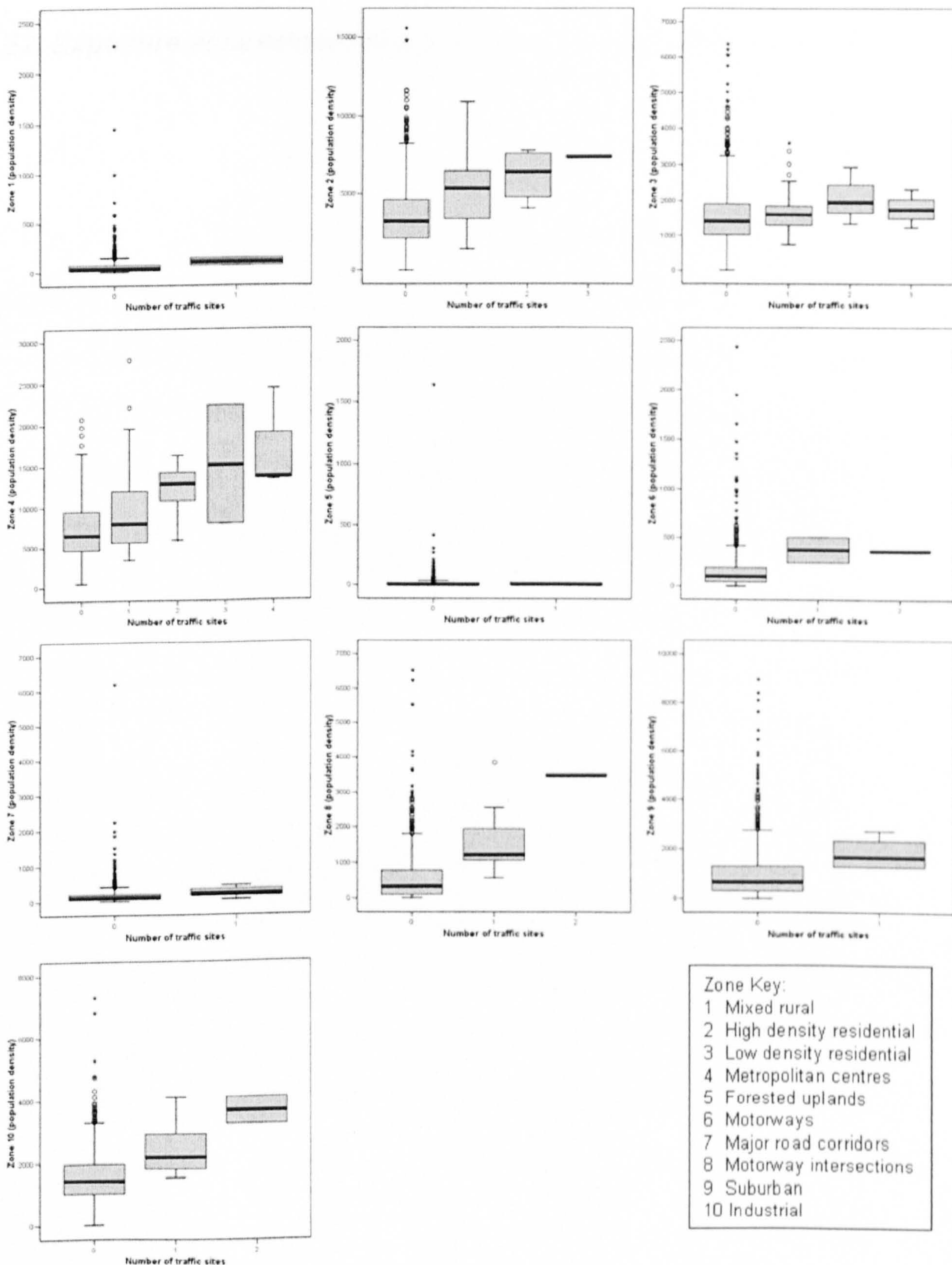


Figure E1.4 PM₁₀ traffic sites by affinity zone population density (50x50 km)

E2. Exposure assessment results

Table E2.1 Breakdown of population by NO₂ exposure class - country
Number of '1000 people per class (Cumulative percent)

	Exposure classes (µg/m³)													% Above guideline
	<5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	>60	
AT	37 (0.5%)	798 (10.4%)	1993 (35.3%)	1729 (56.9%)	1001 (69.4%)	855 (80.0%)	579 (87.3%)	527 (93.8%)	477 (99.8%)	18 (100%)	0 (100%)	0 (100%)	0 (100%)	6.2%
BE	0 (0.0%)	37 (0.4%)	928 (9.6%)	3362 (43.0%)	2908 (71.9%)	1334 (85.2%)	951 (94.7%)	478 (99.4%)	59 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.6%
DE	4 (0.0%)	2101 (2.5%)	17578 (23.3%)	22361 (49.7%)	18240 (71.2%)	14471 (88.3%)	6928 (96.5%)	2200 (99.1%)	649 (99.9%)	75 (100%)	22 (100%)	0 (100%)	0 (100%)	0.9%
DK	1 (0.0%)	211 (3.6%)	1404 (27.5%)	1359 (50.6%)	1184 (70.8%)	796 (84.3%)	508 (93.0%)	273 (97.6%)	139 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	2.4%
ES	254 (0.6%)	1797 (5.2%)	4588 (16.7%)	7022 (34.4%)	6393 (50.4%)	5481 (64.2%)	4248 (74.9%)	3054 (82.6%)	1769 (87.0%)	1393 (90.5%)	1523 (94.4%)	1891 (99.1%)	356 (100%)	17.4%
FI	294 (5.3%)	528 (14.7%)	971 (32.1%)	1124 (52.2%)	1122 (72.3%)	917 (88.7%)	467 (97.1%)	152 (99.8%)	12 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.2%
FR	54 (0.1%)	5030 (8.6%)	13955 (32.2%)	12159 (52.7%)	10213 (69.9%)	7294 (82.3%)	4130 (89.2%)	1753 (92.2%)	1336 (94.5%)	1893 (97.7%)	1278 (99.8%)	113 (100%)	0 (100%)	7.8%
GR	150 (1.3%)	1595 (15.1%)	2623 (37.8%)	1495 (50.7%)	1318 (62.1%)	734 (68.4%)	732 (74.7%)	1116 (84.4%)	1256 (95.3%)	509 (99.7%)	40 (100%)	0 (100%)	0 (100%)	15.6%
IE	37 (1.0%)	923 (25.0%)	728 (43.9%)	568 (58.7%)	510 (71.9%)	474 (84.3%)	466 (96.4%)	114 (99.4%)	25 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.6%
IT	105 (0.2%)	2896 (4.9%)	13896 (27.8%)	16017 (54.2%)	10460 (71.5%)	6969 (82.9%)	4060 (89.6%)	3099 (94.7%)	2815 (99.4%)	353 (100%)	25 (100%)	0 (100%)	0 (100%)	5.3%
LU	0 (0.0%)	7 (1.7%)	74 (20.1%)	106 (46.3%)	88 (68.0%)	100 (92.6%)	30 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
NL	0 (0.0%)	36 (0.2%)	744 (5.0%)	3240 (25.6%)	4975 (57.3%)	3882 (82.1%)	1884 (94.1%)	802 (99.2%)	51 (99.5%)	36 (99.8%)	37 (100%)	0 (100%)	0 (100%)	0.8%
PT	12 (0.1%)	740 (7.4%)	1787 (25.0%)	1912 (43.8%)	1668 (60.3%)	1166 (71.7%)	814 (79.8%)	743 (87.1%)	615 (93.1%)	445 (97.5%)	181 (99.3%)	71 (100%)	0 (100%)	12.9%
UK	87 (0.1%)	1271 (2.3%)	4733 (10.3%)	9754 (26.8%)	13605 (49.9%)	14980 (75.2%)	9353 (91.1%)	3807 (97.5%)	1199 (99.6%)	247 (100%)	8 (100%)	0 (100%)	0 (100%)	2.5%

Table E2.2 Breakdown of population by NO₂ exposure class - affinity zone
Number of '1000 people per class (Cumulative percent)

	Exposure class (µg/m³)													% Above guideline
	<5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	>60	
1	543 (0.6%)	14306 (16.8%)	45250 (68.1%)	21993 (93.0%)	4923 (98.5%)	1039 (99.7%)	195 (99.9%)	55 (100%)	9 (100%)	2 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
2	0 (0.0%)	36 (0.1%)	1334 (3.6%)	7635 (23.7%)	11788 (54.8%)	8909 (78.3%)	4991 (91.4%)	2383 (97.7%)	661 (99.4%)	146 (99.8%)	72 (100%)	0 (100%)	0 (100%)	2.3%
3	0 (0.0%)	37 (0.0%)	7755 (6.0%)	35719 (33.5%)	38274 (62.9%)	28771 (85.0%)	12750 (94.8%)	4512 (98.3%)	1490 (99.5%)	477 (99.8%)	173 (100%)	58 (100%)	0 (100%)	1.7%
4	0 (0.0%)	0 (0.0%)	0 (0.0%)	69 (0.2%)	1424 (4.6%)	4015 (16.9%)	6228 (36.1%)	6694 (56.7%)	6089 (75.4%)	3298 (85.5%)	2488 (93.2%)	1879 (98.9%)	346 (100%)	43.3%
5	491 (5.2%)	3213 (38.9%)	4069 (81.7%)	1479 (97.2%)	228 (99.6%)	30 (99.9%)	4 (100%)	1 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
6	0 (0.0%)	0 (0.0%)	140 (2.5%)	1597 (30.8%)	2220 (70.2%)	1088 (89.5%)	292 (94.7%)	157 (97.5%)	114 (99.5%)	28 (100%)	2 (100%)	0 (100%)	0 (100%)	2.5%
7	1 (0.0%)	259 (1.2%)	6136 (30.6%)	8643 (72.0%)	3522 (88.8%)	1400 (95.5%)	539 (98.1%)	220 (99.1%)	137 (99.8%)	32 (99.9%)	6 (100%)	4 (100%)	0 (100%)	0.9%
8	0 (0.0%)	0 (0.0%)	0 (0.0%)	88 (1.5%)	708 (13.1%)	1526 (38.2%)	1442 (61.9%)	877 (76.4%)	677 (87.5%)	464 (95.1%)	197 (98.4%)	90 (99.8%)	10 (100%)	23.6%
9	0 (0.0%)	113 (0.8%)	900 (6.7%)	1621 (17.6%)	2758 (35.9%)	4265 (64.4%)	3128 (85.2%)	1248 (93.5%)	563 (97.3%)	238 (98.9%)	128 (99.7%)	43 (100%)	0 (100%)	6.5%
10	0 (0.0%)	5 (0.0%)	415 (1.5%)	3365 (13.2%)	7839 (40.7%)	8408 (70.1%)	5579 (89.6%)	1969 (96.5%)	663 (98.8%)	283 (99.8%)	49 (100%)	0 (100%)	0 (100%)	3.5%

Notes: (1) Mixed rural; (2) High density residential; (3) Low density residential; (4) Metropolitan centres; (5) Forested uplands; (6) Motorways; (7) Major road corridors; (8) Motorway intersections; (9) Suburban; (10) Industrial

Table E2.3 Breakdown of population by PM₁₀ exposure class - country
Number of '1000 people per class (Cumulative percent)

	Exposure class (µg/m³)											% Above guideline
	<5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	>50	
AT	0 (0.0%)	27 (0.3%)	269 (3.7%)	2051 (29.3%)	2705 (63.0%)	2576 (95.2%)	381 (99.9%)	3 (100%)	3 (100%)	0 (100%)	0 (100%)	0.0%
BE	0 (0.0%)	0 (0.0%)	9 (0.1%)	404 (4.1%)	2198 (26.0%)	6217 (87.8%)	1227 (100%)	3 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
DE	0 (0.0%)	0 (0.0%)	353 (0.4%)	19729 (23.7%)	50415 (83.3%)	14030 (99.9%)	102 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
DK	0 (0.0%)	0 (0.0%)	1 (0.0%)	2599 (44.3%)	3269 (99.9%)	6 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
ES	0 (0.0%)	3 (0.0%)	129 (0.3%)	2164 (5.8%)	10449 (32.0%)	13091 (65.0%)	5004 (77.5%)	4608 (89.1%)	2139 (94.5%)	1458 (98.2%)	723 (100%)	10.9%
FI	0 (0.0%)	575 (10.3%)	3759 (77.6%)	1243 (99.8%)	10 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
FR	0 (0.0%)	6 (0.0%)	1117 (1.9%)	27158 (47.8%)	23878 (88.1%)	6200 (98.6%)	718 (99.8%)	127 (100%)	3 (100%)	0 (100%)	0 (100%)	0.0%
GR	0 (0.0%)	0 (0.0%)	2 (0.0%)	95 (0.8%)	357 (3.9%)	958 (12.2%)	2787 (36.3%)	4588 (76.0%)	1523 (89.1%)	837 (96.4%)	420 (100%)	24.0%
IE	0 (0.0%)	4 (0.1%)	1062 (27.7%)	2737 (98.9%)	41 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
IT	0 (0.0%)	31 (0.1%)	299 (0.5%)	1175 (2.5%)	4051 (9.2%)	11699 (28.4%)	23500 (67.1%)	13811 (89.9%)	4407 (97.2%)	1346 (99.4%)	377 (100%)	10.1%
LU	0 (0.0%)	0 (0.0%)	6 (1.4%)	388 (97.5%)	10 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%
NL	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (0.0%)	899 (5.7%)	7850 (55.8%)	6848 (99.4%)	84 (100%)	4 (100%)	0 (100%)	0 (100%)	0.0%
PT	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)	70 (0.7%)	848 (9.0%)	4295 (51.3%)	2462 (75.6%)	1958 (94.9%)	420 (99.0%)	101 (100%)	24.4%
UK	0 (0.0%)	4 (0.0%)	3061 (5.2%)	52592 (94.3%)	3356 (100%)	27 (100%)	2 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0.0%

Table E2.4 Breakdown of population by PM₁₀ exposure class - affinity zone
Number of '1000 people per class (Cumulative percent)

		Exposure class (µg/m³)										% Above guideline	
		<5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50		>50
1	Mixed rural	0 (0.0%)	51 (0.1%)	3614 (4.2%)	30494 (38.7%)	18335 (59.4%)	12833 (74.0%)	12280 (87.9%)	7692 (96.6%)	2267 (99.2%)	606 (99.8%)	142 (100%)	3.4%
2	High density residential	0 (0.0%)	10 (0.0%)	542 (1.5%)	6108 (17.5%)	4806 (30.2%)	9440 (55.1%)	8479 (77.4%)	4976 (90.5%)	2336 (96.7%)	882 (99.0%)	377 (100%)	9.5%
3	Low density residential	0 (0.0%)	28 (0.0%)	2212 (1.7%)	43671 (35.3%)	44431 (69.5%)	20379 (85.2%)	9004 (92.1%)	5347 (96.2%)	2862 (98.4%)	1470 (99.5%)	614 (100%)	3.8%
4	Metropolitan centres	0 (0.0%)	0 (0.0%)	42 (0.1%)	5144 (15.9%)	7466 (38.9%)	6800 (59.8%)	7627 (83.2%)	3803 (94.9%)	907 (97.7%)	475 (99.2%)	266 (100%)	5.1%
5	Forested upland	0 (0.0%)	489 (5.1%)	2073 (26.9%)	3933 (68.3%)	1678 (85.9%)	900 (95.3%)	397 (99.5%)	44 (100%)	1 (100%)	0 (100%)	0 (100%)	0.0%
6	Motorways	0 (0.0%)	1 (0.0%)	60 (1.1%)	1355 (25.1%)	1704 (55.3%)	1103 (74.9%)	676 (86.9%)	446 (94.8%)	209 (98.5%)	74 (99.8%)	12 (100%)	5.2%
7	Major road corridors	0 (0.0%)	40 (0.2%)	682 (3.5%)	8118 (42.3%)	6368 (72.8%)	2269 (83.6%)	1360 (90.1%)	1225 (96.0%)	533 (98.5%)	207 (99.5%)	97 (100%)	4.0%
8	Motorway intersections	0 (0.0%)	2 (0.0%)	135 (2.3%)	1001 (18.7%)	2037 (52.2%)	1664 (79.6%)	791 (92.6%)	242 (96.6%)	133 (98.8%)	63 (99.8%)	10 (100%)	3.4%
9	Suburban	0 (0.0%)	0 (0.0%)	227 (1.5%)	5735 (39.7%)	4981 (72.9%)	2360 (88.7%)	1173 (96.5%)	340 (98.7%)	146 (99.7%)	35 (99.9%)	8 (100%)	1.3%
10	Industrial	0 (0.0%)	27 (0.1%)	479 (1.8%)	6779 (25.5%)	9901 (60.1%)	5752 (80.3%)	3078 (91.0%)	1572 (96.5%)	644 (98.8%)	248 (99.7%)	95 (100%)	3.5%

Notes: (1) Mixed rural; (2) High density residential; (3) Low density residential; (4) Metropolitan centres; (5) Forested uplands; (6) Motorways; (7) Major road corridors; (8) Motorway intersections; (9) Suburban; (10) Industrial